Automatic Understanding of Multimodal Content for Web-based Learning

Von der Fakultät für Elektrotechnik und Informatik
der Gottfried Wilhelm Leibniz Universität Hannover

zur Erlangung des Grades
Doktor der Naturwissenschaften
Dr. rer. nat.

genehmigte Dissertation von
M. Sc. Christian Ralf Otto

2023

Referent: Prof. Dr. Ralph Ewerth
Korreferent: Prof. Dr. Johannes Krugel
Tag der Promotion: 20.02.2023

Abstract

Web-based learning has become an integral part of everyday life for all ages and backgrounds. On the one hand, the advantages of this learning type, such as availability, accessibility, flexibility, and cost, are apparent. On the other hand, the oversupply of content can lead to learners struggling to find optimal resources efficiently. The interdisciplinary research field *Search as Learning* is concerned with the analysis and improvement of Web-based learning processes, both on the learner and the computer-science side.

So far, automatic approaches that assess and recommend learning resources in Search as Learning (SAL) focus on textual, resource, and behavioral features. However, these approaches commonly ignore multimodal aspects. This work addresses this research gap by proposing several approaches that address the question of how multimodal retrieval methods can help support learning on the Web. First, we evaluate whether textual metadata of the TIB AV-Portal can be exploited and enriched by semantic word embeddings to generate video recommendations and, in addition, a video summarization technique to improve exploratory search. Then we turn to the challenging task of knowledge gain prediction that estimates the potential learning success given a specific learning resource. We used data from two user studies for our approaches. The first one observes the knowledge gain when learning with videos in a Massive Open Online Course (MOOC) setting, while the second one provides an informal Web-based learning setting where the subjects have unrestricted access to the Internet. We then extend the purely textual features to include visual, audio, and cross-modal features for a holistic representation of learning resources. By correlating these features with the achieved knowledge gain, we can estimate the impact of a particular learning resource on learning success.

We further investigate the influence of multimodal data on the learning process by examining how the combination of visual and textual content generally conveys information. For this purpose, we draw on work from linguistics and visual communications, which investigated the relationship between image and text by means of different metrics and categorizations for several decades. We concretize these metrics to enable their compatibility for machine learning purposes. This process includes the derivation of semantic image-text classes from these metrics. We evaluate all proposals with comprehensive experiments and discuss their impacts and limitations at the end of the thesis.

Keywords: Web-based learning, informal learning, natural language processing, multimodal information extraction, user study, deep learning, knowledge gain prediction, semantic image-text relation, semantic image-text class, semantic gap

Zusammenfassung

Web-basiertes Lernen ist ein fester Bestandteil des Alltags aller Alters- und Bevölkerungsschichten geworden. Einerseits liegen die Vorteile dieser Art des Lernens wie Verfügbarkeit, Zugänglichkeit, Flexibilität oder Kosten auf der Hand. Andererseits kann das Überangebot an Inhalten auch dazu führen, dass Lernende nicht in der Lage sind optimale Ressourcen effizient zu finden. Das interdisziplinäre Forschungsfeld *Search as Learning* beschäftigt sich mit der Analyse und Verbesserung von Web-basierten Lernprozessen.

Bisher sind automatische Ansätze bei der Bewertung und Empfehlung von Lernressourcen fokussiert auf monomodale Merkmale, wie Text oder Dokumentstruktur. Die multimodale Betrachtung ist hingegen noch nicht ausreichend erforscht. Daher befasst sich diese Arbeit mit der Frage wie Methoden des Multimedia Retrievals dazu beitragen können das Lernen im Web zu unterstützen. Zunächst wird evaluiert, ob textuelle Metadaten des TIB AV-Portals genutzt werden können um in Verbindung mit semantischen Worteinbettungen einerseits Videoempfehlungen zu generieren und andererseits Visualisierungen zur Inhaltszusammenfassung von Videos abzuleiten. Anschließend wenden wir uns der anspruchsvollen Aufgabe der Vorhersage des Wissenszuwachses zu, die den potenziellen Lernerfolg einer Lernressource schätzt. Wir haben für unsere Ansätze Daten aus zwei Nutzerstudien verwendet. In der ersten wird der Wissenszuwachs beim Lernen mit Videos in einem MOOC-Setting beobachtet, während die zweite eine informelle Web-basierte Lernumgebung bietet, in der die Probanden uneingeschränkten Internetzugang haben. Anschließend erweitern wir die rein textuellen Merkmale um visuelle, akustische und cross-modale Merkmale für eine ganzheitliche Darstellung der Lernressourcen. Durch die Korrelation dieser Merkmale mit dem erzielten Wissenszuwachs können wir den Einfluss einer Lernressource auf den Lernerfolg vorhersagen.

Weiterhin untersuchen wir wie verschiedene Kombinationen von visuellen und textuellen Inhalten Informationen generell vermitteln. Dazu greifen wir auf Arbeiten aus der Linguistik und der visuellen Kommunikation zurück, die seit mehreren Jahrzehnten die Beziehung zwischen Bild und Text untersucht haben. Wir konkretisieren vorhandene Metriken, um ihre Verwendung für maschinelles Lernen zu ermöglichen. Dieser Prozess beinhaltet die Ableitung semantischer Bild-Text-Klassen. Wir evaluieren alle Ansätze mit umfangreichen Experimenten und diskutieren ihre Auswirkungen und Limitierungen am Ende der Arbeit.

Stichworte: Web-basiertes Lernen, Informelles Lernen, Natürliche Sprachverarbeitung, Vorhersage von Lernerfolg, Multimodale Informationsextraktion, Nutzerstudie, Deep Learning, Semantische Bild-Text Relation, Semantische Lücke

Acknowledgments

I interacted with many people during my work on this thesis, which influenced me to various degrees. I can say with conviction that without their help, this would not have been possible, which is why I want to thank them here.

First, I want to thank my supervisor Prof. Dr. Ralph Ewerth, for granting me the opportunity to embark on this journey by inviting me to Hannover to his newly established Visual Analytics research group. He found a way to resolve my initial doubts about this venture and motivated me to approach the challenges of the scientific world. He helped me find value and purpose in this line of work, be it supervising students, writing papers, or giving talks to rooms full of other researchers.

Parts of this thesis are supported by and have been created in collaboration with partners of the project "*SALIENT*, Search as Learning – Investigating, Enhancing, and Predicting Learning During Multimodal (Web) Search", financially supported by the Leibniz Association, Germany (Leibniz Competition 2018, funding line "Collaborative Excellence" project SALIENT [K68/2017]). The goals of this project were in line with the interdisciplinary ambitions toward improving Web-based learning. I would especially like to thank Dr. Anett Hoppe, Dr. Ran Yu, Georg Pardi, Johannes von Hoyer, and Markus Rokicki for their collaboration.

During my time in the Visual Analytics research group, I had the opportunity to work with several students. In particular, by supervising their bachelor's and master's theses. While I enjoyed working with all of them, I want to thank (in chronological order), especially Justyna Medrek, Hang Zhou, Jianwei Shi, and Markos Stamatakis for their contributions to the publications in this thesis.

As always in life, overcoming obstacles is easier with friends at your side. Here I want to thank Dr.-Ing. Eric Müller-Budack and Matthias Springstein, who were here with me from the beginning. They always provided valuable input if I got stuck with my work in the form of new ideas or the right piece of code for my implementation. Also, I have to thank Eric, in particular, for being a pacemaker in the final stages of my thesis, pushing me toward the finish line.

Lastly, I want to thank my parents, my brother, and most importantly my beloved wife, Lisa, for always being supportive during these years. In the final year of the Ph.D., you brought our son Joschua into our life and completed our family bliss. Even though the short nights put a slight damper on the speed of completion of this thesis, he also provided the final push of motivation necessary. I dedicate this thesis to him.

List of Tables

List of Figures

Contents

1 Introduction

1.1 Motivation

The advent of web-based learning, driven by two decades of digitization, has proven its worth and necessity during the Covid-19 pandemic. Internationally, children, college students, researchers, and employees of sectors like, for instance, public administration or information and communication, were forced to work and, to some extent, learn in front of their computer screens [52]. However, a clear trend was noticeable even before this worldwide situation arose. By 2018 one in three students in the US was enrolled in an online course [205], 22 of the top 25 US universities offer online courses for free [169], and even 45% of elementary school students report their favorite learning methods to be educational games and online videos [282].

Depending on many subjective and topic-related factors, web-based learning has certain disadvantages compared to traditional learning. Disadvantages are, for example, the less motivating, impersonal online classrooms, the associated social isolation, technical inequalities for different social backgrounds [195], and the lack of assessment for a majority of online resources when compared to, for instance, text books [236]. However, there are many advantages connected to this trend as well. At first, it is accessible and, thus, convenient. Given a device that is able to connect to the internet, a learner is free to consume content any time of the day and, with enough network coverage, wherever he or she wants. The second advantage, affordability, underlines accessibility even more. The increasing cost of traditional education, ranging from textbooks over public transportation to college tuition, traditionally prevents students from families with lower incomes from partaking in higher education [61]. Lastly, from an individual's perspective, Web-based learning allows for more flexibility, enabling work-life balance, and can also be adapted to personal preferences and interests.

With this trend comes the need for a better automatic understanding of learning material. That entails, improving computer-based algorithms in their ability to describe complex content similar to humans. Otherwise, it becomes more and more challenging to explore the vast amount of available content. In other words, typical information retrieval (IR) methods in search engines are not tailored toward the learner, but monetary gain in terms of, for example, ad revenue. A learner-focused approach, however, needs reliable ways to generate optimal results for learners. This task is challenging due to multiple factors. First, the given information often only consists of a search query where the learner might not even be sure whether it fits their needs. Occasionally, platforms collect historical data

of previous searches that hint at personal interests or fields of study. In a perfect world, a retrieval algorithm has to recommend an optimal set of resources to achieve the desired learning goal of the learner, given these sparse clues. However, this algorithm requires a thorough and human-like understanding of the respective databases' textual, visual, and audio-visual material. Simple author-provided annotations like keywords or tags, combined with popularity measures and video categorizations, allow for good, superficial results, at least on the entertainment side of video consumption (e.g., YouTube [175]). Nevertheless, techniques like these are prone to generate unsatisfying results due to clickbait titles [270] because their annotations are not guaranteed to have a direct link to the content. On the other hand, semantic feature extraction methods, realized for example through semantic word embeddings [182, 22], directly derive metadata from the content of a given material, which, in theory, circumvents this problem. The trade-off here is the accuracy of the generated labels, which is limited by an algorithm's capabilities to interpret not only text, image, and audio individually, but also their combinations. Further, it has to factor in subjective quality measures for learning resources to make the correct decision when being forced to decide between, e.g., two content-wise identical videos with significant differences in presentation quality.

1.2 Summary of the State-of-the-Art

This section provides a selection of related work to provide context for the research questions and associated contributions in this thesis. Search as Learning is the interdisciplinary research area that deals with all topics surrounding Web-based learning. That entails thoroughly investigating every aspect of learning sessions with an *informational* search intent (besides *transactional* or *navigational* intents [34]), thus implying the intent to acquire knowledge. Search as Learning (SAL) is considered interdisciplinary, because it considers insights and techniques from psychology, educational sciences, and computer science. As stated by Hoppe et al. [113], this encompasses (a) improving the retrieval and ranking process of search engines, (b) predicting and considering individual knowledge states, intents, and needs, and (c) taking all forms of formal and informal learning settings into consideration, especially the wide range of available types of multimodal content (textual, visual and audio-visual). Past research on SAL has, however, widely focused on the exploration of behavioral features (e.g., [47, 78]), and textual features of Web resources [250], neglecting the impact of multimodal data [69]. However, multimodal research that evaluates and measures the importance of visual and audio-visual information for these tasks is still in its infancy. Even though modern approaches show excellent results detecting *what* is seen in an image, SAL attempts to understand *how* the shown information influences the learning outcome *in combination* with other modalities considering the current characteristics of the individual user. These characteristics entail previously acquired domain knowledge [285, 193], Web search literacy [284, 286], or working memory capacity [207], and task characteristics, such as the type of knowledge to be acquired (e.g.,

factual, conceptual, or procedural knowledge) [266] and the cognitive process dimension (e.g., understand or apply) [6].

There is evidence that visual elements can have an impact on both, searching and learning. Several studies have examined how images may help searchers to find the information they are looking for [196, 134]. They show that they can be used to guide their attention and allow for a more efficient identification of relevant (passages in) Web resources.

Research on learning goes into more depth and analyzes how text, images, and videos have to be combined to enable efficient learning. However, the fact that multimedia representation does support many types of learning tasks can be considered as well established [148]. Amadieu et al. [5] state that a combination of hypertext elements, animations, and other multimedia can stimulate "deep processing of the material", but may also lead to problems due to the split attention effect [171].

The risks of multimedia material to cause additional cognitive load for a learner has been discussed by Mayer and Moreno [172]. Therefore, recent research closely analyzes the interplay of different modalities with respect to learning outcomes. For instance, studies explore how the distribution of information on text and image influences their integration [233, 232], the effect of the temporal sequence of presentation [9] and how inter-modality signaling can support learners [223]. Similarly, video learning has been researched with respect to the usefulness of structure elements [97, 4], interaction functionalities, [180, 179, 58], and added functionalities for engagement [186].

In summary, there is clear evidence that the integration of multimedia resources does support human learning in general, and that this transfers to Web environments. However, the composition of learning resources play an important role. It can ensure the efficient communication of learning contents, but it can also lead to cognitive overload and distraction. In consequence, a detailed, large-scale analysis of multimedia features in learning-oriented resources may lead to a better understanding of Web-based learning processes.
More facets of SAL will be discussed in Section 4.2.2.

1.3 Problem Statement and Research Questions

The result of the growth of available online resources is the challenge to automatically understand and, consequently, index the plethora of multimodal information associated with them to recommend optimal learning resources to learners based on their individual needs and knowledge states. The following subsections highlight the individual steps and challenges towards this goal. Motivated by these open questions, we will derive the research questions that we will answer in the upcoming chapters.

1.3.1 Exploiting Textual Metadata to improve Web-based Learning

The automatic analysis of audio-visual content is a challenging task. As pointed out by Beyer et al. [17], a fundamental problem of this research is the *semantic gap* between low-level features and high-level semantics portrayed in the visual domain. To narrow down this gap should enable us to improve the explorative search capabilities, like video recommendation of content summarization, of video search engines like the TIB AV-Portal [259].

Usually, video recommender systems rely on user-based information, for example viewing history [56] or current trends [54]. However, these methods are prone to fail when the search engine does not want to record this type of information, the user utilizes privacy software, or visits the website for the first time [146]. This raises the question whether possible watching interests can be made based on the currently watched video, or, to be more precise, the metadata associated with it. A similar challenge is presented for video summarization techniques in the context of educational videos. Related works aim to generate a short synopsis of a given video that focuses on visual features in the form of key frames or key fragments [8]. They are stitched then together to form a story board. However, in the educational domain (e.g. lecture videos), where the content lacks visual variance, these methods struggle to provide good solutions [290].

A possible avenue to solve this problem is to further process, enrich, or repurpose the textual metadata associated with each video. Metadata is, per definition, structured data about data [82]. Or, in other words, textual information that describe (digital) objects. Consequently, we can apply Natural Language Processing (NLP)-based content analysis methods, like semantic word embeddings [182, 22]. They have the potential to improve educational video Web platforms without the need for extensive data-gathering or high processing power usually connected to algorithms that consider the visual domain. These considerations, together with the mentioned shortcomings of popular video recommendation and video summarization techniques, raise our first research question:

Research Question 1

How can we utilize textual metadata associated with learning content to improve exploratory search in video search portals?

1.3.2 Extraction of Multimodal Features for Knowledge Gain Prediction

There are numerous ways to exploit text-based content for more effective information retrieval algorithms [300, 156, 50] in SAL, especially with the advancements in NLP. Features from visual and audio-visual content are, however, still somewhat underrepresented in information retrieval systems [114]. Commercial video platforms such as YouTube rely on manual metadata such as titles, descriptions, or topic categories [56] in conjunction

with easy-to-compute, text-based video features like continuous bag-of-words [54], which might be too unspecific or insufficient for longer, educational content like lecture videos. This stands in contrast to current learning research which suggests that users may prefer images and videos when tackling certain learning needs (e.g., procedural learning tasks [80, 209]). Since nowadays almost all media is multimodal (i.e., contains textual, visual, and audio-visual information), especially in education, it seems necessary to consider these modalities to understand the meaning of a multimodal document thoroughly. To capture arbitrary layouts of Web-content and consequently, their content, requires a document layout analysis (DLA). Machine learning algorithms that approach this task have been proposed, for example, for historical documents [283, 237], scientific papers [303], or handwritten pages [75]. However, neither a dedicated dataset nor an end-to-end approach have been published for the purpose of DLA on Web documents, yet. Besides the missing visual components of the learning resource, recent works [293] showed that the consideration of behavioral features summarizing the interaction of the learner with the computer, and resource features that provide statistics (e.g., readability, complexity, linguistics) about the content, are beneficial for Knowledge Gain (KG) prediction as well.

Even though multimodal feature extraction is already a complex task, it is only a preprocessing step to improve web-based learning [117], see Figure 1.1.

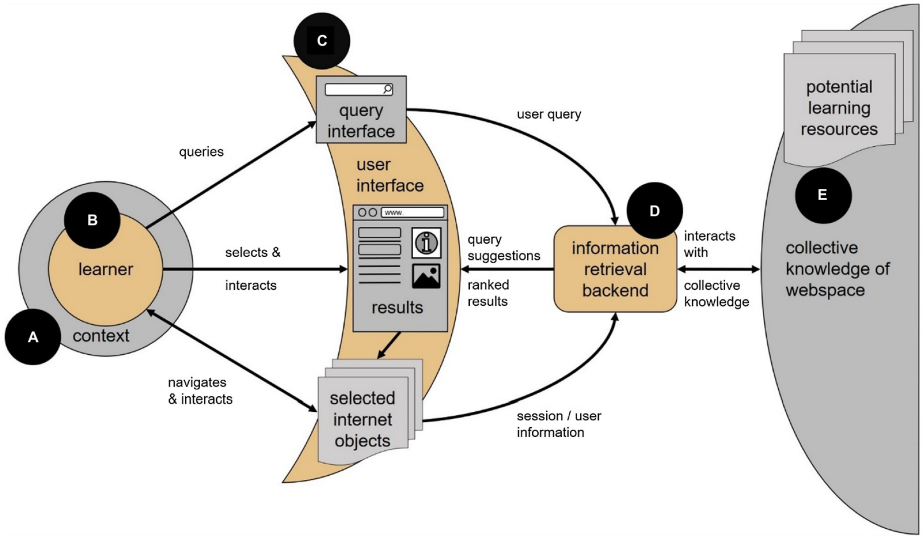


FIGURE 1.1: The *Salient* spaceship model outlining the main components of the *informal* Web-based learning process, namely the learner, the interface and the IR backend (D). This thesis investigates the feature extraction process and the feature processing part of the IR backend and analyzes how the information displayed on the interface (C) influence the learners knowledge gain.

Statistics about multimodal learning resources need to be aligned with realistic learning results to understand how others might benefit from a certain type of learning material [208, 293]. In SAL, this type of data is collected via user studies that resemble *formal* or *informal* learning settings. A formal setting, in this context, means controlled environments where

participants are constrained by design regarding, for example, available websites, learning resources, resource type, or generally how they are supposed to consume knowledge. Brockman and Dirkx [33] found that such a scenario provides critical thinking and independent learning skills that people need to perform well in demanding work situations. It also enhances their ability and desire to learn on their own [158].

Learning in an informal setting on the other hand, is mainly initiated by individuals in their everyday life, characterized by free access to any search engine and available Web resource. In the context of user studies, the challenge is to replicate real-life scenarios as well as possible while simultaneously enforcing a specific learning task. In addition, the participants should not feel constrained in their actions and act naturally. Otherwise, the drawn conclusions may not apply to general learning scenarios. If possible, they should be allowed to use familiar browsers and search engines, and the underlying recording and tracking software should not disturb the learning process.

It is also helpful to choose topics that are somewhat interesting to the participant, which brings up the following challenge: to choose the appropriate task topic and task type [106]. The goal when choosing the task topic is to cover two aspects: on the one hand, ensure that it is not too simple, or, in other words, the average participant does not already know everything about it. Otherwise a gain in knowledge would not be possible. On the other hand, a topic too complex would prevent the person from learning as well. Further, we need to align the task type (informational, navigational, or procedural [34]) with the hypotheses to be proven. Lastly, the pre- and post-knowledge questionnaires must be well chosen to capture the respective knowledge states given a topic, but not too easy to allow for guessing the correct answers by the elimination method. The sum of these parameters influence how well the gathered results generalize to other experiments and, thus, the external validity of this kind of experiment.

Elaborate studies that record all types of interdisciplinary data, meaning knowledge metrics as well as multimodal, gaze, resource, and behavioral features from such carefully constructed studies have neither been published, nor automatically investigated, yet. In summary, automatic multimodal analyses in SAL are still in their infancy in multiple parts of the learning process, which brings up our second research question:

Research Question 2

To what extent can we extract textual, multimedia, and cross-modal features and utilize them for knowledge gain prediction?

1.3.3 Computable Crossmodal Relations

Exploring how different modalities act together to convey an author's intended message might help to get an even better understanding of how information is conveyed from

medium to learner. For the scope of this thesis, we take an in-depth look at the relationship between visual and associated textual information purposely put in place together.

As Bateman states [16], at first glance, combining image and text to convey information of any kind seems to be a natural and easy thing to do. People of all ages do it every day and have done so for hundreds of years. On second thought, by asking *how* these two *modes* work together to create an intended meaning, one quickly comes to realize that image and text are very different tools whose interplay has many potential interpretations. This difference in expressiveness between two linguistic representations, or modalities [84], is called the *semantic gap* [102]. In computer science, Smeulders et al. [243] describe it as the lack of coincidence between the information in visual data and their possible interpretations. Due to the semantic gap, images and text are very rarely able to portray the same information [103, 104]. Conversely, in most cases they are meant to complement each other with one being dominant regarding the amount of information brought into the joint message.

To get a basic understanding of these interplays, it is beneficial to define a set of interpretable, comprehensive, and computable metrics that describe the relationship between visual and textual content in detail. The approach of Henning and Ewerth [103, 104] creates a foundation of such a metric-based distinction, but has not been aligned with research from communication and media science. However, there are various taxonomies and in-depth discussions about semantic image-text classes, which, as this thesis will elaborate on, are based on underlying metrics [15, 166, 167, 264]. So far, a transfer of linguistics knowledge into computer science approaches has only been attempted partially, usually pruned to the problem at hand instead of generally applicable to arbitrary media. Kruk et al. [143], for instance, tailor Marsh and White's taxonomy [166] to measure the author's intent for Instagram posts in terms of two different relationship measures. Zhang et al. [298], on the other hand, investigate only one relationship in the advertisements domain, determining whether a equivalent or non-equivalent parallel information transfer is present. Lastly, diverse, domain-independent, and sufficiently large datasets, which contain a broad range of semantic image-text metrics, do not exist yet. These observations pose our final research question:

Research Question 3

Based on insights from linguistics and visual communications, how can we derive computational models that describe the relationship between image and text?

1.4 Contributions

1.4.1 Improving Video Learning Platforms with Text-Based Features

Chapter 3 explores how, for instance, semantic word embeddings and keyphrase extraction methods, next to others, can be used to improve web-based educational applications to enhance the user experience. In particular, we conduct two user studies on the open-access dataset provided by the *TIB AV-Portal* [259], comprised of scientific videos (e.g., lecture recordings) enriched with metadata composed of a speech transcript and keywords derived from Optical Character Recognition (OCR) and Visual Concept Detection (VCD). We demonstrate and evaluate how this data can be post-processed and extended to provide two improvements to the educational video Web platforms. First, a novel video recommendation tool (Section 3.1) is implemented that suggests related videos based on video content rather than title similarity. Second, we present a visual summary visualization that allows for a more efficient explorative search for learners that try to find a fitting video for their query (Section 3.2).

1.4.2 Prediction of Knowledge Gain with Multimodal Features

Chapter 4 investigates how to assess multimedia educational content from two directions. In a smaller study (Section 4.1) that focuses on MOOC videos, we propose a new feature set comprised of acoustic and visual features but also their cross-modal combinations with the shown textual content. A correlation analysis with the measured learning outcome will indicate their usefulness for an eventual knowledge gain prediction. To further strengthen this assumption, we extend this cross-modal dataset with a wide range of textual features and compare their potential for knowledge gain in a large study.

In our second lab study, carried out by 114 participants, we implemented an informal setting, meaning we extended our scope to unrestricted search on the internet (Section 4.2). To align the chosen, highly diverse resources with the learning outcome, we recorded a plethora of log data. In an attempt to capture the full range of stimuli, we record the learner's gaze, the visited websites in chronological order, the individual behavior (for instance, mouse movements and actions), screen capture, and a wide range of knowledge metrics. We propose an automatic framework that requires minimal manual labor but can extract statistics about the design of the seen websites and classify their content. We use these features in conjunction with a set of behavioral and resource features to achieve state-of-the-art results in knowledge gain prediction.

1.4.3 Categorization of Semantic Image-Text Relations

In Chapter 5 of this thesis, we build upon recent work on image-text metrics and bridge the gap between research in computer science and communication science. Based on previous proposals [103, 15], we define three semantic image-text metrics to describe three

dimensions of the semantic interplay of jointly placed visual and textual information. Subsequently, we derive a categorization of eight semantic image-text classes that combines research from communication science with these metrics and thus, provides a general system to categorize image-text pairs. Next, we show how modern deep learning-based methods, specifically multimodal embeddings, can be utilized to predict these metrics (and classes). Therefore, we employ data augmentation methods to generate a dataset of about 240 000 image-text pairs, which is available to the public (Chapter 5).

1.5 List of Publications

The following papers have been published in the context of this thesis. As in most academic work, other authors have contributed to a certain extent to these publications. Thus, the academic *we* is used throughout this thesis. Below the individual abstracts my contributions to each respective paper are listed under **My Contributions** according to the Contributor Roles Taxonomy (CRediT) [28].

Two papers have been published at conferences ranked A ([202, 201] and three at conferences ranked B ([177, 304, 200] according to the *Australian Computing Research & Education* (CORE[1]) Conference Portal (source: CORE2021). "*Understanding, Categorizing and Predicting Semantic Image-Text Relations*" [200] received the *Best Paper Award* at the *ACM International Conference for Multimedia Retrieval 2019* and was subsequently invited as an extended version called "*Characterization and classification of semantic image-text relations*" [199] in the *International Journal of Multimedia Information Retrieval (IJMIR)*. The paper entitled "*Investigating Correlations of Automatically Extracted Multimodal Features and Lecture Video Quality*" [240] was presented at the *International Workshop on Search as Learning with Multimedia Information* co-located with an A* conference (*ACM International Conference on Multimedia*). In the following section, all publications are outlined and set into context with their respective chapters.
Chapter 3 presents two approaches to improve the TIB AV-Portal [259], a scientific video Web platform, with text-based features and is based on the publications "*Recommending Scientific Videos Based on Metadata Enrichment Using Linked Open Data*" [177] and "*Visual Summarization of Scholarly Videos Using Word Embeddings and Keyphrase Extraction*" [304].

[177] Justyna Medrek, Christian Otto, and Ralph Ewerth. "Recommending Scientific Videos Based on Metadata Enrichment Using Linked Open Data". In: *Digital Libraries for Open Knowledge, 22nd International Conference on Theory and Practice of Digital Libraries, TPDL 2018, Porto, Portugal, September 10-13, 2018, Proceedings*. Vol. 11057. Lecture Notes in Computer Science. Springer, 2018, pp. 286–292. DOI: 10.1007/978-3-030-00066-0_25

Abstract: The amount of available videos in the Web has significantly increased not only for entertainment etc., but also to convey educational or scientific information

[1] http://portal.core.edu.au/conf-ranks/

in an effective way. There are several web portals that offer access to the latter kind of video material. One of them is the TIB AV-Portal of the Leibniz Information Centre for Science and Technology (TIB), which hosts scientific and educational video content. In contrast to other video portals, automatic audiovisual analysis (VCD, OCR, Automatic Speech Recognition (ASR)) is utilized to enhance metadata information and semantic search. In this paper, we propose to further exploit and enrich this automatically generated information by linking it to the Integrated Authority File (GND) of the German National Library. This information is used to derive a measure to compare the similarity of two videos which serves as a basis for recommending semantically similar videos. A user study demonstrates the feasibility of the proposed approach.

My Contributions: Conceptualization, Formal Analysis, Project administration, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing

[304] Hang Zhou, Christian Otto, and Ralph Ewerth. "Visual Summarization of Scholarly Videos Using Word Embeddings and Keyphrase Extraction". In: *Digital Libraries for Open Knowledge - 23rd International Conference on Theory and Practice of Digital Libraries, TPDL 2019, Oslo, Norway, September 9-12, 2019, Proceedings*. Vol. 11799. Lecture Notes in Computer Science. Springer, 2019, pp. 327–335. DOI: 10.1007/978-3-030-30760-8_28

Abstract: Effective learning with audiovisual content depends on many factors. Besides the quality of the learning resource's content, it is essential to discover the most relevant and suitable video in order to support the learning process most effectively. Video summarization techniques facilitate this goal by providing a quick overview over the content. It is especially useful for longer recordings such as conference presentations or lectures. In this paper, we present a domain specific approach that generates a visual summary of video content using solely textual information. For this purpose, we exploit video annotations that are automatically generated by ASR and video OCR. Textual information is represented by semantic word embeddings and extracted keyphrases. We demonstrate the feasibility of the proposed approach through its incorporation into the TIB AV-Portal (http://av.tib.eu/), which is a platform for scientific videos. The accuracy and usefulness of the generated video content visualizations is evaluated in a user study.

My Contributions: Conceptualization, Formal Analysis, Project administration, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing

The KG prediction methods in Chapter 4 are based on two user studies and a total of four publications. Section 4.1 describes the contributions of the papers "*Investigating Correlations of Automatically Extracted Multimodal Features and Lecture Video Quality*" [240] and

"*Predicting Knowledge Gain for MOOC Video Consumption*" [201]. In these two publications we investigate learning in a formal setting (MOOC videos) by conducting a user study and extensive, multimodal feature extraction procedure. Afterward, Section 4.2 explores learning in an informal setting in a larger lab study, the creation of an extensive dataset, and proposes a novel multimedia extraction framework for subsequent knowledge gain prediction. This work is published in the "*SaL-Lightning Dataset: Search and Eye Gaze Behavior, Resource Interactions and Knowledge Gain during Web Search*" [198] and "*Predicting Knowledge Gain During Web Search Based on Multimedia Resource Consumption*" [202].

[240] Jianwei Shi, Christian Otto, Anett Hoppe, Peter Holtz, and Ralph Ewerth. "Investigating Correlations of Automatically Extracted Multimodal Features and Lecture Video Quality". In: *Proceedings of the 1st International Workshop on Search as Learning with Multimedia Information*. SALMM '19. Nice, France: Association for Computing Machinery, 2019, pp. 11–19. ISBN: 9781450369190. DOI: 10.1145/3347451.3356731

Abstract: Ranking and recommendation of multimedia content such as videos is usually realized with respect to the relevance to a user query. However, for lecture videos and MOOC it is not only required to retrieve relevant videos, but particularly to find lecture videos of high quality that facilitate learning, for instance, independent of the video's or speaker's popularity. Thus, metadata about a lecture video's quality are crucial features for learning contexts, e.g., lecture video recommendation in search as learning scenarios. In this paper, we investigate whether automatically extracted features are correlated to quality aspects of a video. A set of scholarly videos from a MOOC is analyzed regarding audio, linguistic, and visual features. Furthermore, a set of cross-modal features is proposed which are derived by combining transcripts, audio, video, and slide content. A user study is conducted to investigate the correlations between the automatically collected features and human ratings of quality aspects of a lecture video. Finally, the impact of our features on the knowledge gain of the participants is discussed.

My Contributions: Conceptualization, Formal Analysis, Project administration, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing

[201] Christian Otto, Markos Stamatakis, Anett Hoppe, and Ralph Ewerth. "Predicting Knowledge Gain for MOOC Video Consumption". In: *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners' and Doctoral Consortium - 23rd International Conference, AIED 2022, Durham, UK, July 27-31, 2022, Proceedings, Part II*. ed. by Maria Mercedes T. Rodrigo, Noburu Matsuda, Alexandra I. Cristea, and Vania Dimitrova. Vol. 13356. Lecture Notes in Computer Science. Springer, 2022, pp. 458–462. DOI: 10.1007/978-3-031-11647-6_92

Abstract: Informal learning on the Web using search engines as well as more structured learning on MOOC platforms have become very popular. However, the automatic assessment of this content with regard to the challenging task of predicting (potential) knowledge gain has not been addressed by previous work yet. In this paper, we investigate whether we can predict learning success after watching a specific type of MOOC video using 1) multimodal features, and 2) a wide range of text-based features describing the structure and content of the video. In a comprehensive experimental setting, we test four different classifiers and various feature subset combinations. We conduct a feature importance analysis to gain insights in which modality benefits knowledge gain prediction the most.

Source Code: https://github.com/TIBHannover/mooc_knowledge_gain

My Contributions: Conceptualization, Formal Analysis, Project administration, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing

[198] Christian Otto, Markus Rokicki, Georg Pardi, Wolfgang Gritz, Daniel Hienert, Ran Yu, Johannes von Hoyer, Anett Hoppe, Stefan Dietze, Peter Holtz, Yvonne Kammerer, and Ralph Ewerth. "SaL-Lightning Dataset: Search and Eye Gaze Behavior, Resource Interactions and Knowledge Gain during Web Search". In: *CHIIR '22: ACM SIGIR Conference on Human Information Interaction and Retrieval, Regensburg, Germany, March 14 - 18, 2022*. Ed. by David Elsweiler. ACM, 2022, pp. 347–352. DOI: 10.1145/3498366.3505835

Abstract: The emerging research field SAL investigates how the Web facilitates learning through modern information retrieval systems. SAL research requires significant amounts of data that capture both search behavior of users and their acquired knowledge in order to obtain conclusive insights or train supervised machine learning models. However, the creation of such datasets is costly and requires interdisciplinary efforts in order to design studies and capture a wide range of features. In this paper, we address this issue and introduce an extensive dataset based on a user study, in which 114 participants were asked to learn about the formation of lightning and thunder. Participants' knowledge states were measured before and after Web search through multiple-choice questionnaires and essay-based free recall tasks. To enable future research in SAL -related tasks we recorded a plethora of features and person-related attributes. Besides the screen recordings, visited Web pages, and detailed browsing histories, a large number of behavioral features and resource features were monitored. We underline the usefulness of the dataset by describing three, already published, use cases.

My Contributions: Conceptualization, Data curation, Project administration, Resources, Software, Visualization, Writing – original draft, Writing – review & editing

[202] Christian Otto, Ran Yu, Georg Pardi, Johannes von Hoyer, Markus Rokicki, Anett Hoppe, Peter Holtz, Yvonne Kammerer, Stefan Dietze, and Ralph Ewerth. "Predicting Knowledge Gain During Web Search Based on Multimedia Resource Consumption". In: *Artificial Intelligence in Education - 22nd International Conference, AIED 2021, Utrecht, The Netherlands, June 14-18, 2021, Proceedings, Part I*. vol. 12748. Lecture Notes in Computer Science. Springer, 2021, pp. 318–330. DOI: 10.1007/978-3-030-78292-4_26

Abstract: In informal learning scenarios the popularity of multimedia content, such as video tutorials or lectures, has significantly increased. Yet, the users' interactions, navigation behavior, and consequently learning outcome, have not been researched extensively. Related work in this field, also called *search as learning*, has focused on behavioral or text resource features to predict learning outcome and knowledge gain. In this paper, we investigate whether we can exploit features representing multimedia resource consumption to predict KG during Web search from in-session data, that is without prior knowledge about the learner. For this purpose, we suggest a set of multimedia features related to image and video consumption. Our feature extraction is evaluated in a lab study with 113 participants where we collected data for a given search as learning task on the formation of thunderstorms and lightning. We automatically analyze the monitored log data and utilize state-of-the-art computer vision methods to extract features about the seen multimedia resources. Experimental results demonstrate that multimedia features can improve KG prediction. Finally, we provide an analysis on feature importance (text and multimedia) for KG prediction.

My Contributions: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing

Finally, the main contribution of Chapter 5 is based on the paper in "*Characterization and classification of semantic image-text relations*" [199], which is an extended journal version of "*Understanding, Categorizing and Predicting Semantic Image-Text Relations*" [200]. In these publications, we propose and define three semantic image-text metrics and a categorization of eight semantic image-text classes. We create a large dataset and suggest a neural network-based method for automatic classification.

[200] Christian Otto, Matthias Springstein, Avishek Anand, and Ralph Ewerth. "Understanding, Categorizing and Predicting Semantic Image-Text Relations". In: *Proceedings of the 2019 on International Conference on Multimedia Retrieval, ICMR 2019, Ottawa, ON, Canada, June 10-13, 2019*. Ed. by Abdulmotaleb El-Saddik, Alberto Del Bimbo, Zhongfei Zhang, Alexander G. Hauptmann, K. Selçuk Candan, Marco Bertini, Lexing Xie, and Xiao-Yong Wei. ACM, 2019, pp. 168–176. DOI: 10.1145/3323873.3325049

[199] Christian Otto, Matthias Springstein, Avishek Anand, and Ralph Ewerth. "Characterization and classification of semantic image-text relations". In: *International Journal of Multimedia Information Retrieval* 9.1 (2020), pp. 31–45. DOI: 10.1007/s13735-019-00187-6

Abstract: The beneficial, complementary nature of visual and textual information to convey information is widely known, for example, in entertainment, news, advertisements, science, or education. While the complex interplay of image and text to form semantic meaning has been thoroughly studied in linguistics and communication sciences for several decades, computer vision and multimedia research remained on the surface of the problem more or less. An exception is previous work that introduced the two metrics *Cross-Modal Mutual Information* and *Semantic Correlation* in order to model complex image-text relations. In this paper, we motivate the necessity of an additional metric called *Status* in order to cover complex image-text relations more completely. This set of metrics enables us to derive a novel categorization of eight semantic image-text classes based on three dimensions. In addition, we demonstrate how to automatically gather and augment a dataset for these classes from the Web. Further, we present a deep learning system to automatically predict either of the three metrics, as well as a system to directly predict the eight image-text classes. Experimental results show the feasibility of the approach, whereby the predict-all approach outperforms the cascaded approach of the metric classifiers.

My Contributions: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing – original draft, Writing – review & editing

The following list shows additional publications, which are only partially related to the topic, and will therefore not be covered in this thesis:

[197] Christian Otto, Sebastian Holzki, and Ralph Ewerth. "Is This an Example Image? - Predicting the Relative Abstractness Level of Image and Text". In: *Advances in Information Retrieval - 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14-18, 2019, Proceedings, Part I*. vol. 11437. Lecture Notes in Computer Science. Springer, 2019, pp. 711–725. DOI: 10.1007/978-3-030-15712-8_46

[117] Johannes von Hoyer, Anett Hoppe, Yvonne Kammerer, Christian Otto, Georg Pardi, Markus Rokicki, Ran Yu, Stefan Dietze, Ralph Ewerth, and Peter Holtz. "The Search as Learning Spaceship: Toward a Comprehensive Model of Psychological and Technological Facets of Search as Learning". In: *Frontiers in Psychology* 13 (Mar. 2022). DOI: 10.3389/fpsyg.2022.827748

[70] Ralph Ewerth, Christian Otto, and Eric Müller-Budack. "Computational Approaches for the Interpretation of Image-Text Relations". In: Oct. 2021, pp. 109–138. ISBN: 9783110725001. DOI: 10.1515/9783110725001-005

[188] Markus Mühling, Nikolaus Korfhage, Eric Müller, Christian Otto, Matthias Springstein, Thomas Langelage, Uli Veith, Ralph Ewerth, and Bernd Freisleben. "Deep learning for content-based video retrieval in film and television production". In: vol. 76. 21. 2017, pp. 22169–22194. DOI: 10.1007/s11042-017-4962-9

[189] Eric Müller, Christian Otto, and Ralph Ewerth. "Semi-supervised Identification of Rarely Appearing Persons in Video by Correcting Weak Labels". In: *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, ICMR 2016, New York, New York, USA, June 6-9, 2016*. ACM, 2016, pp. 381–384. DOI: 10.1145/2911996.2912073

1.6 Organization of this Thesis

Chapter 1 introduces the aspects and challenges associated with multimedia learning that are addressed in this thesis. It defines problems and states research questions that we will answer. **Chapter 2** covers the fundamentals of the key techniques utilized in our methodologies. That entails neural network basics and their relevant applications for uni- and multimodal representations and a selection of classifiers we utilized. **Chapter 3** introduces multiple ways of exploiting textual features to improve educational search engines based on previously extracted text-based metadata. Following up in **Chapter 4**, we contribute two user studies that resemble learning in formal and informal settings to reveal interesting connections between consumed learning resources and knowledge gain. In particular, our workflow entails a comprehensive feature extraction process covering individual methods for audio, visual, audio-visual, textual, and behavioral features, followed by an extensive correlation analysis. Afterward, **Chapter 5** assesses this topic from a different angle. In reality, creators of multimodal content intend their information to be understood in unison, meaning "Which message does, for example, image and text convey *together*? ", rather than individually. So, this final chapter models the cross-modal interplay on a theoretical level to foster future research in the automatic understanding of multimodal content. We propose a categorization of semantic image-text classes derived from communication science and extend with recent proposals of computable image-text metrics from computer science. Finally, we demonstrate the utility of these metrics by evaluating their applicability on two unseen datasets. Finally, **Chapter 6** summarizes the various topics covered by this thesis and consolidates the findings, outlines limitations, and derives avenues for future work.

◇ ◇ ◇

The next Chapter introduces a the most important foundations of the methods and algorithms covered in this thesis.

2 Foundations

2.1 Introduction

This chapter presents a number of approaches and techniques we will leverage in the upcoming chapters. Understanding the foundations of these methods will complement the motivations and design choices in the following methodologies. First, we take a close look at neural networks in general before investigating the neural network-based methods that we use in Chapters 3, 4, and 5, namely semantic word embeddings and autoencoders. Second, this chapter introduces the classification approaches utilized for KG prediction in Chapter 4, namely Random Forest (RF), Naive Bayes (NB), Support Vector Machines (SVMs), and Multilayer Perceptrons (MLPs).

Finally, an introduction to research in the area of the visual-verbal divide as proposed by researchers in communication and media science is given by means of four classification systems that allow for the differentiation of different types of image-text pairs. We end this chapter with a discussion about the limitations of these systems that make a direct adoption from a computer science perspective difficult. In other words, we shed light on the requirements for a categorization of image-text classes that allow us to detect these intricate, semantic relations automatically.

2.2 Neural Networks

This section gives an introduction to artificial neural networks (Section 2.2.1), explains their components and general mechanics and provides a superficial look at the underlying calculus. Afterward, we give a more detailed description of the techniques used in this thesis. This entails semantic word embeddings (Section 2.2.6) and autoencoders (Section 2.2.5).

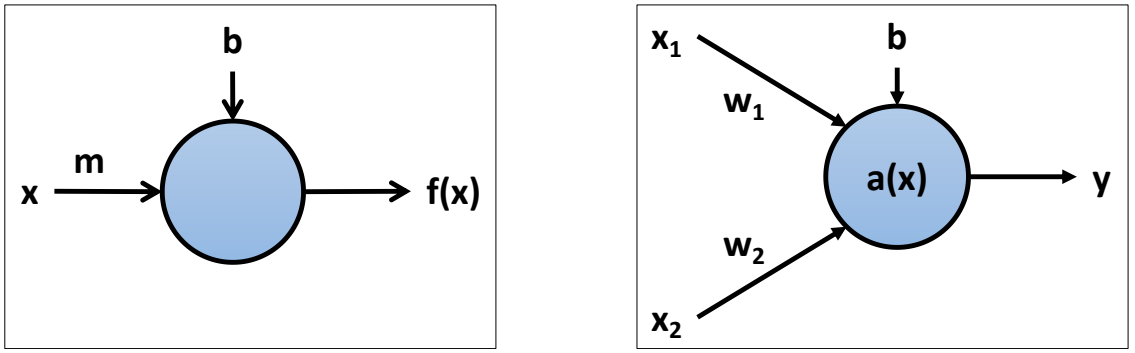
2.2.1 Foundations

As the name suggests, artificial neural networks (NN) consist of artificial neurons, which are inspired by biological neurons, one of the fundamental units of the brain. On their own, they are simple entities that receive inputs, process them, and, eventually, provide an output. In machine learning, these networks can be utilized as universal function approximators. That means they are able to approximate any problem that can be represented by a function, regardless of its complexity. Moreover, the past decade of research has proven that neural networks can be applied to a large variety of real-world

problems, not rarely exceeding human performance [71]. Their ability to do so is mainly attributed to a) their hierarchical structure that allows for the abstraction of tasks similar to the human problem-solving process ("This object has four wheels, windows, and a steering wheel. It must be a car! ") and b) their independence from manually crafting features for each individual task, which was the dominant approach until the advent of *Deep Learning*. The term *Deep Learning* comes from the flexible design of these networks that allows intricate and, therefore, often deep (concerning the number of layers) architectures that are tailored to the problem at hand.

Neurons

The integral building blocks of neural networks are, as stated before, neurons. In their simplest form neurons resemble a linear function $f(x) = w \cdot x + b$, where x is the input, w a weight applied to x, and b a bias value, see Figure 2.1a.



(A) The simple form of a neuron resembling a linear function.

(B) The advanced version of a neuron used in neural networks.

FIGURE 2.1: Two versions of a neuron.

Since that is not very useful yet beyond linear problems, NNs extend this concept in multiple ways, see Figure 2.1b. First, similar to the brain, a neuron is not limited to one input. Instead the number of inputs $X = x_1, x_2, ..., x_n$ is variable, and each input comes with its own weight $W = w_1, w_2, ..., w_3$. The value of the resulting formula is computed by the weighted sum of the inputs plus the bias, see Equation 2.1.

$$y = a(\sum_{i=1}^{|X|} w_i \cdot x_i + b) \tag{2.1}$$

Second, an equally important difference is the choice of the non-linear activation function $a(x)$, which is wrapped around this computation. The activation function allows the neural network to approximate non-linear functions. Linear algebra shows that, regardless of the number of layers and neurons, a neural network with a linear activation can be reduced to a 2-layer version of itself representing, again, a simple linear function $f(x) = w \cdot x + b$. Activation functions will be discussed in more detail in Section 2.2.3.

2.2.2 Types of Networks

With some exceptions (e.g., Gated Recurrent Units (GRUs) [42]), a typical neural network consists solely of neurons arranged into *layers*, hierarchical tiers starting from the input layer, over a variable amount of hidden layers, to the output layer. The desired input data determines the input layer's shape. For example, a visual concept classifier expecting a $30x30$ pixel, 4-channel image requires $30x30x4 = 3600$ input neurons. The output, on the other hand, is determined by the problem to solve. If we label two classes for our visual concept classifier, the neural network would have two output neurons, one for each class, as in Figure 2.2. So-called *feedforward networks* are then trained by *feeding* training samples *forward* through the network, meaning "from input to output without loops". As neural networks are generally supervised approaches, each sample has an associated label. At the end of a forward step, the predicted outcome is compared to the expected output by means of a cost function. This function, which again differs based on the given problem, returns a value representing the difference between the ground truth and prediction value. By propagating this *loss* backward through the neural network, adjusting weights and biases of the neurons along the way, the potential error of the network is reduced the next time it sees a similar sample. We go further into detail in Section 2.2.4.

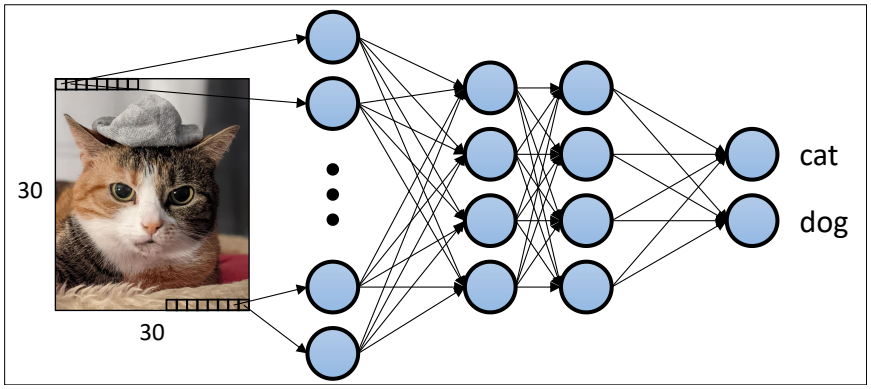


FIGURE 2.2: Simplified example of a neural network for visual concept classification. Recent versions, such as ResNext [287], are much more complex.

Research of the past decade, however, heavily focuses on the intermediate part, the hidden layers. Their design varies depending on the type of data a researcher is working with. Image processing, for instance, makes use of convolutional layers, which can be interpreted as a repeated application of a filter to regions of the original image to receive a feature map of a lower dimension, see Figure 2.3. These filters aim to pick up patterns of various complexities and take advantage of correlations in the image. Hidden layers at the beginning of the network identify low-level features such as lines and edges, while filters at the end detect complex structures such as eyes or entire faces. The most important advantage of these approaches, as compared, for instance, to a Sobel filter [132], is that these filters are automatically learned during training and do not need to be designed manually. After applying the filter and adding the bias, the neural network

passes the values of the feature map, again, through an activation function. Lastly, to further summarize the information of the feature map, a *pooling* operation is applied. Most commonly a *MaxPooling* operation, as shown in Figure 2.3. It only retains the maximum element of a certain region of the feature map and thus, depending on the size of the pooling kernel, reduces the dimensionality of the input even further. Current state-of-the-art approaches are ConvNeXt [157], EfficientNet [256], MobileNetV2 [228], and ResNeXt [287].

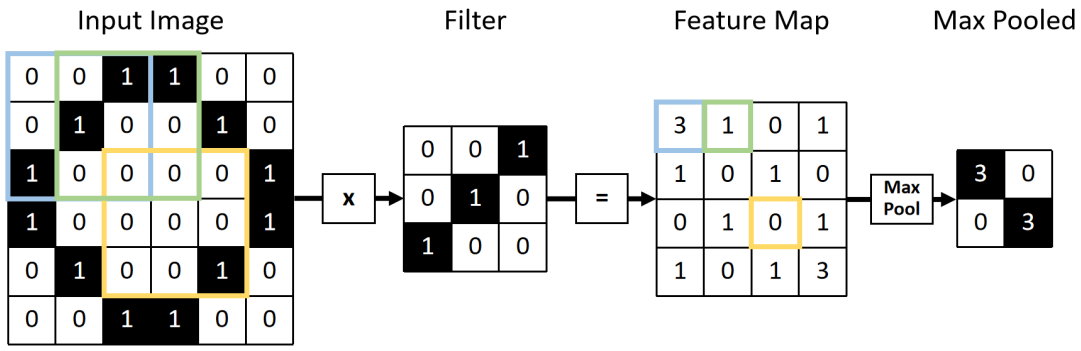


FIGURE 2.3: The general idea of convolutional layers for image encoding. A filter is convolved with the input image to reduce its dimensionality. After pooling the dimensions of the original image are reduced from 36 to 4 while the most important information, with regards to the filter kernel, are preserved.

Sequential data, meaning data with an intrinsic order (e.g., video frames, sentences in a text, temporal data), utilize *Recurrent Neural Networks (RNN)* which can store memory about previously seen tokens of the current sample to make meaningful connections between these tokens. To do this, they contain loops referencing previous parts of the encoding procedure (e.g., the network), which means they are not feedforward networks as discussed up until now. Or in other words, information from previous tokens of the input sequence influence how the current token will be encoded. In practice, this may help give ambiguous words context by allowing the network to consider the entire sentence when encoding, for example, the word "beat" which has multiple meanings: e.g., overcoming a high score, hitting someone, or the noun describing the basic unit of time in a song.

One building block for RNNs, which will also be utilized in Section 5.3.7, is the *GRU* [42]. They are a simpler version of Long-Short-Term-Memory Cells (LSTMs [111]) that is, however, better able to deal with vanishing gradients during training (cf. Section 2.2.3). Simpler because they have only two gates instead of three and, thus, fewer parameters. Gates are the main difference between a GRU and a typical neuron, enabling the network to memorize things it has seen before. In particular, GRUs have an *Update* gate Z_t and a *Reset* gate R_t, which, as their names imply, determine how the memory of the current unit is altered given a certain input. Inputs for GRUs are, besides the current token of the sequence of our input data X_t, the hidden state H_{t-1} of the unit that encoded the previous token in our data, see Figure 2.4. At time step t, both X_t and H_{t-1} have to pass through

the *Reset Gate* where the GRU decides whether to retain the information in H_{t-1} or discard it, In other words, it *resets* what the network has learned so far, see Equation 2.2. This decision is influenced by the activation function a, the weight matrices W_{xr}, W_{hr}, and the bias b_r.

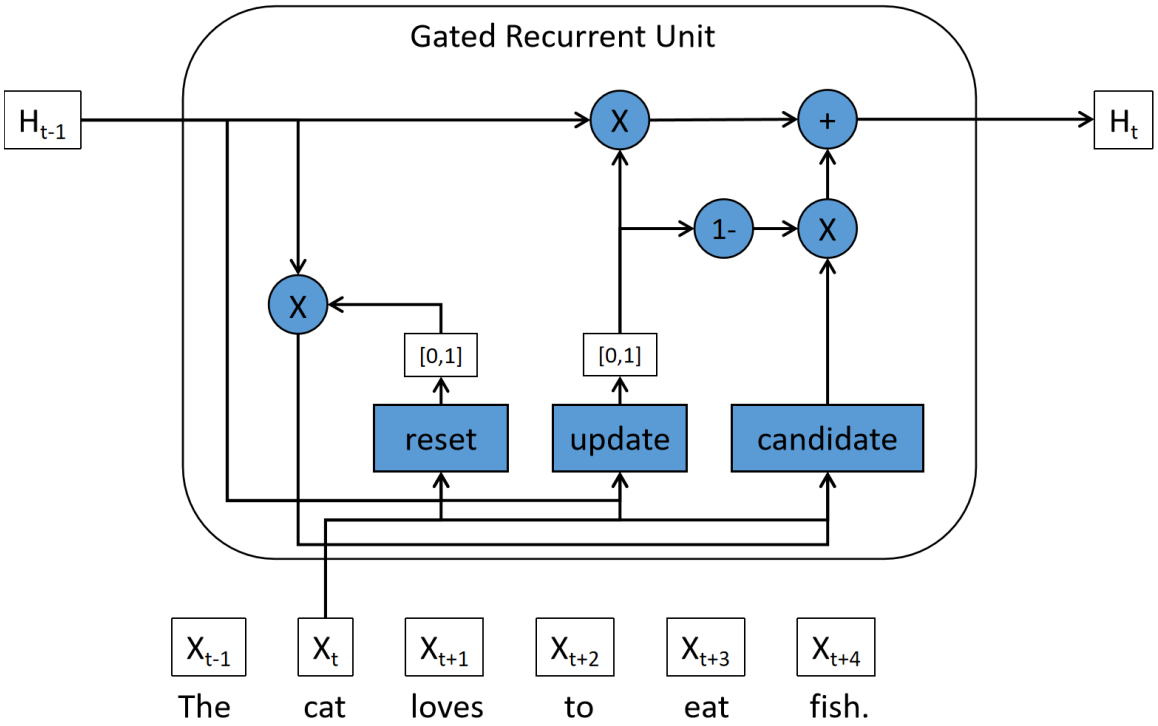


FIGURE 2.4: Workflow of a GRU. Weights and biases are omitted for clarity.

$$R_t = \sigma(X_t \cdot W_{xr} + H_{t-1} W_{hr} + b_r) \tag{2.2}$$

2.1: This equation computes a value between [0,1] to determine whether the current GRU will *reset* (i.e., forget) what has been learned by the network so far.

Next, based on the input X_t and the amount of information from H_{t-1} that passed through the Reset gate, a candidate hidden state $\tilde{H}_t$ is computed, see Equation 2.3. Its content can be interpreted as the amount of new information that is potentially added to the memory of the network given our current sequence token. If $R_t = 0$, only X_t will influence this candidate hidden state. Again, two trainable weight matrics (W_{xh} and W_{hh}), a bias b_h, and a sigmoid function are part of the equation.

$$\tilde{H}_t = a(X_t \cdot W_{xh} + (R_t \bigodot H_{t-1}) \cdot W_{hh} + b_h) \tag{2.3}$$

2.2: This equation computes the candidate hidden state depending on the input X_t and the information retained from H_{t-1}, which is decided by R_t.

Finally, the *Update* gate Z_t determines to what extent the candidate hidden state (or the new information given by the current input X_t) influences the already existing memory from all the previously seen tokens H_{t-1}. Similar to the Reset gate, due to the sigmoid activation, a value between 0 and 1 is calculated, see Equation 2.4. This value enables the computation of the output of the GRU, which is a weighted combination of the previous memory H_{t-1} and the candidate hidden state $\tilde{H}_t$ according to Equation 2.5.

$$Z_t = \sigma(X_t W_{xz} + H_{t-1} W_{hz} + b_z) \tag{2.4}$$

2.3: This equation computes the a value between [0,1] to determine to which extent *new* information influences the already established memory of the network.

$$H_t = Z_t \bigodot H_{t-1} + (1 - Z_t) \bigodot \tilde{H}_t \tag{2.5}$$

2.4: This equation computes output of the GRU combining the previous memory H_{t-1} and the information from the current token $\tilde{H}_t$ according to the value of the *Update* gate Z_t.

In the experiments in Section 4 we utilize a bidirectional GRU to encode our textual inputs forwards and backward at the same time and concatenate the outputs. Thus, giving the network two perspectives on the sentence(s) to capture their semantics even better. State-of-the-art approaches for sequential data-based tasks, such as machine translation, image captioning, or question answering are, for instance, Sentence-BERT [220] and MPNet [246].

2.2.3 Activation Functions

The *sigmoid* function σ was one of the first popular approaches for the activation function for multiple reasons. It satisfies all requirements for an activation function, since it is

- monotonically increasing
- defined everywhere
- continuous
- and differentiable in $\mathbb{R}$.

As a bonus, the derivative required for back-propagation (cf. Section 2.2.4) of σ is simply $\sigma'(x) = \sigma(x)(1 - \sigma(x))$. It fell out of popularity because of the *vanishing gradient problem* [110]. Neural networks learn by applying gradient descent [137] to the results of their cost functions E after passing a training sample through the network. That means they compute the derivative E' and utilize it to adjust the weights and biases in the network. However, *sigmoid* has a meaningful gradient only close to 0 or, in other words, very large or tiny input values return a gradient close to 0 (the gradient "vanished"). The impact

these small values have on the training diminishes the further we go backward through the hidden layers during back-propagation. This leads to the network getting stuck and being unable to find a useful solution during training. The *vanishing gradient problem* also occurs for the *tanh* activation function which is related to *sigmoid* by $tanh(x) = 2\sigma(2x) - 1$.

An even simpler function that does not suffer from vanishing gradients and is therefore commonly used in Deep Learning is the Rectified Linear Unit (ReLU)[191] defined as $a(x) = max\{0, x\}$. ReLUs are nearly linear. That means they preserve many properties that make linear models easy to optimize with gradient-based methods. Even though they are not continuous at $x = 0$, the gradient can still be defined as either 0 or 1 without introducing too much error. While this original design works sufficiently well in general, large derivatives during back-propagation can cause it to get *stuck* returning 0 forever, also called the *dying ReLU problem* [159]. This happens especially during the beginning of the training, where high learning rates cause significant weight swings. To circumvent this problem variations such as *LeakyReLU* [162] or the *ELU* (exponential linear unit) [45] have been introduced, that return negative values for inputs < 0, see Figure 2.5.

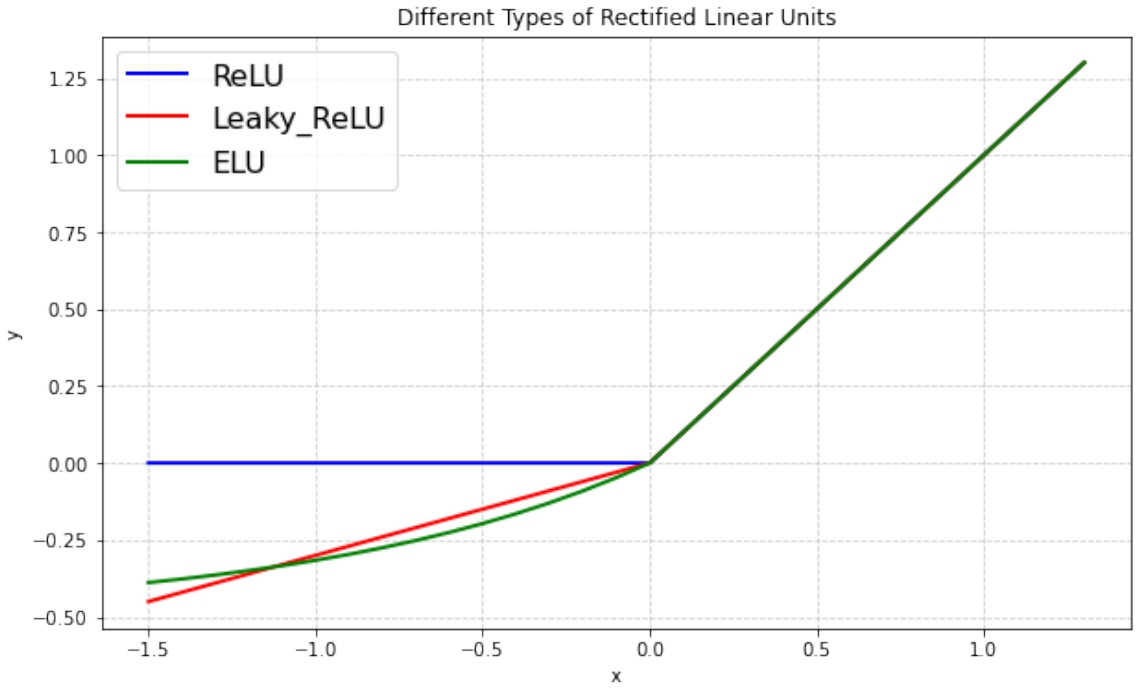


FIGURE 2.5: Different Types of Rectified Linear Units.

2.2.4 Back-propagation and Optimization

As briefly introduced in Section 2.2.2, back-propagation [225] is the first key component of the process responsible for the network learning from its errors during training. It stands for "backward propagation of errors", and with the second key component, an optimization function like *gradient descent* optimizes the networks' ability to solve a given task. It does so by calculating a gradient of the cost function after each sample of the

training dataset, or, if that is not feasible due to hardware restrictions, each sample of the current batch (which equals a shuffled subset of the training data) has passed through the network. The optimizer then decides, according to the chosen learning rate lr, to which extent this gradient will influence the weights $w_{i,j}^k$ and biases b_i^k of the network. These weights and biases are commonly summarized as the parameters $\theta = \{w_{i,j}^k, b_i^k\}$ of the network. Following the slope of the cost function step by step in the direction of this computed gradient reduces the error produced by the dataset predictions P compared to the ground-truth labels $\hat{Y}$. For this explanation, we consider a simple yet commonly used cost function: the mean of the squared errors (MSE), see Equation 2.6.

$$MSE(P, \hat{Y}) = \frac{1}{n} \sum_{i=1}^{n} (P_i - \hat{Y}_i)^2 \tag{2.6}$$

2.5: An example of the mean squared error cost function that averages the squared differences between the predictions P and the ground-truth labels $\hat{Y}$ for each sample of the dataset (or batch).

Mathematically, back-propagation determines the rate of change to our cost function E given an adjustment of the parameters θ. With this, our neural network updates all parameters according to the following Equation 2.7:

$$\theta_{new} = \theta_{old} - learning\_rate \cdot \frac{\partial E}{\partial \theta} \tag{2.7}$$

2.6: Calculation of a parameter update (weights and biases) of a neural network according to the learning rate and the computed gradient.

The old parameters θ_{old} are updated by subtracting (since we want to minimize the cost function) the computed gradient times the learning rate, which is typically a value < 1, e.g., 0.0001. This step is necessary since large gradients can cause the network's performance to drop, because it fails to converge towards a minimum, also known as the *exploding gradient problem*. Computing the gradient in Equation 2.7 requires us to determine and average the rate of change of the cost function considering each individual weight and bias in the network. Luckily, this process can be significantly simplified due to the chain rule in calculus. This chain rule describes how the gradient of a nested function can be computed as the multiple of the derivatives of the nested functions, see Equation 2.8. Calculating each derivative for each weight and bias is significantly sped up by reusing the terms of these chains that form the derivatives.

$$\frac{\partial}{\partial x} f(g(h(x))) = h'(x) \cdot g'(h) \cdot f'(g) = \frac{\partial h}{\partial x} \frac{\partial g}{\partial h} \frac{\partial f}{\partial g} \tag{2.8}$$

2.7: Chain rule in calculus.

Computing the gradient of E follows a similar pattern for two reasons: First, the output of a neuron itself is a nested function (cf. Equation 2.1) since the weighted sum

$z_i = \sum_{i=1}^{|X|} w_i \cdot x_i + b$ is input for the activation function a. Second, considering were are not in the input layer, x_i resembles the output of the activation functions of the neurons in the previous layer, which can be substituted by their nested inputs again. In other words, starting with the output layer, each layer is a function of the activations of its predecessors. Therefore, the entire neural network resembles a nested function as well. In a network with L layers that uses the mean squared error cost function (cf. Equation 2.6) the derivative for a weight w_{jk}^l between neuron k of layer $l-1$ and neuron j of layer l, for instance, is computed as follows:

$$\frac{\partial E}{\partial w_{jk}^l} = a_k^{l-1} \cdot a'(z_j^l) \cdot \frac{\partial E}{\partial a_j^l} \tag{2.9}$$

EITHER: iteratively replace last term with next layer

$$\frac{\partial E}{\partial w_{jk}^l} = a_k^{l-1} \cdot a'(z_j^l) \cdot \sum_{j=0}^{n_{l+1}-1} w_{jk}^{l+1} a'(z_j^{l+1}) \cdot \frac{\partial E}{\partial a_j^{l+1}} \tag{2.10}$$

OR: if last layer

$$\frac{\partial E}{\partial w_{jk}^l} = a_k^{l-1} \cdot a'(z_j^l) \cdot 2 \cdot (a_j^L - \hat{y}) \tag{2.11}$$

2.8: Computation of the influence weight w_{jk}^l has on the cost function. The result of this equation in conjunction with the learning rate is used to update this particular weight.

Verbatim, these equations can be understood as: the influence weight w_{jk}^l has on the cost function E is determined by the outputs of neuron k in layer $l-1$ called a_k^{l-1}, the derivative of the activation function of the neuron it is connected to with respected to the weighted sum, called $a'(z_l^j)$, and, the influence of that same activation function on the cost function E. Afterward, this last term is then substituted for either the derivative of all neurons of the next layer connected to our target neuron or, if we are already in the output layer, simply the derivative of the cost function.

2.2.5 Autoencoders

Autoencoders are a special kind of neural network whose goal is to learn useful representations for any type of input in an unsupervised manner, meaning without the need for labeled training samples. They achieve this by a symmetric architecture that consists of 1) an encoding pipeline that reduces the dimensionality of the input down to the desired level, 2) the hidden embedding layer in the "middle" that, optimally, retains only the most salient information about the input sample, and 3) the decoding pipeline that attempts to reconstruct the input sample as similar as possible based on this embedding. Figure 2.6 outlines these basic components.

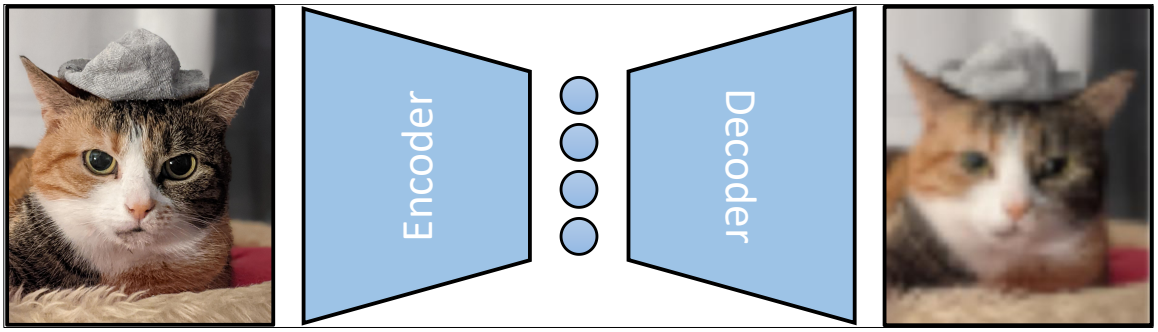


FIGURE 2.6: Simplistic structure of an autoencoder. Detailed architecture of the encoder and decoder component depends on the modality to be encoded.

This functionality has multiple practical applications, which will be briefly outlined next. Early applications utilized autoencoder for denoising all types of imagery [41]. By encoding noisy images and thus reducing them to their key information, the noise will be dismissed, and the reconstruction returns a clean(er) image. Another field of applications are compression algorithms. For most use cases, autoencoder-based compression methods can have disadvantages when compared to traditional algorithms such as JPEG2000 [257] or BPG [26] in terms of general applicability, the requirement for training samples, and compression performance. However, recent state-of-the-art approaches [43, 87] showed comparable performances for the error introduced by compression measure by the peak signal-to-noise ratio (PSNR) as well as the more perceptual MS-SSIM [280] metric that describes the structural similarity of the output image compared to the original. For the purpose of neural network-based machine learning, autoencoders are mainly utilized in two ways. Due to their capabilities of learning salient representations in a self-supervised manner, the required number of labeled samples for an application can be reduced significantly. For example, a classification network, as in Figure 2.2 has to learn how to encode an image *and* how to predict the desired output. With an autoencoder, this representation could be learned beforehand with a much larger number of samples. Then, the decoder part of the network would be replaced with a set of fully-connected layers for classification that will be trained with the labeled samples. Finally, by introducing a probability distribution to the decoding component of the autoencoder, visually similar variants of the input image can be generated as additional training samples. These types of autoencoders are called *variational autoencoders*. Finally, a third, very common usecase for autoencoder is shown in the next section.

2.2.6 Semantic Word Embeddings

Semantic word embeddings are an invaluable building block for modern natural language applications. With them, computers are able to represent not only text representations but also their semantic meaning. This ability enables algorithms to

1. determine whether text passages (words, sentences, paragraphs) are semantically similar or not
2. align visual and textual encodings to describe their relation to one another (cf. Chapter 5)
3. detect antonyms or positive and negative connotations
4. model ties between certain word, e.g., capital - country (cf. Figure 2.7)

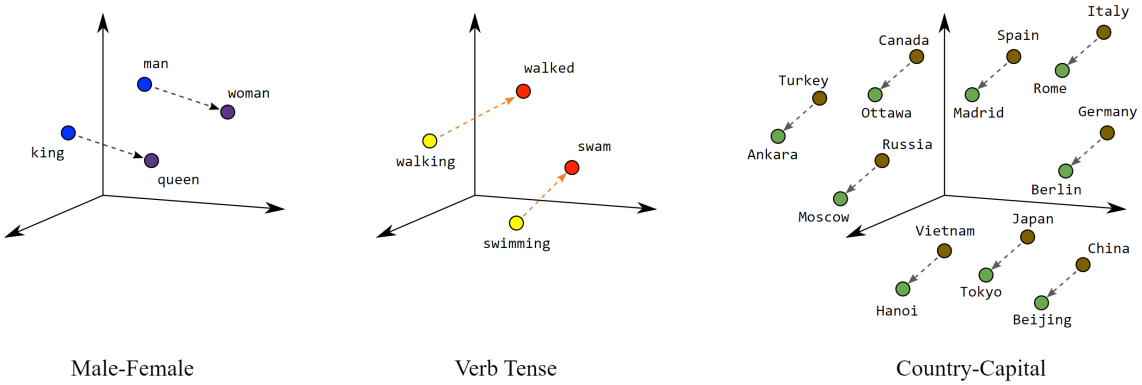


FIGURE 2.7: Example of the relationships semantic word embeddings are able to infer.
(Source: https://developers.google.com/machine-learning/crash-course/images/linear-relationships.svg)

The process of establishing a semantic word embedding relies on the assumption that words that appear in similar contexts have similar meanings [95, 129]. Modern approaches exploit this fact which allows them to convert words into high-dimensional vector representations in a self-supervised manner. The variance and quality of the trained model relies, besides the architecture of the neural network, solely on the chosen text corpora. This is convenient to align a model to a given task or finetune a model without additional, manual labeling. The trick in generating a semantic word embedding is to encode, in addition to the so-called *focus word*, a certain amount of context, and thus, going beyond simple representation encoding. In a pioneering work, Mikolov et al. [182] introduced two approaches to embed contexts for their *word2vec* model called Continuous Bag of Words (CBOW) and Skip-gram (SG).

The quick brown fox jumps right over the lazy dog.

FIGURE 2.8: CBOW and SG consider c=2 context words (blue) on each side of the focus word (green).

As exemplary shown in Figure 2.8, given a context parameter c, CBOW and SG consider a variable amount of context words for each focus word they encode. They differ in the way they try to reconstruct texts: CBOW tries to estimate the focus word given the context, and SG predicts the context words given the focus word. The algorithm is self-supervised because the training process resembles an autoencoder with just one input,

hidden, and output layer. The input and output layer have the dimension of the desired vocabulary, representing the word to be encoded in the input (one-hot encoded), while the output is the probability of each word in the vocabulary appearing close to the input word. The hidden layer has the dimensionality of the desired semantic word embedding. After training, the output layer is dismissed, and the hidden layer is used for further experiments since it returns the desired embeddings. Even though the number of layers is small, the number of parameters is high due to the size of the input and output layers. Moreover, given the nature of the problem, the sample size can exceed billions of samples, see Figure 2.9.

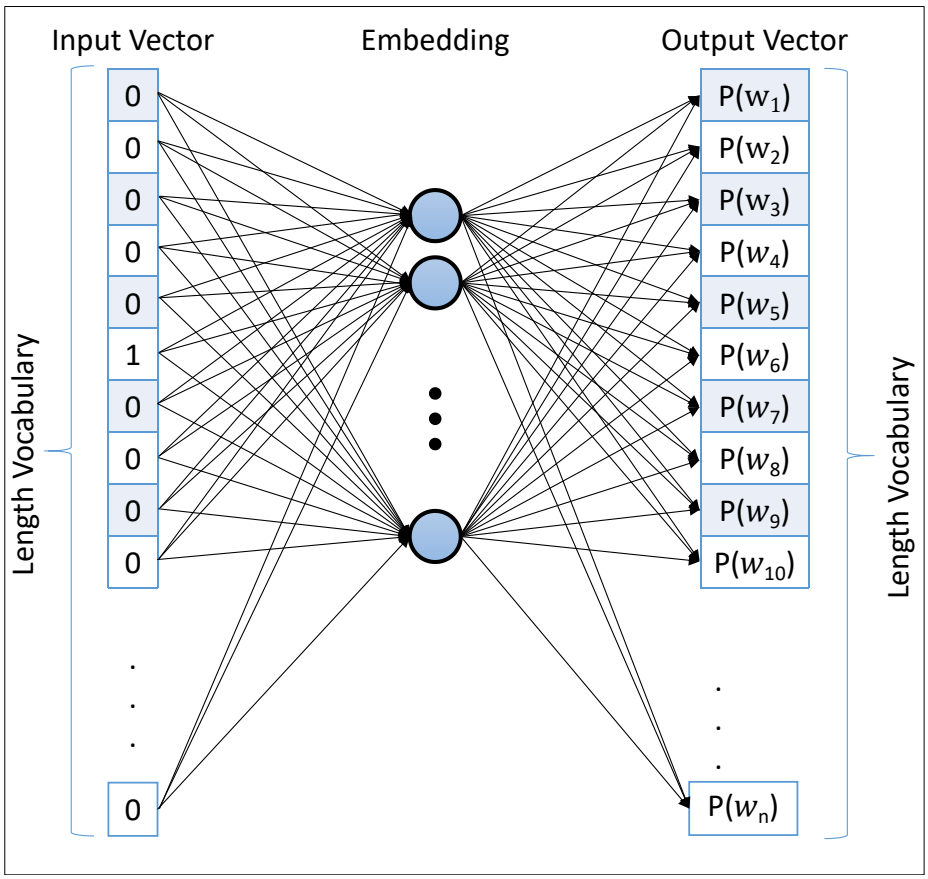


FIGURE 2.9: Word2vec architecture. Input is the one-hot encoded vector of the focus word and the output vector the probabilities of each other word in the vocabulary appearing close to it. During training, the expected label is one-hot encoded as well with a context word that appeared close to the focus word in the underlying dataset.

That means, given a random initialization of the word embeddings, one training step entails the relocation of *all* word vectors in a way that the input focus and context words move closer to each other and the rest moves further away. This computationally heavy process has been improved in the same paper by Mikolov et al. [182]. By subsampling words, which means removing them from an input sentence with a probability proportional to their frequency in the training data, the number of training samples is decreased significantly. Words like *the* or *a* are removed, which a) improves the training time and

b) improves the quality of the sentence embedding in the end since they are not adding valuable context information.

A second method to drastically improve training time is negative sampling. As mentioned, for each training step, the respective ground-truth label of a sample is a one-hot encoded vector the length of the vocabulary, where only 1 is a context word to the current focus word. That means all other bins of the vectors are 0, forcing the neural network to update *all* weights of the model. For negative sampling, the authors suggest selecting only $5 - 20$ bins that are 0 according to a unigram distribution, which chooses more frequent words more often. Only updating the weights associated with this selection of negative words reduces the number of weights to be changed tremendously, depending on the vocabulary size.

In 2016, Bojanowski et al. [22] presented another semantic word embedding approach called *fastText*. As it will be used in multiple sections of this thesis, we will go further into detail about how it works. It extends *word2vec* by a key concept: it encodes each word as a bag of n-grams, which allows the model to harness subword information. Consequently, the model can return valuable encodings for rare words, or even words it has never seen during training but whose parts are similar to other known words. This is especially potent for languages with many compound words, like German. This is another reason fastText is used in this thesis since the experiments in Chapter 3 are based on a german dataset. Additionally, the authors utilize angular brackets to denote the start and end of a word, see Figure 2.10. Lastly, they append the entire word to the bag of n-grams as well. With this, shorter words like <*her*> can be distinguished from words they appear in, like <*gather*>.

fastText(library) = <li, lib, ibr, bra, rar, ary, ry>, <library>

FIGURE 2.10: Representation of the word *library* in fastText using 3-grams. The angular brackets denote the start and end of a word.

2.3 Classifiers

This section gives detailed descriptions of four machine learning classification approaches that are utilized in Chapter 4 in the context of knowledge gain prediction based on a set of input features. To make these explanations more coherent within this context, the following definitions will use this terminology of features as input and knowledge gain as output.

2.3.1 Random Forest

This classifier is a popular and reliable supervised learning algorithm. It is an ensemble approach combining multiple decision trees to partially mitigate their individual drawbacks

and generate a better prediction result. Random Forests can be utilized for regression and classification problems. To understand the prediction process, we first need a definition of decision trees.

Similar to RFs, Decision Trees are also supervised and, in addition, non-parametric. Their goal is to learn simple if-then-else rules to separate and classify a given dataset effectively. How well a rule separates the data is measured as *information gain*, or in other words, the reduction of impurity given by separating a dataset by this rule. How *impure* a dataset is, is measured by its entropy H, see Equation 2.12.

$$H = -\sum p(x)\, log\, p(x) \tag{2.12}$$

A simple example tailored to our use case would be a dataset that contains four learner samples, with two achieving a *high* knowledge gain, while the others achieved a *moderate* and *low* result. Also, one numerical feature $time_{edu}$ was recorded that measures the time spent on educational websites compared to the overall study time. In the beginning, the entropy or impurity of the dataset is

$$H_{input} = -(0.5 \cdot log(0.5)) - (0.25 \cdot log(0.25)) - (0.25 \cdot log(0.25)) = 0.45 \tag{2.13}$$

2.9: Entropy of the input dataset. Label *high* constitutes 50% of the samples while the other two occur only 25% of the time.

Separating the four samples by a rule that divides the participants by, for instance, whether they spent more than 15 minutes on educational websites would separate the data labels into two subsets: *high, high* and *moderate, low*. The resulting information gain is a result of the initial entropy minus the sum of the weighted entropies of the two subsets S:

$$\begin{aligned} InformationGain &= H_{input} - \sum_{i=1}^{|S|} \left(\frac{|S_i|}{|input|} \cdot H(S_i)\right) \\ InformationGain &= 0.45 - \left(\frac{2}{4} \cdot 0 + \frac{2}{4} \cdot 0.3\right) = 0.3 \end{aligned} \tag{2.14}$$

Whichever rule yields the highest information gain is set to be the first if-then branch in the tree. In our case, for a second step, another rule could be established separating the *low* and *moderate* class leading to an optimal result for this trivial example. Alternatively, by defining a minimum requirement for the impurity of the tree, we could stop early, saving computational time.

The simplicity of decision trees bears some disadvantages. They tend to overfit on data with a large number of features, are sensitive to outliers, skew the results towards dominant classes in biased datasets, and are not guaranteed to produce optimal results due to the large potential number of possible rules. RFs mitigate some of these effects by creating multiple decision trees for the same task, evaluating all of them for a given test sample,

and returning the most voted one as the prediction. To be more robust toward outliers, the classifier considers subsets of the original dataset to create its individual trees. Also, based on the training process, the voting result could be weighted according to the error rate of the respective tree. Additional parameters are the number of trees to be generated, tree-depth, or the minimum number of samples per tree node to continue splitting.

2.3.2 Naive Bayes

NB is a classifier that is known for its easy implementation and inference time paired with decent results for specific tasks such as sentiment analysis [261] and spam filters [210]. It is based on Bayes' theorem (Equation 2.15) that computes the probability of an event based on prior knowledge about conditions related to the event. For our task, this translates to: by observing the joint occurrences of a feature together with a high knowledge gain, the classifier learns the probability of that event for each individual feature-knowledge gain combination during training. With this and the given features of a test sample, the classifier computes the probability of each output dimension, returning the highest one as the prediction.

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)} \tag{2.15}$$

The reason for it being called *Naive* are two assumptions the classifier makes about the input data: (1) all features are independent, meaning uncorrelated, and (2) all features have an equal effect on the target variable. They are, however, rarely true for real-world problems. Nonetheless, by ignoring these effects, the classifier predicts binary or multivariate problems as follows: We consider y as our knowledge gain variable with three dimensions (low, moderate, high) and X as a feature set with individual features $X = (x_1, x_2, ..., x_n)$. By substituting this into Equation 2.15 we get:

$$P(y|x_1, ..., x_n) = \frac{P(x_1|y)P(x_2|y)...P(x_n|y)P(y)}{P(x_1)P(x_2)...P(x_n)} \tag{2.16}$$

In Equation 2.16 the denominator is always static, and since we are not interested in the actual probabilities, just an overall ranking of scores, we can remove it. This step turns the equation into a proportionality. Consolidating the product in the numerator yields:

$$P(y|x_1, ..., x_n) \propto P(y) \prod_{i=1}^{n} P(x_i|y) \tag{2.17}$$

Here, $P(y)$ describes the prior knowledge we have about the probability of the knowledge gain dimensions. A common guess is the distribution of the classes in the training data, but it is possible to replace this value with more educated guesses with the goal of not skewing the final result. Lastly, to get a classification result for our input sample X we compute all three probabilities (one for each knowledge gain dimension), and the highest score is our guess.

$$y = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y) \tag{2.18}$$

To summarize: the Naive Bayes classifier computes the probability of each knowledge gain dimension by considering a priori knowledge about that dimension together with the likelihood of each individual input feature being present given that dimension. As stated above, training this classifier is fast. However, neither of the two assumptions stated above are true for our task of knowledge gain prediction. First, in an ensemble of, e.g., textual features, it is highly possible that two of them are correlated and, thus, not independent. Second, the premise of the experiments conducted in Chapter 4 is to determine the difference in importance of a given feature set with the knowledge gain. And as the results of the feature importance analysis will show, the impact of the individual features varies heavily.

2.3.3 Support Vector Machines

Another supervised learning technique are SVMs, suitable for linear and non-linear regression and classification problems. The core idea is to separate n-dimensional data into two classes with an (n-1)-dimensional hyperplane. For two-dimensional data, this translates to finding a line that separates the given classes optimally, see Figure 2.11. Optimal refers to the fact that even though there are infinite lines that separate linear separable classes, SVMs aim to predict the one solution that maximizes the margin (the "corridor") between the classes. The margin is defined as the minimal distance between a sample of the class and the found hyperplane (also called support vector), so a solution where both classes are equidistant to the nearest sample of each class is considered optimal. This approach is also called maximal margin classification, and it allows no misclassification, which makes it impractical for noisy data and outliers.

SVMs, however, are not maximal margin classifiers. They utilize a soft margin that allows misclassifications to find a better solution regarding margin size. For example, if the blue sample x in Figure 2.11 was considered for a support vector, the classifier performance would have increased only marginally while the margin's size had been roughly divided by four.

For data that is not linear separable, SVMs utilize the *Kernel Trick*. By purposefully adding one or more dimensions to the samples, it is possible to find a hyperplane in a higher dimension that allows for the separation of the data. Figure 2.12 gives an example that shows two classes that are not linearly separable (left). Adding a temporary third dimension $z = x^2 + y^2$ to each sample resembling the distance to the center of the coordinate system returns the middle image. As we are now in 3-dimensional space, a 2-dimensional plane is able to separate the data. Transforming the intersection between the plane and the 3-dimensional space returns the image on the right, again showing the optimal hyperplane, margin, and support vectors.

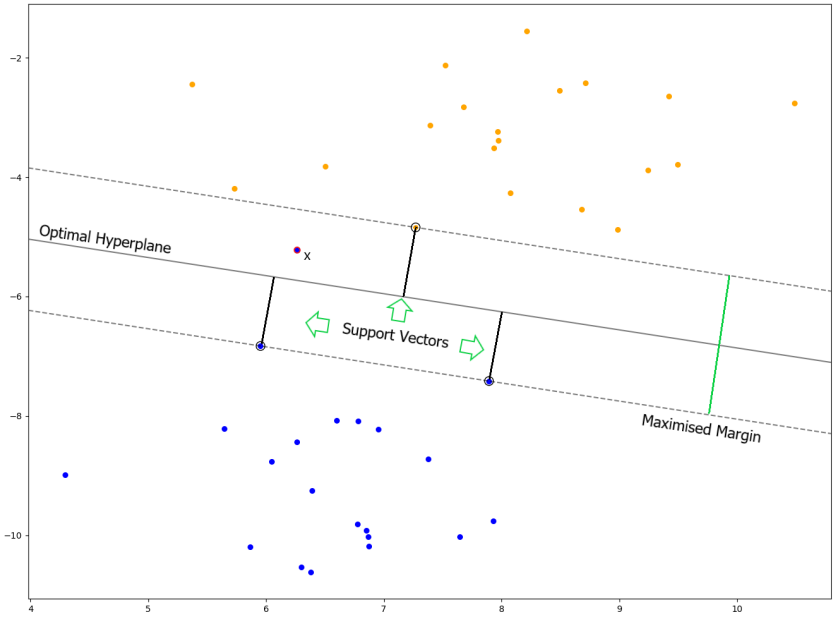


FIGURE 2.11: A hyperplane dividing two sets of two-dimensional points. Support vectors and margin "corridor" are visualized as well.

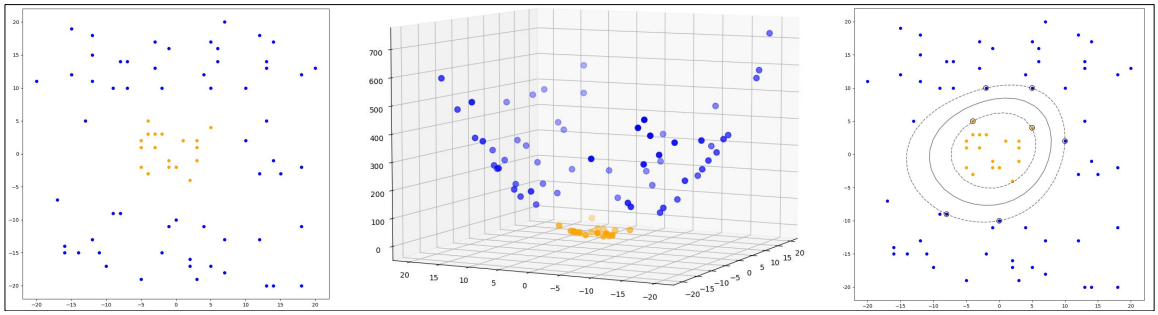


FIGURE 2.12: Separation of non-linear data by utilizing the *kernel trick*.

In practice, the training and, therefore, estimation of an optimal solution is NP--complete, entailing a time complexity of $O(N^3)$, where N is the number of samples. Consequently, it becomes inefficient for large datasets. One approach to solve this issue is *Sequential Minimal Optimization (SMO)*, proposed by Platt [214]. The core idea is to, instead of trying to find a solution considering all data points at once, only consider two variables simultaneously in an iterative manner, optimizing the solution step by step. Selecting these variables can be done by various heuristics, starting from random choice. Their explanation goes beyond the scope of this thesis, however.

2.4 Image-Text Taxonomies

As outlined in Section 1.4, Chapter 5 proposes an interdisciplinary approach to model semantic image-text relations. Interdisciplinary because we built the analysis upon research from communication and media sciences. In particular, we consider multiple approaches that categorize image-text pairs into meaningful taxonomies. This Section introduces, largely based on Bateman [16], some of the most impactful works and briefly discusses their advantages and limitations since we are just referring to parts of them later in the thesis.

For the earlier parts of the 20th century, monomodality (as in *text-only*) was the predominant approach for the analysis of meaning-making in linguistics [269]. In a pioneering work, Barthes [15] questioned this practice by arguing that, in order to deal with multimodal artifacts of everyday life such as advertisements and film, a mere textual view is not sufficient to describe the respective message(s). For example, one effect he mentions that had yet to be described is the 'floating' or 'vague' meaning of images. An image of, for instance, a baby with puffy cheeks eating a snack titled "my little sunshine enjoying his food" portrays an entirely different story when paired with the caption "Child obesity in country XY on an all-time high". Fixing the intended interpretation of the image by providing an appropriate caption is called **Anchorage** by Barthes, which is inherently important in media such as news. As the text is a mere tool to fixate the image's meaning, it is, according to Barthes, subordinate to the image, which shows their unequal relationship. Conversely, he also describes the (up to this point in time) traditional role of the image, namely providing a visual aid to the dominant text modality, as *Illustration*. That means the image *realizes* the text by showing a concrete instance of the entities or concepts described in the text. In this function, the image plays a subordinate role. Finally, there are also instances where image and text provide an equal amount of information to the overall message by the author called **Relay**. This relationship is characterized by the modalities complementing and subsequently depending on each other to make sense. The resulting classification of image-text relations is shown in Figure 2.13.

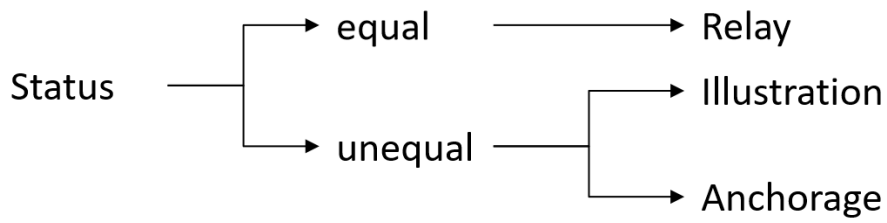


FIGURE 2.13: Image-Text class distinction by Barthes [15].

This categorization discretizes the importance of both modalities in arbitrary constellations. However, it only superficially talks about *how* information is conveyed. In 2005, Martinec and Salway [167] constructed their own classification system with the goal of being

able to assign each image-text pair to a distinct class. For this, the authors propose to describe each image-text pair based on two already developed dimensions. First, Barthes' distinction regarding the relative importance (called *Status*) as explained above and so-called *logicosemantic* relations, see Figure 2.14.

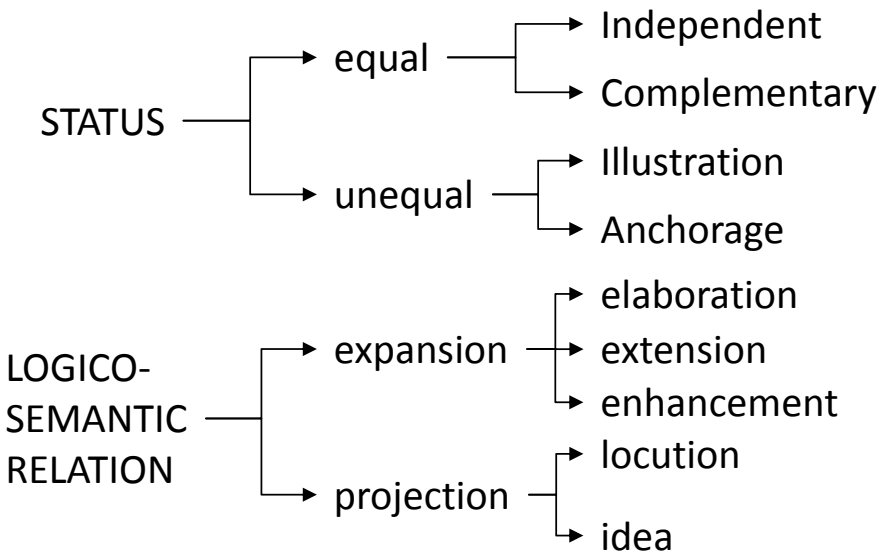


FIGURE 2.14: The image-text classification by Martinec and Salway [167] describing image-text pairs by means of the *Status* and *Logicosemantic* Relation.

First of all, Martinec and Salway extend Barthes' *Status* relation by arguing that image and text are also of equal importance when both modalities portray the same information and are therefore *independent*. However, as Henning and Ewerth [103] pointed out, it is debatable whether real independence occurs in practice since the two modalities are two different to portray *identical* information. The *logicosemantic* relations are based on work from Halliday and Matthiessen [94] who developed this system for relating clauses in English grammar but adopted it to relate images and texts. It distinguishes between whether the modalities expand each other or when the content that has been presented in one modality is re-represented in the other modality by means of *projection*.

According to [167], *Projection* mainly appears in two contexts: comic strips and labeled diagrams such as Venn diagrams or technical drawings. In other words, instances where the image itself contains text. The differentiate then between *Meaning*, where the content of one modality is given in a different form in the other modality, e.g., a diagram about the amount of rain per month in Berlin and the associated text which explains and interprets the diagram. Conversely, the *Idea* relationship is present when, according to Martinec and Salway, the text reports an approximate meaning. For instance, in comic strips when a character expresses a thought in form of a speech bubble about the current situation.

The different forms of *expansion* subsequently describe which type of information is added by the other modality. Considering an image of a woman in a suit, an *Elaboration* text would provide further details about this person, for example, her name and profession.

On the other hand, an *Extension* would provide additional information about this woman's actions: "The woman is leaving the building and walks to the nearest train station". Finally, *Enhancement* provides additional information about the situation's circumstances or environment, for example, "After receiving news about the bad quarterly earnings, the CEO leaves the building". All three of these relations describe similar situations, namely information being added by the opposite modality. As criticized by [16], due to a lack of annotation studies, an inter-coder agreement was not established that would prove how distinct these classes really are, and one could argue how it can be difficult to assign an image-text class to only one of Martinec and Salway's categories.

Unsworth [264] identified certain shortcomings of this approach with respect to educational materials. He argues that since Martinec and Salway [167] focus their attention on advertisements and online news, their systems lack the robustness required for a more generalized application. They state, however, that their work is the subject of ongoing work. Basically, as can be seen in Figure 2.15, Unsworth extends (and partially renames) the subcategories of *expansion* in order to fill in identified gaps in [167]'s classification system.

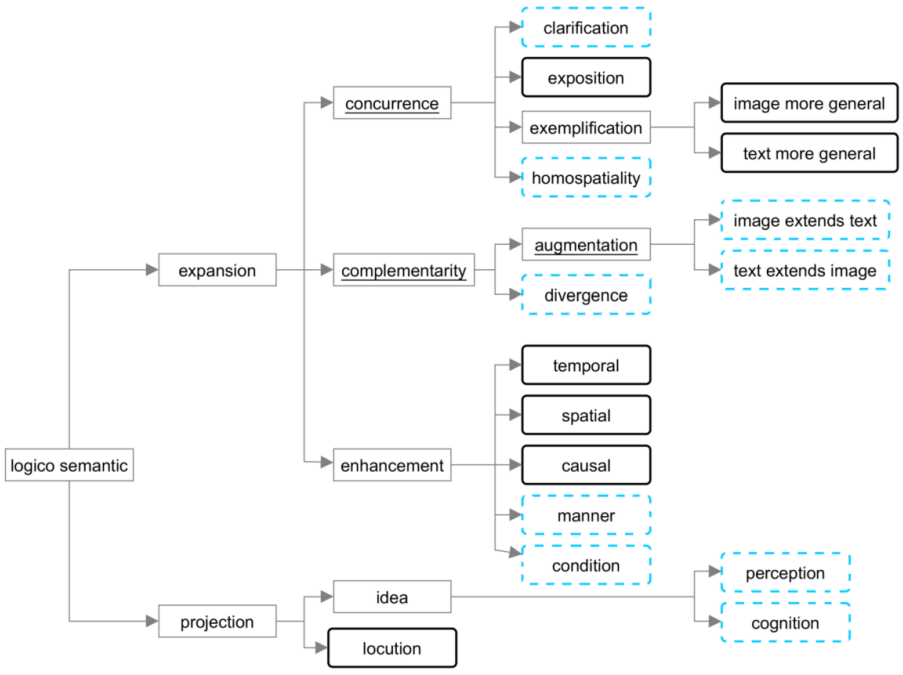


FIGURE 2.15: Unsworth's extension of Martinec and Salway's [167] system in blue dashed borders, while underlined classes were renamed, but kept their meaning.

Unsworth makes some significant additions to the work of Martinec and Salway for the context of this thesis. First, the *divergence* class under *complementarity* (former: extension), which was first considered by [140]. It describes circumstances where image and text 'pull in different directions, i.e., appear to convey incoherent messages, according to Bateman [16]. Inspired by [103] we talk in Section 5.3 about how these arrangements (intended or unintended) can be described by means of an image-text relation. Further,

Unsworth adds *Exemplification* to Martinec and Salway's *Elaboration*, which he calls *Concurrence*. Lastly, the relationship called *augmentation*, which differentiates between examples where an image adds new information to those given in the text and vice versa, will be considered in Section 5.3.

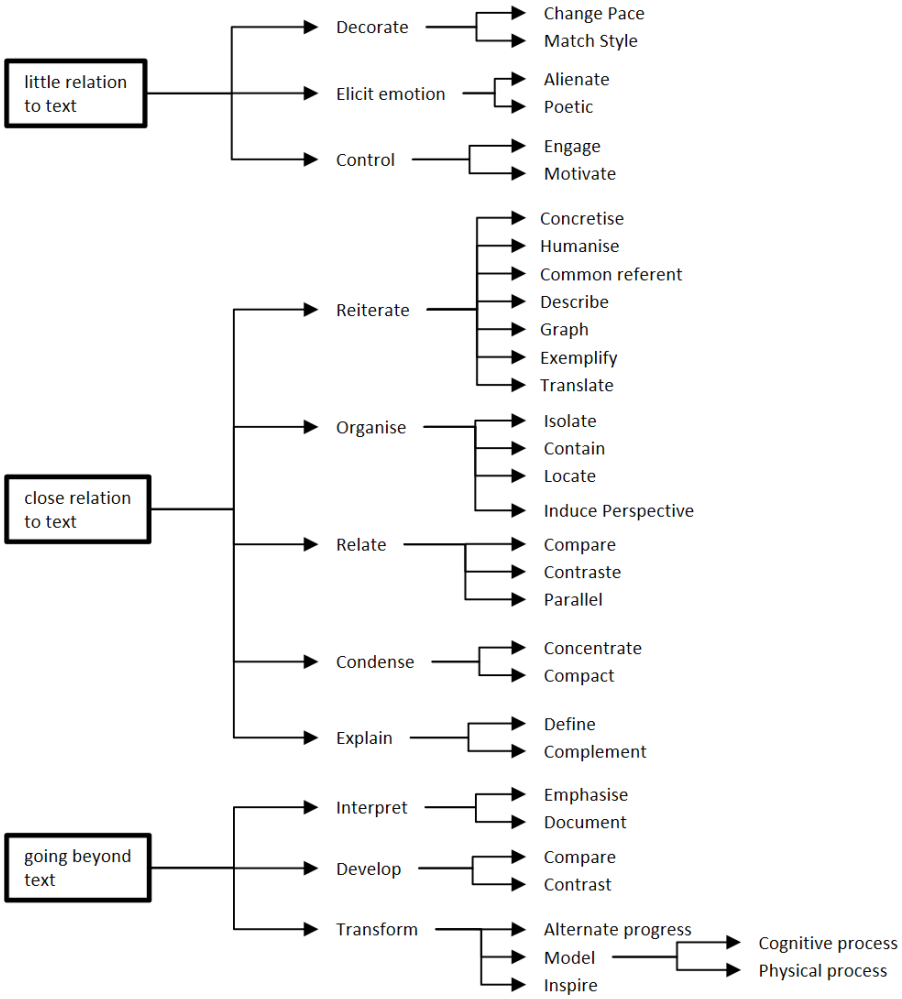


FIGURE 2.16: Marsh and White's classification system that consists of three subtrees distinguished by how closely visual and textual information are connected.

The classification system presented by [167] and the extension by [264] already consider a substantial amount of semantic image-text. Consequently, the process of assigning a label to an image-text sample is tricky, especially for longer texts and similar class definitions (e.g., *extension*, *elaboration*, *enhancement*). However, in 2003, Marsh and White [166] established a system of 46 distinct image-text classes to describe even more intricate rhetorical figures in print and digital media. Their system, see Figure 2.16, contains three categories measuring how closely the image is related to the text in portraying the overall message.

Interestingly, this is a unidirectional version of what Henning and Ewerth [104] propose with their finegrained semantic correlation (SC) metric. We discuss their work in Chapter 5. Basically, *little relation to text* implies that the image has a subordinate role, for example, as a form of *decoration*. Next, a *close relation to text* encapsulates forms of relations similar to

Unsworth's *concurrence* and *complementarity* subtrees describing different ways of *reiterating* information in the opposite modality, *clarifying* details of examining whether the content is, for example, of *parallel* or *contrasting* nature. The last category, *going beyond text*, summarizes instances where the image goes beyond the information given in the text by *interpreting*, *developing*, or *transforming* them. While the plethora of image-text classes makes it seem as if a majority of possible combinations can be covered and the authors themselves claim that their system is "largely complete", Bateman [16] points out that this was not evaluated based on strict criteria but rather by applying the framework to random websites and see whether the taxonomy covers all occurring visuals. Marsh and White circumvent the aforementioned challenge of assigning an image-text pair to a distinct class by describing them with multiple relations of their classification. This, however, makes the process of determining a definitive set of labels even more subjective.

From a computer science perspective, the number of different classification systems, the heterogeneity of approaches towards differentiating between the relations, and the ambiguity between classes within the individual systems make adopting research from media and communication science challenging. From our perspective, an optimal taxonomy of semantic image-text classes has the following attributes:

- not limited to a media domain
- allows every possible image-text pair to be assigned to exactly *one* class
- the assignment process by multiple people should achieve high inter-coder agreement
- image-text classes should be derived from measurable metrics and not from a non-representative set of image-text pairs

In Chapter 5, we propose a novel categorization of semantic image-text classes based on basic, interpretable metrics to establish an entry point to computable image-text relations from a computer science perspective.

◇ ◇ ◇

The next Chapter presents the first category of contributions surrounding research question 1. We present two approaches to improve exploratory search in the TIB AV-Portal, a learning-oriented video platform.

3 Improving Video Learning Platforms with Text-Based Features

This Chapter presents two approaches that focus on improving an educational video platform called the TIB AV-Portal [259]. Given the unique metadata provided, we introduce methods to utilize this textual information to improve the exploratory search capabilities of the platform. Our goal is to answer the first research question, namely:

Research Question 1

How can we utilize textual metadata associated with learning content to improve exploratory search in video search portals?

Section 3.1 introduces a ranking algorithm for related videos based on semantic word embeddings in conjunction with linked open data from the GND. Since educational videos tend to be longer on average, we propose a method to give an overview of the content independent of video length in Section 3.2. Section 3.3 concludes our findings and discusses the implications of this Chapter.

3.1 Recommending Scientific Videos based on Metadata Enrichment using Linked Open Data

3.1.1 Motivation

Videos hold a great potential to communicate educational and scientific information. The growing influence of e-Learning platforms such as Udacity [263] or Coursera [53] reflects that [169]. However, a growth in available content makes it more challenging for providers of e-Learning websites to recommend relevant results. Optimally, retrieval algorithms have to align search queries, which can be short and imprecise, with hours of video content while ensuring that the retrieved documents are of good quality, recent and tailored towards the learners' assumed state of knowledge [117]. To narrow down the semantic gap [17] between the query and the high-level semantics in, e.g., video content, can be expensive and time-consuming.

Because of this, recommender systems in online shopping platforms or video portals mainly rely on user-based information such as the viewing history [56], current trends [54], or item similarity [100]. There is also another type of Web portals that offer exclusively scholarly videos, one of them being the TIB AV-Portal [259] of the Leibniz Information Centre for Science and Technology (TIB). Researchers can provide, search, and access

scientific and educational audio-visual material, while benefiting from several advantages compared to other portals. First, the TIB AV-Portal reviews submitted videos to check whether they contain scientific or educational content. Second, videos are represented in a persistent way using Digital Object Identifier (DOI), potentially even at the segment and frame level, making it easy and reliable to reference them. Finally, they apply audio-visual content analysis in order to allow the user to not only search for terms in descriptive metadata (e.g., title, manually annotated keywords) but also in the audio-visual content, i.e., in the speech transcript, in the recognized overlaid or scene text through video OCR, and keywords derived from VCD.

In this section, we investigate how similar videos can be recommended based on their metadata, particularly by using automatically extracted metadata from audio-visual content analysis. This is relevant, for example, when users do not agree that their search behavior is tracked or a sufficient amount of user data is not available. Particularly, we propose to exploit and enrich the entire set of available metadata, be it created manually or extracted automatically, to improve recommendations of semantically similar videos. In the first step, we utilize a Word2Vec approach [130] to make the semantic content of two videos comparable based on title, tags, and abstract. Then, we enrich automatically extracted metadata about the audio-visual content by linking them to the Integrated Authority File (in German: *GND* - Gemeinsame Normdatei) of the German National Library (in German: *DNB* - Deutsche Nationalbibliothek). We use these two kinds of information to derive a measure to compare the content of two videos which serves as a basis for recommending similar videos. A user study demonstrates the feasibility of the proposed approach.

First, we give a brief overview of related work in Section 3.1.2. The proposed approach to generate video recommendations is presented in Section 3.1.3. Section 3.1.4 describes the conducted user study to evaluate the proposed approach.

3.1.2 Related Work

Scientific Video Portals

Yovisto is a scientific video portal that allows the user to search for information via text-based metadata [274, 275]. Learners can reduce the number of search results by refining their query via additional criteria and grouping videos by language, organization, or category. On the contrary, to increase the scope of possible results, a tool for exploratory search reveals interrelations between different types of videos to present a broader spectrum of results to the user. Their approach is to exploit an ontology structure, which is part of every video element and Linked Open Data (LOD) resources, namely DBpedia [57]. Marchionini [165] describes a similar portal that automatically feeds the uploaded content into a data analysis chain. This process assigns semantic entities to each video segment resulting in a storyboard comprising the video content. In contrast to the AV-Portal, their

portal hides this information from the user. Marchionini's approach focuses on providing a good exploratory search tool, i.e., a user should find what s/he is looking for even when unsure about the correct phrasing.

Recommendation Systems for Scientific Videos

Clustering semantically similar videos is a possible approach to providing video recommendations based on a currently watched video. A fundamental problem of this research is the semantic gap between low-level features and high-level semantics portrayed in visual content [17]. One approach towards solving this problem is using textual cues in addition to visual content. These can be manually added tags by the video author or automatically extracted keywords by machine learning algorithms. Either way, they are often superficial, noisy, incomplete, or ambiguous, making clustering a challenge. Vahdat et al. [265] enrich the set of tags by modeling them from visual features and correcting the existing ones by checking their agreement with the visual content. They show that this method outperforms previous ones that use either modality and even the naive combination. Wang et al. [278] discover that by incorporating hierarchical information – instead of considering a "flat" tag taxonomy – the semantics of a video can be described even better. Despite only using two levels of abstraction in their hierarchical multi-label random forest model, they find strong correlations between ambiguous visual features and sparse, incomplete tags.

3.1.3 Framework

Next, we present our approach to enrich metadata with open data sources. First, we describe the set of available metadata before the acquisition of additional information from an open data source. Second, we derive a similarity measure to compare videos based on a Word2Vec representation and enriched metadata. The overall workflow is displayed in Figure 3.1. The input of our system consists of manually generated and automatically extracted information, where the former comprises abstract and title. Additional inputs are the following automatically extracted **Tags** (see Figure 3.1) derived from: 1) Transcript based on speech recognition, 2) Results of video OCR, and 3) results of visual concept and scene classification. They all have a representation in the German National Library, which is the key requirement for the enrichment process.

Acquiring Additional Information from Open Data Source

Automatically generated tags usually contain a certain amount of errors and noise. Although state-of-the-art algorithms can achieve human performance [239] in specific tasks and settings, issues with audio quality in lecture rooms or hardly legible handwritings can cause errors. We try to circumvent this problem by evaluating additional information provided by the German National Library. Besides information such as synonyms and related scientific publications, they provide the *Dewey Decimal Classification (DDC)* for every tag. The DDC is a library classification system, which categorizes technical terms

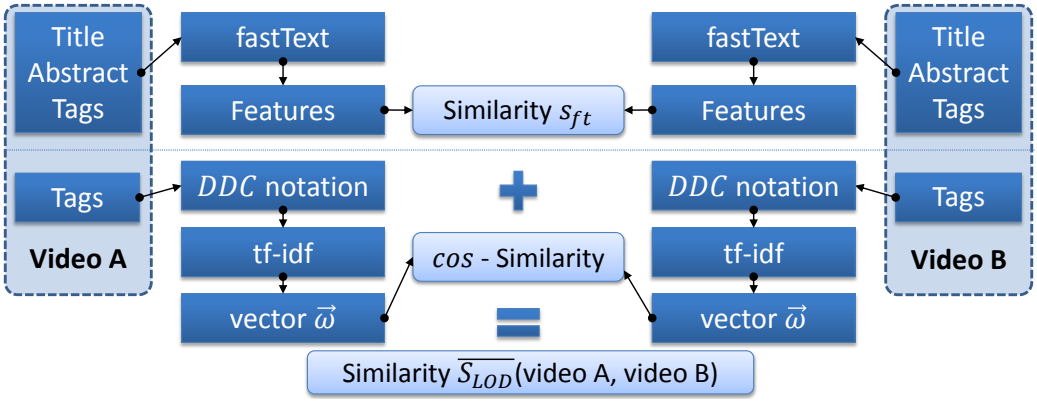


FIGURE 3.1: The general workflow of the approach combining the *method without LOD* (upper half) with the features from the DDC notation (lower half).

into ten classes via three-digit Arabic numerals [221]. We further divide these main classes into subcategories denoted by the decimals after these three digits, where additional decimals depict a more specific subject. For instance, *SPARQL* is contained in *006.74 - Markup Language*, *005.74 - Data files and Databases* and *005.133 - Individual Programming Languages*, which yields valuable contextual information.

Defining a Similarity Measure for Scientific Videos

Simply comparing two videos for mutual tags is not sufficient to determine semantic similarity. Even if two sets of tags have little to no overlap, they might be highly correlated when considering their context. We address this issue by utilizing fastText [130] to generate word embeddings, which has several advantages for this task. First, they represent semantically similar words mathematically closer to one another so that a distance measure like cosine similarity indicates the correlation of two words. Second, since fastText works on substrings rather than whole words, it can produce valuable features even for misspelled or words unknown to the word embedding. Finally, a pre-trained model is available for a large number of languages. Title, tags, and abstract are taken from the metadata and processed via fastText. It generates a 300-dimensional feature vector for every word in the metadata. The average of these vectors is our representation for a particular video. This approach is our baseline and denoted as *method without LOD* in the sequel. The improvement of this already powerful feature extraction method is the main contribution of this section. We achieve this by incorporating the information provided by the DDC notation in addition to the fastText embeddings. As a preprocessing step, we need to create a vector ω, which consists of all DDC tags in our dataset and will be assigned to every video entry v. Since the DDC notation encodes the upper-level classes in the codes of the classes at lower levels, we divide them accordingly. Therefore, the length of ω equals the total number of these tag fragments. For instance, if the video corpus would only contain tags 005.74 and 005.133, we would split them into $5_1, 57_2, 51_2, 574_3, 513_3, 5133_4$ (indices mark the level in the hierarchy), resulting in a vector

ω of length 6. If a particular tag fragment occurs in a video, we set the corresponding bin in ω to the *term frequency-inverse document frequency (tf-IDF)*, or zero otherwise. This strategy ensures that the more specific, and thus more informative, DDC classes have a greater influence on the result. For example, if two tags share the main DDC class *Science and Mathematics*, it does not mean that they are necessarily closely correlated, but if both share the class *Data Compression*, they most likely cover a similar topic. For the "method with LOD" the two vectors ω_i and ω_j of video v_i and v_j are compared via cosine similarity. It is important to note that this method also uses the fastText features of the *method without LOD*. In order to compute the overall similarity, we apply both methods and use their average to calculate s_{LOD} (see Figure 3.3).

3.1.4 Experiments and Results

We used videos crawled from the TIB AV-Portal in the experiment. The complete stock of metadata that falls under the Creative Commons License CC0 1.0 Universal is made available by the TIB as Resource Description Framework (RDF) triples. To extract the necessary annotations, we utilized SPARQL. In a first step, it was necessary to keep only videos that allowed *derivate works* in addition to the CC0 1.0 license, since content analysis is applied. 2 066 samples satisfied these conditions[1]. Unfortunately, we can not directly compare word embeddings of two different languages forcing us to use a subset of videos with the same language (German, 1 430 videos). Annotations are represented in JSON format to make them easily accessible for future tasks without rebuilding the RDF graph. After gathering all tags of an entry, we employed another SPARQL query assigning a GND (German: *Gemeinsame Normdatei*, English: Integrated Authority File) link to each tag, which is the key part of linking it to the data of the German National Library (DNB) and retrieving the corresponding DDC notations.

We evaluated the quality of our similarity measure by conducting a user study with eight participants, five men, and three women. A random selection of 50 videos was presented to every participant along with ten video recommendations, randomly either entirely provided by the *method without LOD* or the *method with LOD*. We integrated the video recommendations into the live system with a Greasemonkey script in the Firefox browser. Every participant had to rate each of the ten recommendations from *None* to *High*, i.e., *None*: not relevant; *Low*: low relevance; *Medium*: medium relevance; *High*: highly relevant. The results are displayed in Figure 3.2.

The results show that the *method with LOD* increases the number of video recommendations with medium (4.56%) and low relevance (11.29%), while the effect is small (0.97%) for the highly relevant recommendations. However, the *method with LOD* significantly decreases the number of irrelevant recommendations (by 18.17%). A Chi-Square test (Chi-Square=15.1471, p-value=0.001695) indicates that this method is significantly better than the text-based method, most likely due to the hierarchical nature of the DDC notation.

[1] as of June 16, 2017

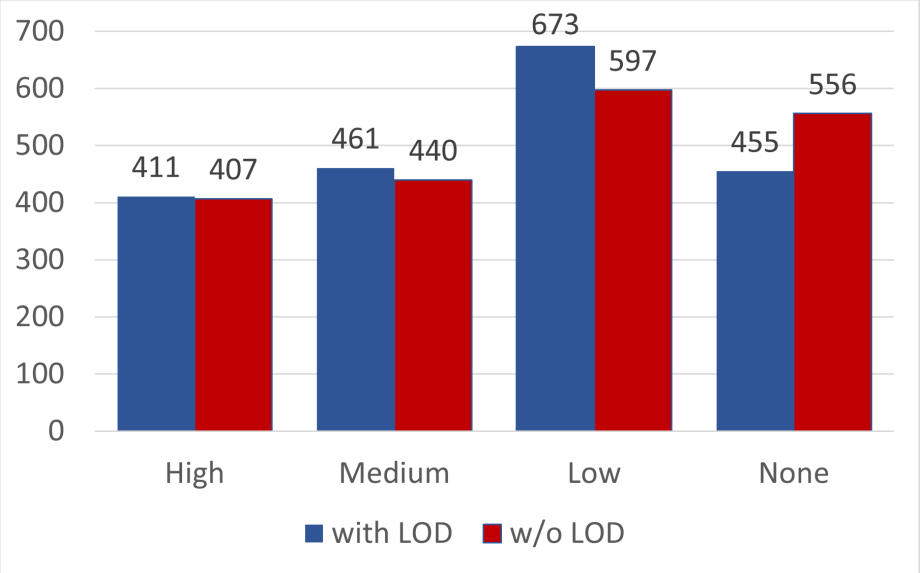


FIGURE 3.2: Absolute number of votings for each relevance level in the user study.

We assume that the relatively slight improvement for the very relevant recommendations results from the restrictions we had to oblige due to license and language, i.e., the relatively small set of remaining videos (1 430) does not contain more highly relevant samples.

3.1.5 Summary

In this section, we have proposed a method to generate recommendations for scientific videos based on noisy, automatically extracted tags by utilizing linked open data to weave in hierarchical semantic metadata. This enables users to find relevant information more quickly improving their overall learning experience. Next, we explore how we can give the learner a first quick impression of an unseen video's content for them to decide whether to watch it or move on to the next without having to scroll through the entire video.

3.2 Visual Summarization of Scientific Video Content

3.2.1 Motivation

The massive growth of online video platforms underlines the role of audio-visual content as one of the most commonly used sources of information for entertainment and learning-related scenarios. Exploring an extensive collection of videos to find the most relevant candidate for a specific learning intent can be overwhelming and, therefore, inefficient. This is especially true for longer videos if the title alone cannot capture all parts and aspects of the content. Approaches for *video summarization* address this problem by analyzing the visual content and generating an overview by the combination of identified key sequences and frames [8]. However, such approaches struggle with videos, where the visual content lacks variance or is mainly comprised of concepts with low *visualness* [290], e.g., abstract concepts. Scientific and educational videos often share this characteristic, for example, tutorials or lecture recordings of the STEM subjects (Science, Technology, Engineering, and Mathematics) like chemistry or computer science.

After discussing related work in the areas of video summarization and keyphrase extraction in Section 3.2.2, we propose an interactive visualization approach to summarize the content of scientific or educational videos in Section 3.2.3. The goal is to provide an approach that facilitates the exploratory search capabilities of respective video portals, thus making learning for the end-user more efficient and satisfying. Our approach uses automatically extracted video annotations and entities, which significantly enrich the usually available, conventional metadata. As described in Section 3.1, these entities are generated from the 1) ASR, 2) VCD, and 3) text extracted using OCR. The TIB AV-Portal [259] makes this type of metadata publically available. We choose the TIB AV-Portal as the primary platform for these reasons and incorporate the proposed approach there. Our system utilizes these data and generates a comprehensive, interactive visualization by combining semantic word embeddings and keyphrase extraction methods. We demonstrate how to display the visualization on the actual website with a *GreaseMonkey* script, which is also a pre-requisite for our user study (Section 3.2.4) that investigates the usefulness of the proposed approach for video content visualization.

3.2.2 Related Work

Video Summarization

The vast majority of video summarization algorithms rely on visual features and are very domain-specific (e.g., movies, sports, news, documentary, surveillance), resulting in a large number of different approaches. The focus of these approaches can be dominant concepts [203], user preferences [160], query context [279] or user attention [161]. A typical result of these approaches is a sequence of keyframes or a video excerpt comprising the most important parts of a video [8]. More recent methods treat video summarization as

an optimization problem [296, 86, 64] or they utilize recurrent neural networks [297, 301] based on, for instance, long short-term memory cells (LSTMs), which can capture temporal or sequential information very well. Another use case for LSTM is proposed by Mahasseni et al. [163], who suggest a generative adversarial network (GAN) consisting of an LSTM-based autoencoder and a discriminator. Some methods include textual information (e.g., tags [112] or full documents [149]), which result in a storyboard that provides short titles for each key shot, which is particularly useful for news summarization. Scientific or scholarly videos provide a more significant challenge in this respect since their visual content often lacks visualness. Consequently, summarization techniques focus even more on textual metadata. Chang et al. [38] combine image processing, text summarization, and keyword extraction techniques resulting in a multimodal surrogate. They generate a word cloud displaying more important words with a bigger font size together with a set of three to four thumbnails with a short transcription.

In this work, we go one step further and show how to summarize the content solely based on textual information. The core techniques to create a video summarization utilized in this section are keyphrase extraction and measures for semantic text similarity. We describe related work in these respective areas next.

Keyphrase Extraction

Hasan and Ng [96] describe that keyphrase extraction techniques generally consist of two steps. First, they identify a list of possible candidate phrases, and then these candidates are ranked according to their importance. They categorize these ranking approaches into supervised and unsupervised methods. Early supervised algorithms rely on, for instance, decision trees [262]. Hulth [120] extends this approach by adding linguistic features to a bagged decision tree classifier while also extending previous work by filtering incorrectly assigned keywords with different feature pairs. Another approach [68] utilizes lexical chains based on a WordNet ontology, which is associated with features such as first occurrence position, last occurrence position, and word frequency. Additionally, support vector machines [276], maximum entropy classifiers [138, 154], conditional random field models [295], logistic regression [88] and neural networks [277, 125] have been used to solve the task of finding the most important phrases in a document.

The techniques mentioned above share a drawback: the training data requires manual labeling, which generally introduces unrealistic experimental data and is time-consuming and resource-intensive. Thus, unsupervised approaches moved into the focus of attention. Their task is to automatically discover the underlying structure of a dataset without human-labeled keyphrases. To summarize, the two most popular methods are *graph-based ranking* and *topic-based clustering*. The idea behind graph-based algorithms is to construct a graph of phrases connected with weighted edges that describe their relation derived from the frequency of their co-occurrence [181]. Topic-based clustering methods use statistical language models, which contain the probability of all possible sequences of

words [20]. Recently, fusions of these two directions gain attention, namely TopicRank [25], PositionRank [76] and MultiPartiteRank [24]. The latter one, which we also use in this work, first builds a graph representation of the document and then ranks each keyphrase with a relevance score. In addition, it adjusts the edge weights to capture information about the word's position in the document in an intermediate step.

3.2.3 Framework

In this section, we describe our approach for video content summarization solely based on textual information. The necessary process to summarize a scientific video and display this information in an efficient way 3.3 consists of four steps: 1) pre-processing, 2) semantic embedding of content-related information to generate a bubble diagram, 3) creation of a keyphrase table from the speech transcript, and 4) combining diagram and table to form a visualization. The utilized video dataset from the TIB AV-Portal is available at [260], including the associated metadata as RDF triples (under Creative Commons License CC0 1.0 Universal).

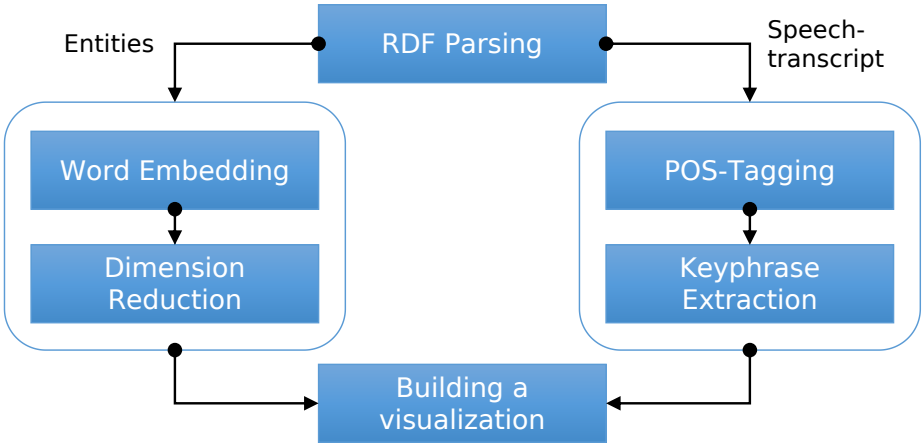


FIGURE 3.3: Workflow diagram of the proposed visualization approach.

To build the RDF graph we use Python 3.6 and the *rdflib* library. Next, we use the query language SPARQL to select videos that contain automatically extracted metadata (this applies only for videos related to the six core subjects of the TIB[2]). An exemplary query can be seen in Listing 3.1.

```
1 PREFIX dcterms: <http://purl.org/dc/terms/>
2 PREFIX oa:      <http://w3.org/ns/oa#>
3 SELECT DISTINCT ?url
4 WHERE {
5     ?annotation oa:annotatedBy asr_link.
6     ?annotation oa:hasTarget ?videofragment .
7     ?videofragment dcterms:isPartOf ?url .}
```

LISTING 3.1: SPARQL-query that returns all videos which contain automatically analyzed speech transcripts ASR and recognized entities.

[2]engineering, architecture, chemistry, computer science, mathematics, physics

This query yields a list of 1756 videos from multiple languages, which we then query further for the embedded metadata, in particular the **key entities** which are the result of VCD, OCR, and ASR. Additionally, we crawl the unfiltered speech transcript from the website using the *BeautifulSoup* library. We use fastText to generate word embeddings from the extracted key entities. fastText's tri-gram technique embeds words by their substrings instead of the whole word. For instance, it decomposes the word *google* into the following tri-grams: <go, goo, oog, ogl, gle, le>. This decomposition is a valuable feature for multiple reasons. First, it enables the system to encode misspelled or unknown words. Secondly, the quality of embeddings of the generally longer or compound words of the German language is improved, too. We use the pre-trained model for German language [74], which contains the vocabulary of the German Wikipedia and encodes each word w in a 300-dimensional vector f_w.

The visualization of the embedded feature vectors requires dimension reduction to project data onto a two-dimensional space. We apply a principal component analysis (PCA) instead of a non-linear algorithm like t-distributed stochastic neighbor embedding (t-SNE) since we intend to keep the semantic arrangement laid out by fastText and refrain from clustering the keywords further. Input for the keyphrase extraction process is the unfiltered speech transcript, which is already divided into time segments by the TIB AV portal. The required format of the textual information is given by the *pke* toolkit [23] which is shown in Listing 3.2. Requirements are tokenization and part-of-speech (POS) tagging, which is the assignment of lexical categories such as nouns, verbs, adjectives, and adverbs. For this process, we use the Python *Natural Language Toolkit (NLTK)*, in particular, the Stanford POS-Tagger, which also comes with a pre-trained model for the German language [248].

```
1 wenig/PRON Speicher/NOUN es/PRON kommen/VERB [...]
```

Listing 3.2: POS-Tagged speech transcript labeled with lexical categories.

Results of the POS tagging process are then passed to the Multipartite Rank [24] algorithm of the *pke* library in order to perform keyphrase extraction. As stated in Section **Related Work**, this technique models topics and phrases in a single graph, and their mutual reinforcement together with a specific mechanism to select the most import keyphrases are used to generate candidate rankings. We only consider nouns, adjectives, personal pronouns, and verbs ('NOUN', 'ADJ', 'PROPN', 'VERB') and dismiss all words given by *NLTK*'s collection of German stop words. The remaining parameters are $alpha = 1$, which controls the weight adjustment mechanism, and the $threshold = 0.4$ for the minimum similarity for clustering (default: 0.25). We decide to set this value to 0.4 due to the high similarity of topics in a single video. The linkage method was set to *average*. Finally, we choose to retrieve the 20 highest ranked keyphrases of every time stamp for our keyphrase table that will become part of the visualization.

Finally, we display the recognized, embedded entities in an interactive graph with the properties shown in Table 3.1 and combine it with the keyphrase table generated in the last section.

Components	Meaning	Approach
circle	key topics	recognized entities
the size of a circle	importance of the topic	the frequency of the entity
arrangement	similarity between topics	word embeddings
table	timestamp summary of speech transcript	keyphrase extraction

TABLE 3.1: Overview over the properties of the visualization.

We choose a bubble diagram as opposed to Chang et al.'s [39] word cloud. This allows us also to illustrate and emphasize the distance between related or unrelated keywords, which reflects (dis)similarity. In addition, minor differences in area sizes are visually easier to perceive than font sizes. We decided against other alternative implementations such as TextArc [204] since we aimed for a more intuitive approach. The inclusion of the temporal dimension using ThemeRiver [98] did not deliver consistent results for short videos or contained only a few keywords. In addition, ThemeRiver is less suitable to represent the similarity of several entities. The actual implementation is done in Javascript and the *Plot.ly* API[3]. As displayed in Figures 3.4 and 3.5, the visualization entails circles of different sizes, each representing a topic and its importance. An interactive toolbar is displayed on the upper right allowing the user to explore the graph easily. At the bottom, a keyphrase table indicates the main topic of each time segment.

3.2.4 Experiments and Results

We conducted a user study to evaluate the quality and usefulness of the proposed visualization approach. Ten participants were recruited, of which eight were male and two female. Their ages were between 21 and 30, and their educational levels were between high school and master's. Seven participants study computer science, one mechanical engineering, and one mathematics. All of them are fluent in German, with four of them being native speakers. **Task I** of the study investigates how precisely the visual summary represents the video content. Therefore, we randomly assigned ten videos with a duration of 5 to 30 minutes to each participant. Then, the user had to rate how well the presented visualization matches the video content, based on the following options: "0" - no correlation, "1" - slight match, "2" - good match, "3" - exact match. **Task II** aimed to evaluate if the visualization is a valuable tool to provide a quick overview of the video content or if it is no improvement over the current state of the website. The participants could choose one of the following options to rate the usefulness: "0" - not helpful at all, "1" - slightly helpful, "2" - moderately helpful, "3" - very helpful, "4" - extremely helpful, and had to give a short statement about their reasoning. Figure 3.6a shows the distribution of the 100 gathered ratings, while Figure 3.6b shows the results of Task II.

[3] https://plot.ly/api/

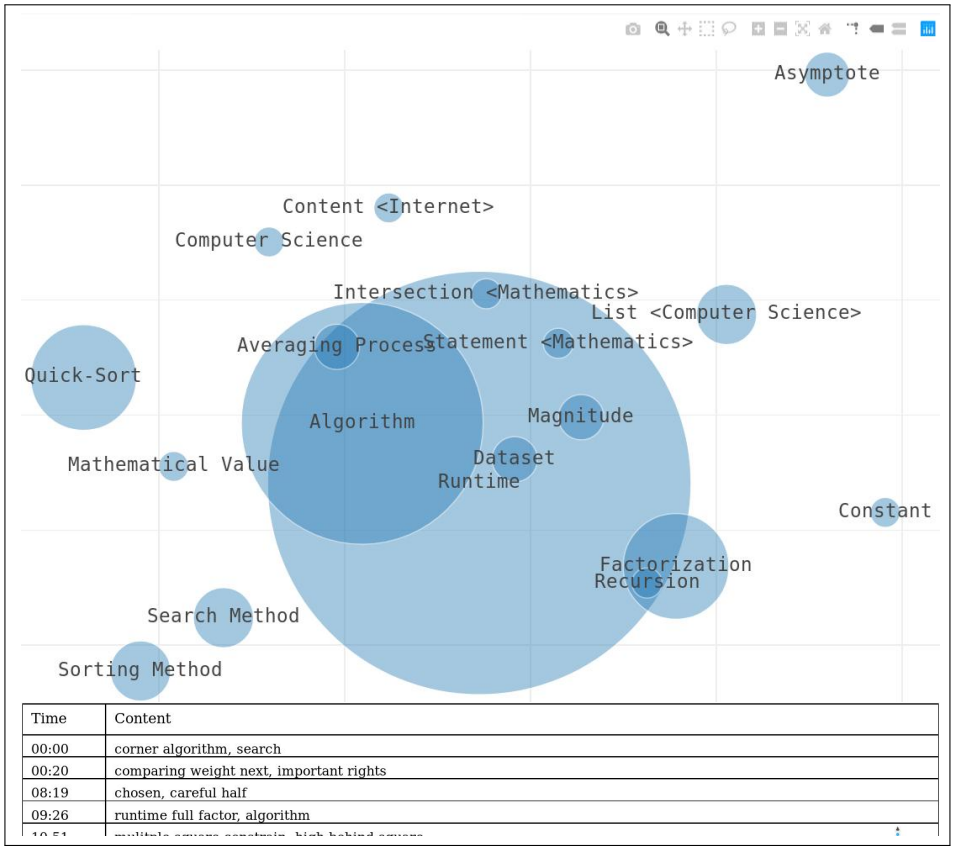


FIGURE 3.4: Visualization of video https://av.tib.eu/media/9557 titled "Bubblesort, Quicksort, Runtime" incorporated via GreaseMonkey in the live website as portrayed during the user study comprised of the visualization itself, a toolbar and the keyphrase table. Note: Translated for better comprehensibility.

Discussion

Figure 3.6a shows that 68% of the visualizations were good or better, while 26% only provided a slight match or did not correlate at all to the video content (6%). As shown in Figure 2, positive examples successfully provide the user with a summarization of the video content. The first example, video 9557, explains the runtime behavior of the sorting algorithms Bubblesort and Quicksort. The largest circle in the visualization is Runtime ("Laufzeit") and represents the main topic well. Also, the visualization groups different subtopics related to computer science that are present in the video. For instance, it groups sorting methods ("Sortierverfahren"), algorithm ("Algorithmus"), and Quick-Sort itself on the left, while related topics from mathematics like factorization ("Faktorisierung"), asymptote ("Asymptote"), and statement ("Aussage <Mathematik>") are on the right. Another positive example (video 10234), which talks about eigenvalues and eigenvectors, is mainly represented by the entity matrix multiplication ("Matrizenmultiplikation") and vector ("Vektor"), but also shows more detailed aspects of that topic, namely vector algebra ("Vektorrechnung"), inverse matrix, gradient and of course, eigenvector and eigenvalue.

The results of the keyphrase extraction, as can be seen in Figure 3.4, were less helpful. We

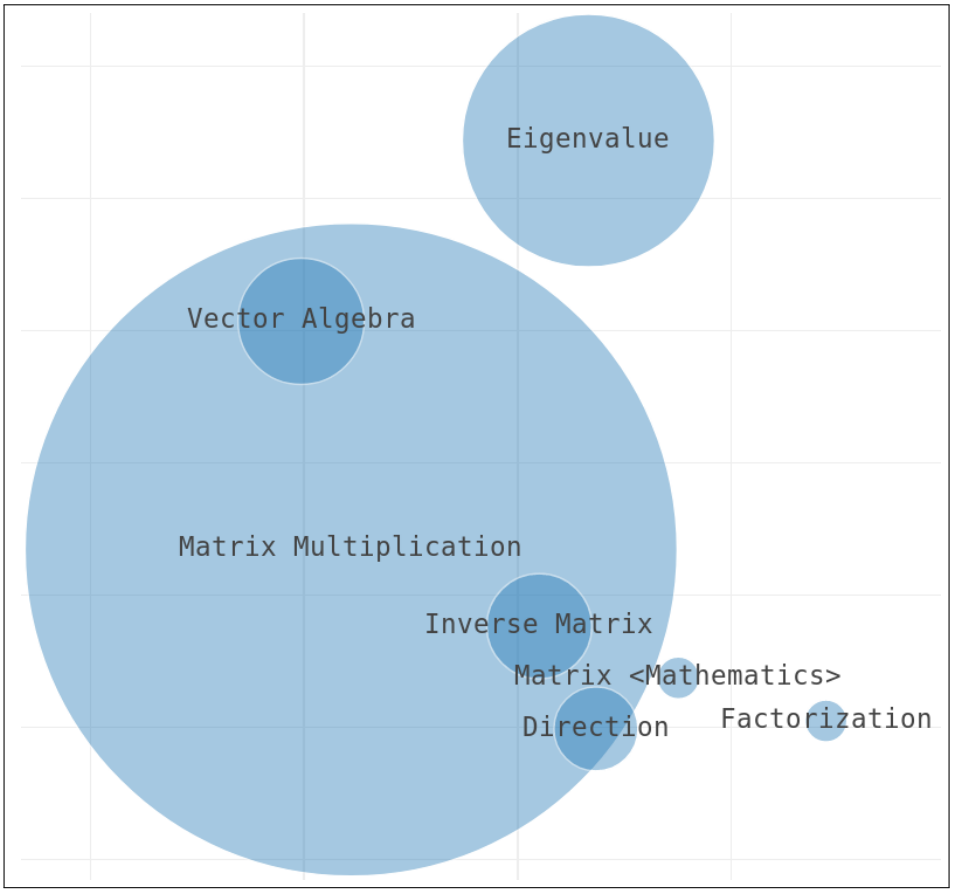
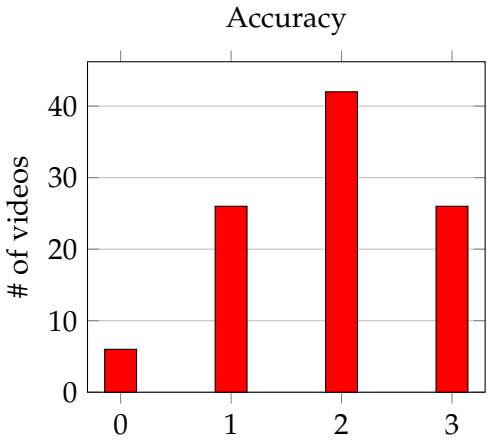
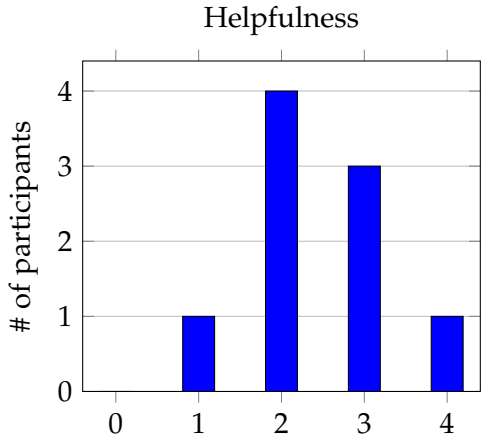


FIGURE 3.5: Visualization of video https://av.tib.eu/media/10234 titled "Eigenwerte, Eigenvektoren" (eng: "eigenvalues, eigenvectors"). Note: Entities were translated for better comprehensibility.



(A) Results of Task I of the user study evaluating the correlation of the visualization to the video content. From "0" - no correlation to "3" - exact match.



(B) Results of Task II of the user study showing the perceived helpfulness of the visualization. From "0" - not helpful at all to "4" - extremely helpful.

assume the main reason for this effect is the nature of the automatic speech transcripts, which usually differ from written text. They often consist of incomplete sentences, misspelled words, missing punctuation, and falsely recognized words that can completely change a sentence's interpretations. Since standard keyphrase models are suited for proper

textual content, there is still room for improvement in our scenario.

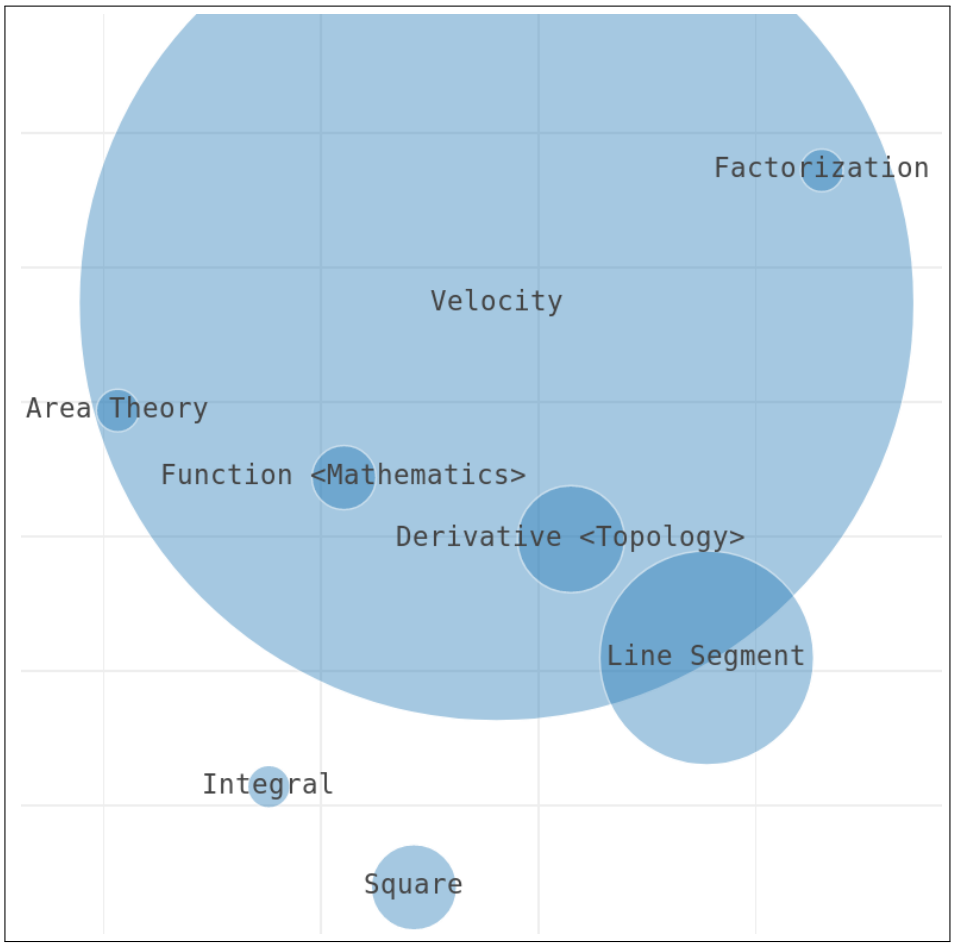


Figure 3.7: Visualization of video https://av.tib.eu/media/9915. It demonstrates the effect of very common entity "Geschwindigkeit" (eng: velocity), which was used frequently by the speaker during an example scenario, but is misleading since the video talks about arc length computation. Note: Entities were translated for better comprehensibility.

3.2.5 Summary

In this section, we have presented a system that summarizes and displays the content of scholarly videos in order to support semantic search in video portals. Based on entirely automatic video content analysis as conducted in the TIB AV Portal, we have proposed an approach that leverages the resulting metadata and generates an interactive visualization and a keyphrase table to outline the content of a video. Different techniques like POS-Tagging, semantic word embeddings and keyphrase extractions were exploited in our approach. The usefulness of the visualization was evaluated in a user study that demonstrated the feasibility of the proposed visual summarization, but also indicated areas for future work. For instance, we plan to implement reliable filters for keywords that are not closely related to the content to provide a better user experience.

3.3 Summary

Research Question 1

How can we utilize textual metadata associated with learning content to improve exploratory search in video search portals?

This section outlined two examples of how previously established textual information can be enriched and post-processed to obtain rich(er) semantic features. We demonstrated a video recommendation tool and a content visualization method that we tested by overlaying the TIB AV-Portal website with a JavaScript plugin. The plugins' usability has been evaluated by one user study each, reporting moderate to good results regarding accuracy and helpfulness. Thus, regarding *Research Question 1*, employing semantic word embeddings to represent textual metadata of educational resources can help to improve exploratory search. This effect is limited by the quality of the respective metadata as revealed by Section 3.2.4.

◇ ◇ ◇

The next chapter extends this topic in various directions. We consider more modalities for our feature sets, utilize data of more extensive user studies and evaluate our methods regarding the capability to predict the potential KG of a learning resource.

4 Prediction of Knowledge Gain with Multimodal Features

As indicated by the results in Chapter 3, textual features can represent the content of educational resources in a way that resembles a (more or less) human-like *understanding*. That means being able to set them into context with each other to a certain degree. The following chapter explores, with respect to research question 2, how well textual features can be utilized for knowledge gain prediction and, thus, to assess the general capability of a website to convey information. The word *general* in this context means *for the average learner* since we are not factoring in individual preferences of the proband, which are, however, a significant part of the human learning process as outlined by Hoyer et al. [117].

Research Question 2

To what extent can we extract textual, multimedia, and cross-modal features and utilize them for knowledge gain prediction?

In a first study (Section 4.1), we establish a substantial list of textual features that capture various aspects of the given texts. In addition, we extend our experiments to other modalities. We consider visual, audio, audio-visual, and lecturer-specific, cross-modal features and compare their impact on the measured knowledge gain. Second, for Section 4.2 we gather a much larger dataset based on a user study. We add design- and content-specific features to an ensemble of (textual) resource features by employing different CNN-based approaches. We evaluate all features regarding their usefulness toward knowledge gain prediction by considering different classification approaches. Section 4.3 summarizes our findings.

4.1 Predicting Knowledge Gain for MOOC Video Consumption

4.1.1 Motivation

Today's Web environment has become a valuable resource for human learning, with content available to explore an abundance of knowledge – from sophisticated science topics like particle physics to everyday tasks such as changing a bike's tire. Especially video platforms such as YouTube gain more and more momentum in this field – a study states that about 50% of the daily views target some kind of learning resource [244]. However, with 500 hours of new content uploaded to YouTube alone every minute [44], it is obvious that learners require effective search and recommendation tools to find fitting content. However, the automatic assessment of this content with regard to predicting

(potential) knowledge gain has not been addressed by previous work yet. That entails, for example, objective features that assess the nature of the presenter's voice or the slides' design. A human viewer evaluates the quality of a learning resource based on all available information. In common lecture videos, this includes the textual, oral, and visual modalities. The learning process is supported by 1) visual elements on the slide, 2) spoken words, and 3) the gestures of the lecturer. While some of these features have been explored in isolation [49, 250], as we will discuss in Section 4.1.2, a comparison of their impact on the task of knowledge gain prediction has not been conducted yet. This, however, would presumably give a good indication on which video to recommend to a learner that is faced with multiple options.

In this section, we go beyond previous work by 1) gathering the real-life data from a user study that instructed learners to watch videos of a Mass Open Online Course website in Section 4.1.3, 2) proposing a novel set of intuitive unimodal and cross-modal features that do not rely on tracking the subject's body, in Sections 4.1.4 - 4.1.5, and 3) conducting an extensive set of experiments and evaluate different classification approaches as well as combinations of these features to predict the potential KG of a video in Section 4.1.6. We also consider that the user's capabilities might play an essential role in this context.

4.1.2 Related Work

Previous work, like Guo et al. [84], investigated how the design of lecture videos impacts viewers' engagement and provided recommendations to optimize the content accordingly. Chen et al. [40] used multimodal sensing to assess the quality of a presentation. They extracted speech, body movement, and visual features from the slides shown. They applied Principal Component Analysis to human ratings to address the two primary modalities of the presentation: 1) recital skills, including, for instance, voice information and body language, and 2) slide quality, with regards to grammar, readability, and visual design. The authors used Pearson correlation to measure the relation between the different features. Haider et al. [93] proposed a system for automatic video quality assessment, which is the most similar to our approach, focusing on prosodic and visual features. They extracted the complete set of audio features from the ComParE challenge [234] and a total of 42 features related to hand movements of the speaker. The employed Multimodal Learning Analytics (MLA) dataset [194] contains 416 oral presentations (in Spanish) and the respective metadata regarding speech, facial expressions, skeletal data (acquired from tracking the learner's body) extracted from a Microsoft Kinect, as well as the shown slides. The authors labeled each video with ten individual ratings and an overall score related to the quality of the slides. A correlation study (discriminant analysis) was employed, which found that prosodic features can predict *self-confidence and enthusiasm* (of the speaker) as well as *body language and pose*, which is a quality measure their participants had to label. Their visual features showed similar results but with less accuracy.

Research on automatic assessment of learning resources has targeted several possible dimensions, such as predicting the user engagement towards a certain learning resource [37]. Another dimension for the assessment of educational resources is the learning success that a user can achieve with them, often measured as the KG [49, 266, 79, 293, 131]. Research on KG prediction largely focused on the analysis of user behavioral features [293, 79] and the influence on textual contents [250].

4.1.3 User Study

We conducted a user study to estimate the expressiveness of our automatically extracted features. Our goal was to get human ratings for different quality aspects of lecture videos to later align them with the achieved learning outcome. In addition, we ask for an overall rating of the seen videos. Furthermore, every participant was asked to fill in a knowledge test before (pre-test) and after (post-test) watching the video, aiming to measure the capability of a video to convey knowledge.

Data Acquisition

Our dataset consists of 22 videos (with associated slides and speech transcripts) from edX [62]. The course materials are Copyright Delft University of Technology and are licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License [55]. edX provides the available course materials in the following formats: videos as MP4, slides as PDF, and transcriptions as SRT. We chose this source since it does not require any further pre-processing, it is open access and the speech transcript is of high quality (presumably due to manual review).

Participants and Task

The subject of the 22 videos is software engineering. Each video has one presenter, while the entire dataset has nine different presenters with varying slide designs. We employed 13 participants (ten men, three women) from our university with a computer science background, an average age of 25.8 ± 2.4 years, and asked everyone to watch and assess nine videos. With an average video length of 8 minutes and the required time to fill out the evaluation forms, the experiment took $1.5 - 2$ hours. We rewarded the participants with 13 €/Hour. We made sure that each set of videos contained as many different presenters as possible and that each video was viewed at least by five different people. We chose to gather multiple ratings for the same video instead of investigating a more extensive set of videos to be more robust against outlier ratings.

Experimental Setting

A common way to estimate the KG of a participant during a learning session is to conduct a knowledge test before and after a controlled learning session (e.g., Yu, Gadiraju, Holtz,

Rokicki, Kemkes, and Dietze [293]). The difference between these scores indicates how much the person learned. Even though the potential KG depends on the participant, we try to circumvent this problem by choosing a subject that is most likely unfamiliar to a majority of people. It is, however, important to ensure that the participants have a chance to understand the content. Otherwise, the KG will be low again. Therefore, we selected the topic *Globally Distributed Software Engineering*[1]. On the one hand, it is a computer science topic related to the studies of our participants and a specific area that is not part of their curriculum. Thus, everyone had a chance to understand the topic based on their prior knowledge, therefore favoring a positive KG.

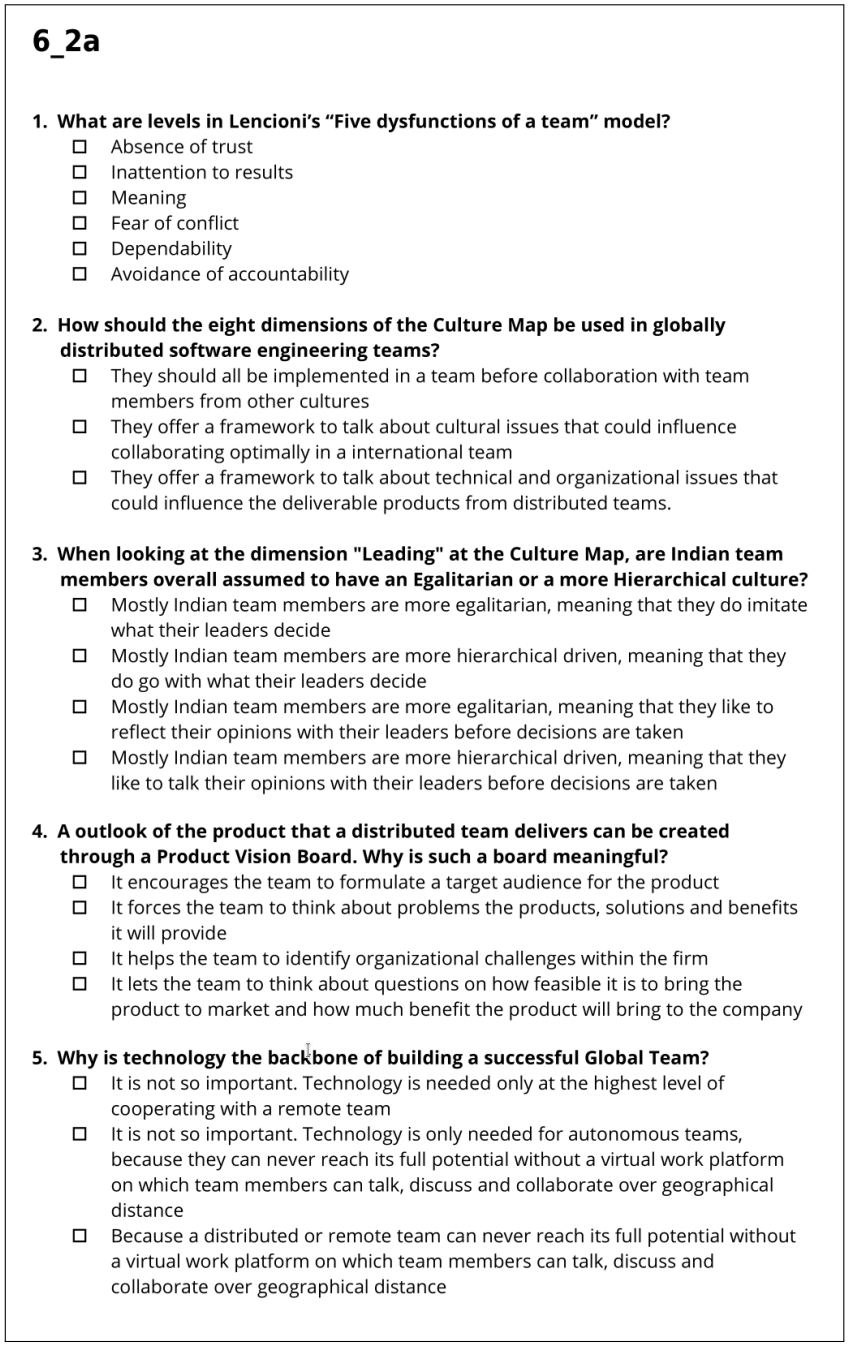
A negative effect of the pre-test is that it might influence the user behavior by providing hints on what to focus on in a video because participants will try to get a good score on the post-test. We gathered a set of relevant questions inspired by the intermediate quizzes in the course material. However, we made sure to amend and change them since their reuse is restricted. We ask two to four questions after each video with a similar amount of unrelated questions from other videos. Also, we put the videos in random order, so it was hard to guess which of the questions would be the relevant ones. In addition, the number of possible answers was different every time. Exemplary, the knowledge test for video 6_2a can be seen in Figure 4.1.

After filling out the pre-test, we instructed the participant to watch the entire video without pausing, rewinding, or taking notes. The reason for that is that we wanted the participants to get a complete impression of the presentation instead of, again, just skipping to the relevant parts for the knowledge test to get a good score. Similarly, we assumed that when we allowed people to take notes, they would write reminders down about the pre-test and focus solely on their appearance in the video. Admittedly, this is slightly different from a realistic setting, but we applied it in favor of the KG measurement. After watching the video, the person is asked to answer the same questions again and also to fill out an evaluation form (see Figure 4.2) with questions that are related to different quality aspects, see Table 4.1. We assess the items using a Likert scale from 1-5.

Knowledge Gain Scoring

This paragraph describes how we scored the knowledge test to treat each video equally, independent of 1) the number of relevant questions per quiz and 2) the number of possible answers per question. First, the score for an unanswered question will be treated as zero since we gave the participants the option to skip a question to discourage random guessing. If the participant answered the question, we would calculate the score for each answer option by increasing (decreasing) the score by 1 for a correct (false) answer. Thus, a question with five answer options can yield the following scores: $-5, -3, -1, 0, 1, 3, 5$.

[1] https://www.edx.org/course/globally-distributed-software-engineering-2



6_2a

1. **What are levels in Lencioni's "Five dysfunctions of a team" model?**
   - ☐ Absence of trust
   - ☐ Inattention to results
   - ☐ Meaning
   - ☐ Fear of conflict
   - ☐ Dependability
   - ☐ Avoidance of accountability

2. **How should the eight dimensions of the Culture Map be used in globally distributed software engineering teams?**
   - ☐ They should all be implemented in a team before collaboration with team members from other cultures
   - ☐ They offer a framework to talk about cultural issues that could influence collaborating optimally in a international team
   - ☐ They offer a framework to talk about technical and organizational issues that could influence the deliverable products from distributed teams.

3. **When looking at the dimension "Leading" at the Culture Map, are Indian team members overall assumed to have an Egalitarian or a more Hierarchical culture?**
   - ☐ Mostly Indian team members are more egalitarian, meaning that they do imitate what their leaders decide
   - ☐ Mostly Indian team members are more hierarchical driven, meaning that they do go with what their leaders decide
   - ☐ Mostly Indian team members are more egalitarian, meaning that they like to reflect their opinions with their leaders before decisions are taken
   - ☐ Mostly Indian team members are more hierarchical driven, meaning that they like to talk their opinions with their leaders before decisions are taken

4. **A outlook of the product that a distributed team delivers can be created through a Product Vision Board. Why is such a board meaningful?**
   - ☐ It encourages the team to formulate a target audience for the product
   - ☐ It forces the team to think about problems the products, solutions and benefits it will provide
   - ☐ It helps the team to identify organizational challenges within the firm
   - ☐ It lets the team to think about questions on how feasible it is to bring the product to market and how much benefit the product will bring to the company

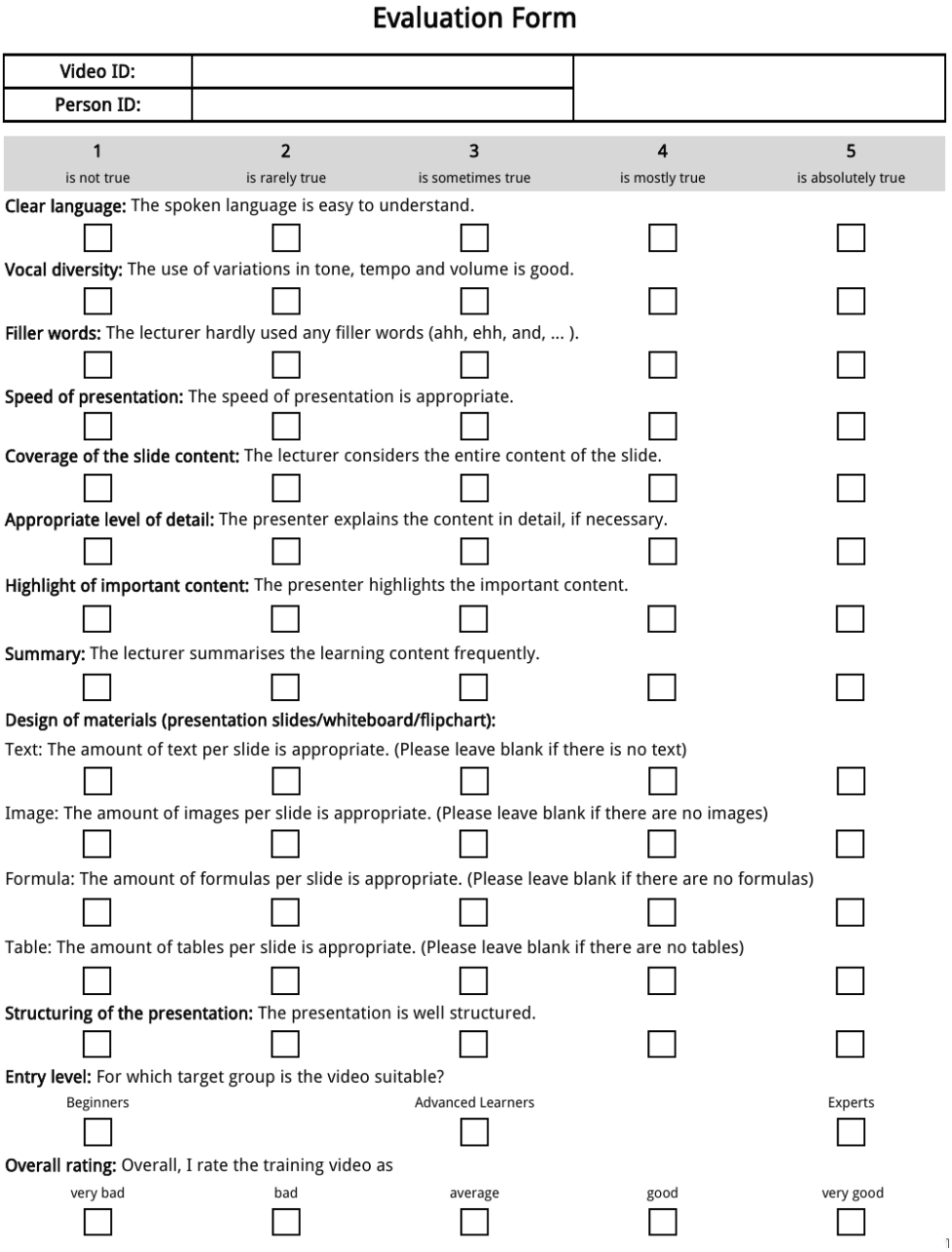
5. **Why is technology the backbone of building a successful Global Team?**
   - ☐ It is not so important. Technology is needed only at the highest level of cooperating with a remote team
   - ☐ It is not so important. Technology is only needed for autonomous teams, because they can never reach its full potential without a virtual work platform on which team members can talk, discuss and collaborate over geographical distance
   - ☐ Because a distributed or remote team can never reach its full potential without a virtual work platform on which team members can talk, discuss and collaborate over geographical distance

FIGURE 4.1: The questionnaire for the pre- and post test of video 6_2a. Questions 1 and 3 are relevant to this video.

We calculate the KG of participant's after watching video v as the difference between the pre-test score PB_{vs} and post-test score PA_{vs}. Let n_v be the number of participants who watched video v.

We start by computing the mean KG of participants for each video:

$$\mu = \frac{\sum_{j=1}^{n_v}(PA_{vs} - PB_{vs})}{2n_v}. \tag{4.1}$$



Evaluation Form

Video ID:		
Person ID:		

1	2	3	4	5
is not true	is rarely true	is sometimes true	is mostly true	is absolutely true

Clear language: The spoken language is easy to understand.

Vocal diversity: The use of variations in tone, tempo and volume is good.

Filler words: The lecturer hardly used any filler words (ahh, ehh, and, ...).

Speed of presentation: The speed of presentation is appropriate.

Coverage of the slide content: The lecturer considers the entire content of the slide.

Appropriate level of detail: The presenter explains the content in detail, if necessary.

Highlight of important content: The presenter highlights the important content.

Summary: The lecturer summarises the learning content frequently.

Design of materials (presentation slides/whiteboard/flipchart):

Text: The amount of text per slide is appropriate. (Please leave blank if there is no text)

Image: The amount of images per slide is appropriate. (Please leave blank if there are no images)

Formula: The amount of formulas per slide is appropriate. (Please leave blank if there are no formulas)

Table: The amount of tables per slide is appropriate. (Please leave blank if there are no tables)

Structuring of the presentation: The presentation is well structured.

Entry level: For which target group is the video suitable?

Beginners | Advanced Learners | Experts

Overall rating: Overall, I rate the training video as

very bad | bad | average | good | very good

FIGURE 4.2: The full evaluation form the users had to fill out for each video.

Automatic features	**Human-rated aspects**
Audio	Clear language Vocal diversity
Linguistic	Filler words Speed of presentation
Visual	text/image/formula/table design Structure of the presentation
Cross-modal	Coverage of the slide content Appropriate level of detail Highlight of important content Summary Overall rating

TABLE 4.1: Automatically extracted features and corresponding items in the evaluation form of the user study.

Next is the standard deviation of the KG:

$$\sigma = \sqrt{\frac{1}{2n_v}[\sum_{s=1}^{n_v}(PB_{vs} - \mu)^2 + \sum_{s=1}^{n_v}(PA_{vs} - \mu)^2]}. \tag{4.2}$$

Based on the mean and standard deviation value, the scores are normalized. PB'_{vs} is the normalized score which is computed by subtracting the mean and dividing by the standard deviation. The same applies for PA'_{vs}:

$$PB'_{vs} = \frac{PB_{vs} - \mu}{\sigma}. \tag{4.3}$$

Consequently, the normalized KG of participant s for video v is:

$$KG_{vs} = PA'_{vs} - PB'_{vs}. \tag{4.4}$$

KG_v, the overall KG for video v is finally calculated by the average of all participants' KG :

$$KG_v = \frac{1}{n_v}\sum_{s=1}^{n_v} KG_{vs}. \tag{4.5}$$

The next section outlines our approach for the extraction of in total 22 features from lecture videos including audio, linguistic, as well as a set of chosen multimodal features. An overview of our feature set is depicted in Figure 4.3. Since we are dealing with educational videos, we assume that for each data sample a video file is available, together with a PDF file of the shown presentation as well as a speech transcript.

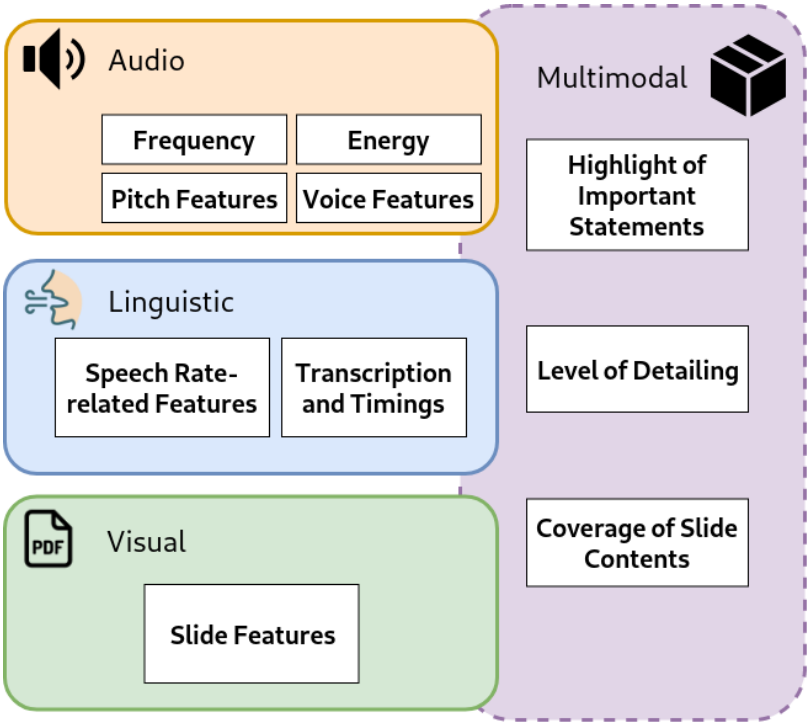


FIGURE 4.3: Overview over the feature sets extracted for the automatic assessment algorithm.

4.1.4 Multimedia Features

Audio Features

We utilize the openSMILE-toolkit [72] to extract the audio features, except for the pitch variation information. We compute these according to Hincks [107]. We selected the feature subset of the ComParE challenge (6.373 dimensions, around 70 low-level descriptors with multiple features each) and reduced it to nine features and their arithmetic mean, see Table 4.2. For our study, we selected those features that have either impacted the audio quality before (Jitter, F0 Harmonics ratio) or are very likely to influence the perceived quality of the audio (e.g., Energy, Loudness, Harmonics-to-Noise Ratio).

Eloquence Features

The extraction of eloquence features describes how the presenters articulate themselves using syllable duration and speaking rate. De Jong and Wempe's [128] Praat [21] script was used to extract these features, namely *speech rate*, *articulation rate*, and average syllable duration (*ASD*). We derive these features from the number of vowels or syllables per time interval because they indicate if the speaker is talking too fast or too slow. The video transcript also contains a lot of helpful information regarding speech quality. These features are, together with the slide content, considered in Section 4.1.5.

Feature	Description
Loudness	Sum of a simplified auditory spectrum
Modulated loudness	Sum of a simplified RASTA-filtered auditory spectrum (RelAtive Spectral TrAnsform, Hermansky et al. [105])
Root-Mean-Square energy	Square root of mean of the discrete values of the sound pressure
Jitter	Deviation from true periodicity of a presumably periodic signal
Δ Jitter	Normalized average length deviation from true periodicity of a presumably periodic signal
Shimmer	Amplitude variation of consecutive voice signal periods
Harmonicity (spectral)	Ratio between the minima and the maxima in relation to the amplitude of the maxima from a magnitude spectrum
Logarithmic Harmonics-to-Noise Ratio	Logarithmic scale of the ratio harmonic to noise component in the wave signal
Pitch Variant Quotient	Standard deviation of the pitch, which is divided by the mean of the pitch (cf. Hincks [107])

TABLE 4.2: Automatically extracted audio features.

Visual Features

For the visual content, we examine the PDF files of the presentation slides with the motivation to find whether a certain composition of visual and textual content amplifies KG. With the bash command `pdftotext` we extract text layout information from the PDF slides. The command extracts the position of each text element as well as the size of the slide. The elements of a slide are stored hierarchically, starting with the biggest text element, which contains multiple text lines, and each line consists of multiple words. We convert this information into an XHTML file. Similarly, we use the `pdftohtml` command to extract the image position and size of the slide, which we store in an XML file. The generated files are then parsed to JavaScript Object Notation (JSON) since the format is more convenient for data handling.

Based on this representation, we compute two features related to the design of the slides, which are *text ratio* and *image ratio*. They describe, how much slide space is covered by each of the modalities according to Formula 4.6.

$$TextRatio = \frac{\sum_{i=1}^{n} TextArea_i}{Area_{slide}}. \tag{4.6}$$

Also, for each file we store the mean and sample variance of the text ratio and image ratio values of all slides.

Cross-modal Features

This section presents a set of cross-modal features that model specific quality aspects of a presentation. We base them on criteria important to us humans, for instance, the way and frequency the presenter highlights important aspects on the slides. If we can capture these metrics, we can rank videos with similar content according to their presentation-quality, providing an optimal recommendation.

Highlight of Important Statements: This feature indicates how often important statements are emphasized per slide and over the complete slide set. To identify the text boxes most likely containing the critical components of a slide, we use the information from the document layout analysis, which we stored in JSON format earlier. In this procedure, we use the following natural language processing functions, which were adapted from Bird et al. [19]:

- N(): return a list of nouns from a sentence
- LEM(): return a set of lemmas from a list of words
- SYN(): return a set of synonyms from a list of words

Our assumption for the identification of *important* text is that font size is often proportional to importance. Since we do not have the font-size information for each slide, we sort the text lines by the area they cover. However, simply choosing the n largest text areas does not yield good results because there are often bullet points of similar importance that cover text areas of different sizes. So we cluster the text areas according to the following rule: For each text area starting from the biggest one, if the area difference to the next biggest text area is smaller than n% of the slide size (in our experiments 1%), add it to the cluster, otherwise create a new one. We consider all text areas in the two largest clusters to contain important statements. The text area of the title and all headlines are usually in these clusters of the highest category. For each selected text line, we first extract the sentence(s) (St_{imp}) from the respective JSON file. Then, we extract the nouns and their synonyms from the text. Finally, we lemmatize the nouns and their synonyms:

$$St = LEM(N(St_{imp}) \cup SYN(N(St_{imp}))). \tag{4.7}$$

Locating Emphasized Transcriptions: We designed this feature to capture the ability of the presenter to consider important statements shown on the slides as well as their emphasis through his or her voice. If so, there should be a corresponding local maximum in the audio signal. To get this information, we need to align the speech transcript with the audio signal in the time frame where the currently observed slide was covered. The segmentation of the video according to single slides is done manually. An automatic cut detection was not reliable enough as the difference in visual content is not high enough when changing

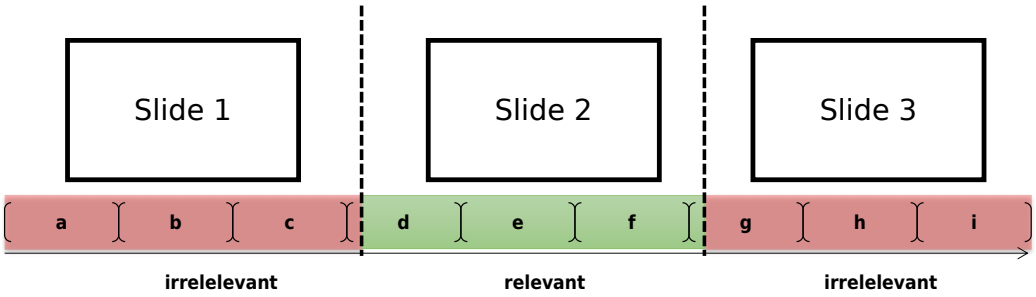


FIGURE 4.4: Visualization of the overlap between the speech transcript blocks $a, ..., i$ and the slides $1-3$. We remove a certain percentage of words from each block based on their percental overlap with the slide.

from slide to slide. For each slide and its associated time frame, we need to find the corresponding segment of the speech transcript. The given transcript is segmented into blocks of 10 seconds, which is a common interval in speech analysis (cf. Hincks [107]). A slide segment of arbitrary length can contain multiple of these blocks, and it is essential to find the correct one for each highlighted statement, see Figure 4.4. We do this by using the audio signal. We encoded the audio information via three metrics: F_0, loudness, and energy. If all of these features have a maximum at around the same timestamp, we assume the presenter emphasized the word said at that moment. Locating the exact words at that moment is done by choosing the speech transcript block whose temporal center is closest to the found maximum. We store the list of these presumed important statements in *EmphasizedTranscriptions*, and we lemmatize them again:

$$Tr = LEM(EmphasizedTranscriptions). \tag{4.8}$$

Finally, the fraction of highlighted statements is calculated as:

$$Highlight = \frac{|St \cap Tr|}{|St|}. \tag{4.9}$$

Level of Detailing: Another possible measure of quality for the presentation is the overlap of spoken text with the information on the slide. As literature from video learning suggests [27], audio and visual contents should provide complementary information rather than being overly redundant. Thus, we examine if the speaker only read the information already on the slide or if the oral explanation provides further detail, giving an appropriate amount of additional information. For this purpose, we calculate the ratio of said words to shown words on the slide.

Again, we use the speech transcript blocks from the previous section to count the number of words said during the time frame when the corresponding slide was visible. We consider each speech transcript block overlapping with the duration of the slide. We cut off blocks at the interval boundaries appropriately. If a transcript block overlapped 70% at the end of a slide and contained ten words, we would consider the first seven words

and dismiss the last three. Therefore, we store all found words in $Words_{said}$. We gather the number of words on the slide $Words_{slide}$ from the process explained in section 4.1.4. Subsequently, we calculate the **level of detailing** as the ratio of said words to words on the slide, see Formula 4.10:

$$LevelOfDetailing = \frac{|Words_{said}|}{|Words_{slide}|}. \tag{4.10}$$

Also, we calculate the mean and sample variance of these values for later usage for each video.

Coverage of Slide Contents: This metric encapsulates whether the speaker talks about all the information shown on the slide or if he or she skipped some parts. Again, the motivation is to capture if the presentation is well structured or if it appears rushed since the presenter left out some of the shown information without explanation. We can reuse the $Words_{slide}$ from the previous section. For the words said by the speaker, we reuse the already established $Words_{said}$ from the previous metric. We lemmatize the words in $Words_{slide}$ and $Words_{said}$ again, which enables an easier comparison. The **Coverage** of slide content is then calculated by us as the ratio of the number of common words in $Words_{said}$ and $Words_{slide}$ to the total number of words on the slide:

$$Coverage = \frac{|Words_{said} \cap Words_{slide}|}{|Words_{slide}|}. \tag{4.11}$$

Similarly, we calculate the mean and sample variance of the values of all slides for later usage. The complete list of manual and automatic features can be seen in Table 4.3.

4.1.5 Textual Features

This section describes an extensive feature extraction process *per lecture video* applied to the slides (PDF) and transcript (SRT) of each video of the dataset yielding 387 features. The set of textual features comprises five subcategories: *syntax*, *lexical*, *structure*, *semantics*, and *readability*. The full feature list is in Appendix B.

Syntactic Features (308)

For the extraction of the syntactic features, we used the library *Stanza*[2]. It supports 66 languages, tokenization, lemmatization, and POS tagging. In addition, it provides an interface for the Stanford CoreNLP library [164] that generates syntax trees to represent the structure of a sentence. For our experiments, we consider the full list of **word types** given by the *Universal POS tags*[3], their average frequency per sentence in each video, once for the PDF and for the SRT files, and the ratio compared to the total number of words in the sentence per word type. Since nouns and pronouns are a majority in the given types,

[2] https://stanfordnlp.github.io/stanza/
[3] https://universaldependencies.org/u/pos/)

Human-rated Quality Aspect	Automatic Features
Clear Language	Loudness avg.
Vocal Diversity	mod. Loudness avg.
Filler Words	RMS Energy avg.
Speed of Presentation	f0 avg.
Coverage of the Content	Jitter avg.
Level of Detail	Δ Jitter avg.
Highlight of imp. Content	Shimmer avg.
Summary	Harmonicity avg.
Text Design	log. HNR avg.
Image Design	PVQ avg.
Formula Design	Speech Rate
Table Design	Articulation Rate
Structure of Presentation	avg. Syllable Duration
Entry Level	Text Ratio avg.
Overall Rating	Text Ratio var.
	Image Ratio avg.
	Image Ratio var.
	Highlight of imp. Statements
	Level of Detailing avg.
	Level of Detailing var.
	Coverage of Slide Content avg.
	Coverage of Slide Content var.

TABLE 4.3: Overview of the recorded non-textual features.

we record their ratio as well. Words that can not be assigned to a class are denoted as *X* in the *Other* class. This results in 110 calculated features. Next, we extract 88 **temporal features**. As reported by Kurdi et al. [145], less complex texts prefer simpler tenses and thus can be used to measure text complexity.

We utilize XPOS annotations (also generated by *Stanza*) and a list of rules, that we derived from an English teaching website[4]. We determine and count the appearing tenses for every word of a clause. We analyze clauses individually by splitting sentences at commas or semicolons so their tenses do not interfere with each other. Lastly, we account for passive and active forms individually and also their respective frequencies compared the entirety of occurrences is considered as a feature. The next category are **phrases**, where we identify 86 different features. 14 different phrase types were considered, extracted by *Stanza's* constituency parser and defined by *Penn Treebank* [258]. "Wh-words" are interrogatives like *where, who, when*, but also *how*. Similar to the previous feature types, we compute the raw count, the ratio of each individual phrase type to the entirety of occurrences, and the average frequency per sentence. Additionally, the average number of phrases per sentence is determined. Also, 24 **other syntactic features** are extracted. According to Kurdi et al. [145], who found that tri-grams and tetra-grams influence the textual complexity, we count the average number of n-grams per sentence. Furthermore, the number of characters

[4] https://www.englisch-hilfen.de/grammar/englische_zeiten.htm

Tense	Active clause	Passive clause
Simple Present	VB/VBZ/VBP; do/does + VB	am/is/are + VBN
Present Progressive	am/is/are + VBG	is + being + VBN
Present Perfect	has/have + VBN	has/have + been + VBN
Present Perfect Progressive	has/have + been + VBG	-
Simple Past	VBD; did + VB	was/were + VBN
Past Progressive	was/were + VBG	was/were + being + VBN
Past Perfect	had + VBN	had + been + VBN
Past Perfect Progressive	had + been + VBG	-
Will Future	will/shall + VB	will/shall + be + VBN
Future Progressive	will/shall + be + VBG	-
Future Perfect	will/shall + have + VBN	-
Future Perfect Progressive	will/shall + have + been + VBG	-
Conditional Simple	would + VB	would + be + VBN
Conditional Progressive	would + be + VBG	-
Conditional Perfect	would + have + VBN	would + have + been + VBN
Conditional Perfect Progressive	would + have + been + VBG	-
Present Participle	VBG	being + VBN
Perfect Participle	having + VBN	having + been + VBN
Present Infinitive	to + VB	to + be + VBN
Perfect Infinitive	to + have + VBN	to + have + been + VBN

Table 4.4: Used tenses and their rules for active and passive clauses

per video for slides and transcripts are computed, and in a similar fashion, the number of words as well as the minimal, average, and maximal number of words per modality. Finally, the number of expressions and questions as well as their ratio to the total number of sentences is computed as a feature.

Readability (12)

Stajner et al. [247] and Brigo et al. [32] suggest readability indices to measure text complexity. Accordingly, we implement Flesch-Reading-Ease [249], its successor the Flesch-Kincaid and the Gunning-Fog Index [35], the SMOG [174], Coleman-Liau [48] as well as the Automated Readability Index (ARI) [245].

The input parameters for these indices involve, besides the number of letters, words, sentences, and syllables, the number of *long* (4+ syllables) and *difficult* words. *Difficult* words have based on Brucker's [35] definition, at least three syllables, not counting common suffixes like *es, ed, ing* while not being a name or compound word. To identify these two types of words we use the *CompoundWordSplitter* [66]).

Lexical Features (36)

The first two lexical features are related to **word frequencies**. Intending to determine whether repetitions of important words influence the quality of a learning resource, we compute word frequencies [145]. This is realized by checking if the resulting word can be found in a list of 79 672 words given by the *English Lexicon Project* [13, 67], after filtering stopwords (via *CoreNLP*) and lemmatization of plural forms. If the word exists, we gather

additional metadata from this website, namely the number of syllables, the acquisition age (cf. Section 4.1.5) of the word, and the POS tags for later usage. Lastly, we divide the frequencies of each remaining word by the number of total words. Symbols and digits are not considered in this list and receive the frequency 0. Composite words with a dash are associated with the frequencies of their respective parts. Next, we identified six features related to the age of aquisition (AoA). This feature denotes the average age a human learns a certain word. This age can vary heavily for different words and, thus, indicate its difficulty. For reference, Kuperman et al. [144] created a list of 30 000 nouns, verbs and adjectives that we extended with a list of 50 000 articles, pronouns, and inflections [2]. If an AoA rating is not present in either list, we assign the average value of 10.36 years. For our classification, we collect the earliest, average, and latest reported AoA for each word. Again, we ignore digits and symbols. For the average AoA per video, we divide the sum of all average ages by the total number of words with an AoA.

Due to their important role in the context of readability, we choose to examine the **number of syllables** even further. We focus on words with one, two, polysyllabic, and the aforementioned *difficult* words (cf. Section 4.1.5). If possible, we get the number of syllables from the *English Lexicon Project* (cf. Section 4.1.5), otherwise, the *SyllaPy* library [252]. It contains a (smaller) word list that is referenced if possible but can also compute the number of syllables for unknown ones. In theory, this method is still not a 100% accurate since the pronunciation sometimes omits syllables, but it is sufficient for our task. We compute the total number of syllables, the average number of syllables per word, the number of words containing one, two, or 3+ syllables, and the number if *difficult* words. The ratio of all these measures to the total number of words is computed to normalize the features. Finally, we want to investigate **word variations** in the text. *Stanza's* lemmatization method allows us to analyze the variety in phrasing used by the author of the video, thus indicating a less repetitive, more vivid textual content. For this purpose, we create two sets, i.e., lists without duplicates. The first list contains all word occurrences in the text that a POS tag can be assigned to, while the second one represents the result of lemmatizing all words in the first one. The resulting features for speech transcripts and slides, respectively, are the lengths of both lists and the ratio comparing their lengths to the total number of words in the respective file type.

Structural Features (24)

We extract structural features from the two input file types: presentation slides and the SRT transcript files. Naturally, we have to apply different measures for the two file types since they vary strongly in their layout.

Slides PDF For the presentation slides, we count the total number of lines in the entire presentation and the minimum, average, and maximum for (a) the number of lines per slide, (b) the number of words per slide, (c) the number of words per line, and (d) the number of letters per line. Furthermore, we count the total number of slides. We utilize

the *PyMuPDF* library to access the textual components of PDF documents. It is important to extract the content in natural reading order, see Algorithm 1.

Algorithm 1: Sort text in reading direction

```
Data: Slides
Result: Lines per slide sorted in the natural reading direction
Load slides of a presentation;
while not all slides processed do
    Read next unprocessed slide with all words objects;
    Sort objects by block_no, line_no and word_no;
    Combine objects into one that have the same block_no and only a maximum
     difference of 2 between y_0 and y_1 coordinates;
    Sort by y_1 and x_0;
    Remove all word objects with one character;
    while first word object of word objects list not reached do
        Take last non checked word object of list;
        Check the correctness of the position by examining previous word objects
         with a y_0 value smaller by a maximum of 5. If the x_0 is greater then the
         value of the current word object, position the current one beforehand;
    Store list of lines as part of a bigger list;
```

Speech transcript SRT Lines in an SRT file are enumerated sections of the subtitles with individual start and end timestamps. We extract the number of subtitles, number of sentences, and their collective display time. We also compare the display time with Ziefle's [305] reading speed of 180 words per minute to gain insight into whether it is possible to read the subtitles within the given time frame. We further record information about the subtitles by computing the minimum, average and maximum number of letters and words per sentence. We read the SRT files via Python 3's *SRT* library and concatenate the individual subtitles before reassembling the underlying sentences with the *Stanza* library.

Semantic Features (6)

Earlier approaches, as, for example, suggested by Stajner et al. [247], counted the number of possible interpretations of a word to get an idea of its complexity. Nowadays, semantic word embeddings are the state of the art to represent words numerically. We choose *sentence transformers* [220, 219] to semantically represent various parts of our textual features. From the large set of pre-trained, multilingual models we revert to the *roberta-large-nli-stsb-mean-tokens* model [235], which achieved the highest score in the "Semantic Textual Similarity" benchmark. It returns embeddings with a dimension of 1024, ignores punctuation but reacts more sensitive to changes in tenses, replacement of core nouns of a sentence, or position changes of words in a sentence. To compute the embedding representing an entire video's speech transcript as well as slide content, we first encode each sentence separately and average the results afterward, entailing three features:

embed_srt and *embed_slide* and their distance in the embedding space. To reflect the semantic distance between the individual sentences, we record each pair's average distance for both modalities. Also, we compare these two distances by capturing their difference.

User-specific Features (1)

By design, all our extracted features are independent of the user, since they are based on the educational resource alone. However, our goal of KG prediction is also influenced by the learner's cognitive capabilities as well. For instance, some users might generally obtain better learning results after watching the video. In order to investigate the influence of user identity in our experiments, we add another feature subset, the person ID (*USER* from here on). To prevent linear dependencies between these IDs, we represent them as one-hot-encoded vectors (13 dimensions).

4.1.6 Experiments and Results

This section describes the two KG prediction experiments that we conducted on all combinations of our feature categories. Figure 4.5 gives an overview over the setup. Shi et al's user study [240] yielded 111 individual learning sessions based on 13 participants that watched eight to nine videos each. The extraction of features from these sessions yields, however, only 22 unique (feature vector) samples, that is one for each video $v_i, i \in 1,..,22$. They differ only in their target variable, the KG score KG_i. However, we follow Yu et al. [293] and do not predict KG scores directly, but rather assign one of three classes. We normalize the scores by transforming them into Z-scores with $\overline{X}$=0 and σ=1. The KG classes are defined as follows: 1.) *Low* KG, if $X < \overline{X} - \frac{\sigma}{2}$; 2.) *Moderate* KG, if $\overline{X} - \frac{\sigma}{2} < X < \overline{X} + \frac{\sigma}{2}$; and 3.) *High* KG, if $X > \overline{X} + \frac{\sigma}{2}$. This results in a dataset composition of 6 low, 10 moderate and 6 high for **V22** and 40 low, 40 moderate and 31 high KG samples for **V111**. For the first set of experiments (**V22**), we try to predict the average achieved KG class per participant that saw video v_i. We establish a challenging *KG baseline V22* by estimating the performance of participant p_k on video v_i. Therefore, we average the KG scores of all other participants p_l with $l \neq k$ who saw v_i and convert it to the appropriate class afterwards, but only on videos $v \neq v_i$. Thus, this baseline has strong hints about the learning outcome of different participants that are not available to our classifiers. It achieves an accuracy of 45.45%. In our second set of experiments (**V111**) we add the person ID as a one-hot encoded vector to the respective video feature vectors to make them unique again, giving us the original 111 samples. Target variable is the recorded KG class of the learning session. Again, we derive another challenging *KG baseline V111*. To estimate the KG class that user u achieved on video i we average his/her score on the $n-1$ other videos seen by him/her and, again, convert it to the appropriate class. This baseline is also challenging (accuracy = 43.24%) because the information about the user-specific learning performance is not available to our classifiers; as mentioned above, our user-specific feature is simply the encoded person-ID.

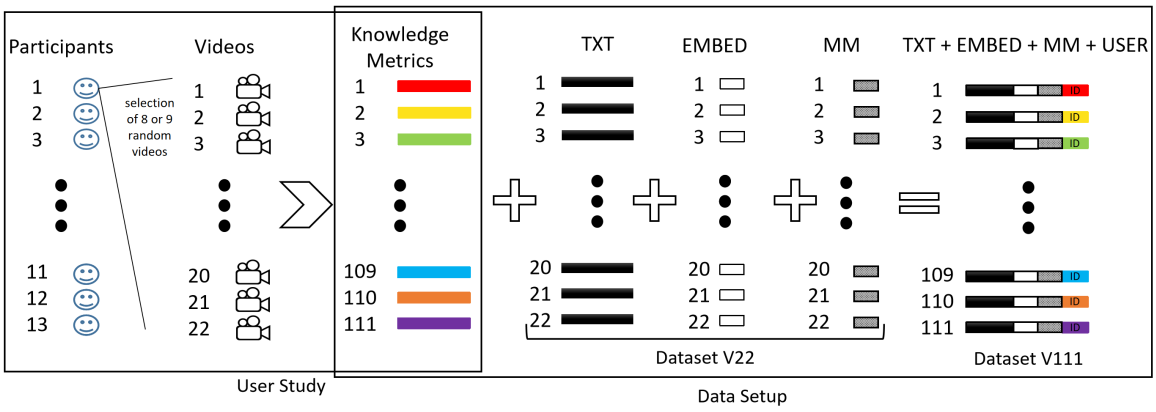


FIGURE 4.5: The workflow of our approach detailing the composition of our datasets for experiments **V22** and **V111** (best viewed in color).

Data Preprocessing

The correlation analysis of Shi et al. [240] (note: they did not attempt KG prediction) investigates the relationship of their multimodal features and KG. We want to examine if their features (*MM* from here on) allow for KG prediction, and how our suggested features (*TXT + EMBED + USER*) are suitable for this task, as separate feature sets and in combination. Consequently, we have seven feature combinations as inputs for experiments **V22** and **V111**: TXT, EMBED, MM, TXT+EMBED, TXT+MM, MM+EMBED, and all of them together TXT+MM+EMBED. For **V111** all of these categories also contain the one-hot-encoded person id (*USER*) of the respective learner. We translate and scale all features with *sklearn's* MixMaxScaler such that it is in the range between 0 and 1.

Dimension Reduction for Sentence Embeddings: Since the majority of features are (single) scalars, the high-dimensional sentence embeddings (1024) most likely outweigh the rest. Thus, we conduct a PCA (*scikit-learn*) with target dimensions of $3, 8, 16, 32$. We decide to use 16 dimensions for the sentence embeddings since they yield 93.73% explained variance for slide text, as compared to 38.26% (3 dim.), 69.29% (8 dim.) and 99.99% (32 dim.). The results for the transcripts were similar. For the final decision, whether to use 16 or 32 dimensions, we investigated the trade-off between loss of information and accuracy in the following classification. Preliminary results showed that 16 dimensions retained better results, even though they contain around 7% less information.

Data Filtering: As a last pre-processing step we remove 47 features that are zero for every sample and thus, contain no information towards the classification. The reason is that not all occurrence-based information, e.g., tenses and word types, appear in the text.
Finally, for both experiments the samples are randomly split into approximately 80% training and 20% test. For experiment **V111** we made sure that no video seen in training was used in test.

Feature Selection

Breiman [31] discusses the two categories of feature importance computations. The first category examines model parameters to identify what is most important towards the result, while the other one treats the model as a black box and compares simply how changing the input impacts the output. Recent work on KG prediction [293] utilizes Pearson correlation to estimate the most influential features, which falls into the first category. However, Breiman describes the issues related to such measures as follows: 1) analyzing what the model does assumes that the model is the right fit for the problem, 2) the amount of trust regarding these results is tied to the performance of the model, and 3) this feature analysis does not tell whether the model is biased.

Therefore, we decide to resort to a feature importance technique of the second category, namely *Drop-Column Importance*, a more computational expensive type of *Permutation Feature Importance*. The idea is that, for every feature f, to train one model from scratch by dropping f. Then, the decrease (or increase) in performance, compared to a baseline model that contains all features, shows how important f is for the process. This technique makes the approach model-agnostic, which is useful since we are investigating multiple classifiers. The implementation we used can be found on GitHub[5]. Negative importance values imply that the performance increases when the feature is not considered. Thus, we only keep features that have values ≥ 0 (**V22**: 40, **V111**: 191) and do our final run of both experiments afterwards. We keep the 13-dimensional person id vector for the feature selection process, since each bin represents one person and we want to investigate whether the models utilize information about the individual performances of the participants. We omit the PCA-transformed sentence embeddings for the importance analysis, since they are hard to interpret. The results of the feature importance analysis are discussed in Section 4.2.6.

Results and Discussion

We use four classifiers in our experiments, *NB, SMO, RF, and MLP* implemented by the *Weka* machine learning software. For each classifier (set to default hyperparameters), each feature category, and both experiments we conduct a 5-fold cross-validation and average the results per fold in terms of precision, recall, F1-score, and accuracy. Also, for each fold, a separate feature importance analysis and feature selection is conducted to avoid bias. The following two Tables 4.5 and 4.6 show the best performing combinations of classifier and feature category for the experiments **V22** and **V111**. The overall scores are macro recall, precision, and F1.

In **V22**, all models clearly outperform random guessing, but only one feature combination outperforms the challenging baseline of 45.45% regarding overall accuracy: the combination of our textual features and Shi et al's multimedia features [240] with a Random Forest approach. It was able to distinguish between all three classes. In second place,

[5] https://github.com/parrt/random-forest-importances/blob/master/src/rfpimp.py

		Low			Moderate			High			Overall			
Feature Category	Class.	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1	Acc. in %
Random Guess Baseline	-	-	-	-	-	-	-	-	-	-	-	-	-	33.33
KG Baseline V22	-	0.00	0.00	0.00	0.45	1.00	0.62	0.00	0.00	0.00	0.15	0.33	0.21	45.45
EMBED (slide)	SMO	0.10	0.20	0.13	0.45	0.80	0.57	0.00	0.00	0.00	0.18	0.33	0.23	42.0
EMBED (srt)	MLP	0.10	0.20	0.13	0.47	0.50	0.48	0.30	0.60	0.40	0.29	0.43	0.34	42.0
MM	RF	0.20	0.20	0.20	0.50	0.60	0.55	0.30	0.30	0.30	0.33	0.37	0.35	43.0
TXT	NB	0.00	0.00	0.00	0.60	0.70	0.65	0.17	0.40	0.24	0.26	0.37	0.30	41.0
MM+EMBED (slides)	RF	0.20	0.20	0.20	0.58	0.70	0.63	0.07	0.20	0.10	0.28	0.37	0.31	42.0
TXT+EMBED (both)	NB	0.00	0.00	0.00	0.58	0.80	0.67	0.17	0.40	0.24	0.25	0.40	0.30	45.0
TXT+EMBED (slide)	NB	0.00	0.00	0.00	0.60	0.80	0.69	0.17	0.40	0.24	0.26	0.40	0.31	45.0
MM+TXT	RF	0.10	0.20	0.13	0.55	0.60	0.57	0.23	0.60	0.34	0.29	0.47	0.35	**46.0**
MM+TXT+EMBED (both)	NB	0.07	0.20	0.10	0.65	0.70	0.67	0.10	0.20	0.13	0.27	0.37	0.30	42.0
MM+TXT+EMBED (slide)	NB	0.07	0.20	0.10	0.65	0.70	0.67	0.10	0.20	0.13	0.27	0.37	0.30	42.0

Table 4.5: Best results for each classifier in the **V22** experiment on the respective feature category.

our set of textual features together with the semantic sentence embeddings achieved 45% accuracy with a Naive Bayes classifier. However, they failed to detect the *Low* KG samples entirely. Even though the other results do not outperform the strong KG baseline, they are noticeably better than random guessing. In summary, **V22** indicates that a multimodal approach with a Random Forest classifier is a good approach for this problem.

		Low			Moderate			High			Overall			
Feature Category	Classifier	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1	Acc. in %
Random Guess Baseline	-	-	-	-	-	-	-	-	-	-	-	-	-	33.33
KG Baseline V111	-	0.50	0.18	0.26	0.39	0.85	0.54	0.70	0.23	0.34	0.53	0.42	0.38	43.24
EMBED (srt)+USER	MLP	0.35	0.33	0.34	0.48	0.47	0.47	0.39	0.56	0.46	0.41	0.45	0.42	**44.74**
MM+USER	NB	0.35	0.39	0.37	0.44	0.34	0.39	0.35	0.50	0.41	0.38	0.41	0.39	39.03
TXT+USER	MLP	0.36	0.29	0.32	0.31	0.34	0.32	0.36	0.35	0.35	0.34	0.33	0.33	34.66
MM+EMBED (srt)+USER	NB	0.37	0.33	0.35	0.45	0.42	0.44	0.35	0.47	0.40	0.39	0.41	0.40	39.00
TXT+EMBED (slide) + USER	MLP	0.45	0.35	0.40	0.46	0.49	0.47	0.39	0.33	0.36	0.43	0.39	0.33	40.83
MM+TXT+USER	MLP	0.24	0.26	0.25	0.48	0.45	0.47	0.30	0.39	0.34	0.34	0.37	0.35	36.06
MM+TXT+EMBED (srt)+USER	SMO	0.35	0.31	0.33	0.43	0.48	0.46	0.27	0.32	0.29	0.35	0.37	0.36	36.15

Table 4.6: Best results for each classifier in the **V111** experiment on the respective feature category. Each experiment considered the one-hot-encoded person id *USER* as an additional feature.

For **V111**, the best result of 44.74% has been achieved by the sentence embeddings generated from the speech transcript (EMBED (srt)) fed into an MLP. The model clearly outperforms random guessing, and is slightly better than the KG baseline. Second best performance was achieved by our textual features combined with the sentence embeddings extracted from the slide content (TXT+EMBED (slide)). These results indicate that focusing on textual features, in a syntactic as well as semantic manner, is beneficial for KG prediction when the individual is represented as a variable. Also, neural network based approaches scored the highest in four of the seven feature categories, perhaps due to the large number of features. The other categories fell below 40.00% accuracy. However, they remained above random guessing by means of accuracy. Additionally, as the feature importance analysis will highlight, the sentence embeddings generated by the speech transcript and slides were of high importance, so neither should be neglected for this task.

In summary, experiment **V111** suggests that semantic text features that describe the content of a MOOC video, are a better choice for this task than syntactic features that objectively describe the video. In comparison with **V22** that had a slightly stronger focus

on multimedia features that describe the objective quality of the video, this finding could be explained by the following: On one hand, to predict the user-independent (average) learning outcome of a MOOC video (as in **V22**) it is beneficial to consider multimodal features describing general quality aspects. On the other hand, the prediction of the individual KG (**V111**) depends on a combination of content-features and the preferences of the person itself. We tried to capture this personal influence with our one-hot-encoded person id feature. The results of the feature importance analysis will underline its impact on the classification result.

Feature Importance Analysis

Tables 4.7 and 4.8 show the average of the five drop-column feature importances (generated by the 5-fold cross-validation) conducted on the feature categories MM+TXT+EMBED for **V22**, and MM+TXT+EMBED+USER for **V111**.

Type	Feature	FI
MM	img_ratio_var	0,09
MM	log_HNR_avg	0,05
MM	average_syllable_duration	0,05
MM	coverage_of_slide_content_avg	0,05
MM	f0_avg	0,05
MM	PVQ_avg	0,05
MM	harmonicity_avg	0,05
MM	highlight	0,05
MM	jitter_avg	0,05
MM	rms_energy_avg	0,05
MM	shimmer_avg	0,05
MM	level_of_detailing_avg	0,05
MM	level_of_detailing_var	0,05
MM	delta_jitter_avg	0,05
MM	articulation_rate	0,05

Table 4.7: Importance of the top-10 features of the **V22** experiment.

Type	Feature	FI
TXT	ratio_VP_sli	0,05
TXT	VP_sli	0,05
USER	Person_ID_5	0,04
TXT	PP_sli	0,04
TXT	amount_main_verb_sli	0,04
TXT	sum_tok_len_tra	0,04
TXT	avg_adj_sli	0,04
TXT	amount_adj_sli	0,04
TXT	ratio_adj_sli	0,04
TXT	amount_adpos_sli	0,04
TXT	avg_VP_sli	0,04
TXT	amount_one_syl_tra	0,04
TXT	avg_adpos_sli	0,03
TXT	WHNP_sli	0,03
TXT	ratio_sim_pres_sli	0,03

Table 4.8: Importance of the top-10 features of the **V111** experiment.

The feature importance analyses of the two experiments show significant differences. In the experiment **V22**, the important features are dominated by the multimedia features [240]. From the 40 features yielding a feature importance ≥ 0 only 14 were of the textual category. This is reflected in Table 4.5, where MM+TXT achieved the best performance. In experiment **V111** the textual features from our approach obtain the highest importance scores, with a slightly stronger focus on the slide content ("_sli" suffix). Out of the 191 most important features with a value ≥ 0 the first multimedia feature has rank 50. Rank 3 is of type *USER* highlighting the importance of this bin in the 13-dimensional one-hot-encoded vector. This suggests that our models identified that this learner's individual performance gave hints about the eventual learning outcome in the other videos he or she saw.

In summary, the results of the feature importance analyses do not favor a certain modality. This suggests that it is beneficial to follow a workflow of our approach, that is to initially consider a broad range of features and assess their importance for the classification. Focusing on a single modality from the start may not yield optimal results as the impact of the selected features may vary heavily depending on the target scenario.

4.1.7 Summary

In this section, we have investigated whether we can predict KG for MOOC videos based on their content. For this, we have presented an exhaustive multimodal feature analysis and analyzed the individual and combined impact of these modalities for the task of KG prediction. However, as the sample size is relatively small, the implications of these results are limited. In the next section, we utilize data of a significantly larger user study which is conducted in an informal learning setting.

4.2 Predicting Knowledge Gain During Web Search

4.2.1 Motivation

Investigating the learning process in informal settings is important because it helps us understand how people learn in contexts that are outside of traditional formal educational institutions, such as schools and universities. These processes are ubiquitous, since people engage in informal learning activities throughout their entire lives. As pointed out by Johnson and Majewska [127], informal learning can have an holistic impact on learners, influencing affective, cognitive, and social aspects [10]. Also, this learning type adapts to the needs and interests of individual students, especially to the work pace. This allows, for example, learners with a lower pace to not feel rushed in contrast to a more rigid, time-pressured formal curriculum [178]. Finally, with the rising popularity of multimedia content, such as video tutorials or lectures, comes the need for a better understanding of informal learning. Yet, as we will highlight in Section 4.2.2, the users' interactions, navigation behavior, and consequently learning outcome, have not been researched extensively.

Previous work has studied the relationship between learning progress and text content or behavioral features collected from search sessions. For instance, Collins-Thompson et al. [49] studied the influence of distinct query types on knowledge gain, and found that intrinsically diverse queries are correlated with knowledge gain. On the other hand, Syed and Collins-Thompson [250] explored a range of text and resource-based features and their impact on short-term and long-term learning outcome, but did not investigate multimedia content. Closing this research gap, however, is challenging due to the fact that available datasets are scarce and usually do not cover the entire range of modalities.

In this section, we contribute to this research field by first gathering an extensive dataset based on an study conducted in an informal learning setting (Section 4.2.3). This study, again, recorded the pre- and post-knowledge states of the participants through multiple-choice questionnaires. Also, participants' Web search sessions were recorded, including query and navigation logs. Afterward, Section 4.2.4 outlines the implementation of a data processing pipeline. We analyzed all visited Web pages to gather a set of features regarding consumed multimedia content, e.g., document layout, image size and type. This novel feature set allows us to investigate the role of multimedia features for knowledge gain prediction in Section 4.2.5. Therefore, we train a supervised learning model (random forest) to predict knowledge gain based on text and multimedia features. Experimental results demonstrate the feasibility of the approach.

4.2.2 Related Work

Educational Material Datasets

Research on Search as Learning relies on study-based data that has to capture (a) search behavior of various nature and (b) knowledge metrics of users (through pre- and posttests). As they have to be conducted in controlled environments, their design and execution is costly. First, they have to reflect realistic scenarios in order to be indicative for real-life applications, i.e., omit a restriction of web pages or intrusive recording equipment. Second, assembling meaningful questionnaires for pre- and post-tests, for instance, is a challenging task, as topic domain and item difficulty have to be well calibrated. Lastly, the logging process itself is non-trivial since available software, to our knowledge, usually only covers part of the features of interest. Therefore handcrafted, non-intrusive logging mechanisms need to be implemented manually.

To the best of our knowledge, there are currently only two SAL-focused datasets available: Proaño-Ríos and González-Ibáñez[216] provide a set of 83 expert-generated learning paths on a diversity of topics. Each expert assembles a set of three web resources useful towards a certain learning goal, including a justification of their choice. However, data on real-world user behavior is not included. Gadijaru et al. [78] present a dataset comprised of 420 crowdsourced learning sessions, investigating the information needs on the search behavior and KG of users. Our dataset improves their contribution by presenting data gathered in a controlled lab study and it captures behavioral, resource, and gaze data. Other resources focus on either search or learning: (1) *Search Focus* – Datasets for the conception and optimization of search systems provide the basis for improved automatic analysis of queries [90, 89], identification of user tasks [273, 92], the influence of found resources on the user's viewpoints [136], and novel interaction methods such as conversational search [211]; (2) *Learning focus* – Datasets from the educational domain often explore recommendation tasks [272], provide data on user behavior in restricted learning environments [83] or specific instructional practices [281]. Finally, there is an active area of research on predicting the memorability of visual resources [122, 46] and the impact of resource modality on learning success [108]. The scope of these datasets is usually limited to a single feature type; none of them collects user behavior information in a realistic and open, learning-related Web search scenario.

Prediction of Knowledge Gain

Previous work has studied the relationship between learning progress and observable features in a search session. By matching the learning tasks into different learning stages of Anderson and Krathwohl's taxonomy [6], Jansen et al. studied the correlation between search behaviors of 72 participants and their learning stage [123]. They showed that information searching is a learning process with unique searching characteristics corresponding to particular learning levels. Zhang et al. [299] explored using search

behavior as an indicator for the domain knowledge level of a user. Through a small study ($n = 35$), they identified features such as the average query length or the rank of documents consumed from the search results as being predictive. Karanam et al. [133] conduct a study that gave insight into query reformulation techniques utilized by different age groups, revealing that older participants have more trouble narrowing down the search path towards the intended information. Cole et al. [47] observed that behavioral patterns provide reliable indicators about the domain knowledge of a user, even if the actual content or topics of queries and documents are disregarded entirely. Eickhoff et al. [63] investigated the correlation between a number of features of the search session as well as the Search Engine Result Pages (SERPs) with learning needs related to either procedural or declarative knowledge. Collins-Thompson et al. [49] studied the influence of distinct query types on KG, finding that intrinsically diverse queries lead to increased KG. Moraes et al.'s [187] work compared the learning outcome of instructor-designed learning videos against three instances of search ("single-user", "search as support tool", "collaborative search") in order to find the most efficient approach for their learning scenario. Hagen et al. [91] revealed that query terms can be learned while searching and reading through investigating the relation between the writing behavior and the exploratory search pattern of writers. Vakkari [266] provided a structured survey of features indicating learning needs as well as user knowledge and KG throughout the search process.Syed and Collins-Thompson [251] explored the possibility of using regression models and features extracted from user accessed document content to predict user knowledge change on vocabulary learning tasks [250]. Gwizdka et al. [85] proposed to assess learning outcomes in search environments by correlating individual search behaviors with corresponding eye-tracking measures. Gadiraju et al. [78] described the use of knowledge tests to calibrate the knowledge of users before and after their search sessions, quantifying their KG, and investigated the impact of search intent and search behavior on KG of users. In a follow-up work, Yu et al. [293] proposed to use user interaction features to build classification models to predict user knowledge state and KG in search sessions. Bhattacharya et al. [18] investigated the relationship between users' search and eye gaze behaviors and their learning performance. In a recent work, Roy et al. [224] investigated at which time during a search session learning occurred, and found that the learning curve is largely influenced by a user's prior knowledge on the searched topic. Kalyani et al. [131] explored this direction further by designing search tasks that fit into the different learning stages of the revised Bloom's taxonomy. Through knowledge tests before and after each search session, they found significant impact of the learning stage on a user's search behavior and KG. Liu et al. [155] adopted mind maps to capture user's knowledge change process and hence identified four types of knowledge change styles.

The aforementioned works consider a limited set of features. Just recently, the attention shifted towards the analysis of multimedia features, such as images and videos embedded in Web documents and the user's interactions with them. Yuan et al. [294] conducted a

study on realistic Web pages to train a neural network that predicts how easy it is for a user to find a specific multimedia object. Martinez-Maldonado et al. [168] investigated in their work how multimodal data convey information by organizing it into meaningful layers. They claim it is naive to expect that simply rendering information visually, a learner can make sense of them. The framework shown in this chapter provides a way of exploring the influence of multimodal content design and content directly to the learning outcome.

4.2.3 User Study

Participants and Task

The participants (N=114, 95 female, μ_{age}=22.88, σ_{age}=2.93), German speaking university students from different majors were asked to solve a realistic learning task to understand the principles of thunderstorms and lightning. This topic has been used before to study multimedia learning (e.g., [171, 230]) and has been chosen since it requires learners to gain knowledge about different physical and meteorological concepts and their interplay, i.e., they need to learn about causal chains of events and acquire declarative as well as procedural knowledge [6].The acquisition of information about such task a can be accomplished through studying different representation formats, such as text, pictures, videos, or combinations of those. This circumstance is beneficial for our goal to get a general idea about optimal multimedia learning resource design, especially in SAL scenarios.

Procedure and Measurements

The experiment consisted of an online and a laboratory part. In the online part, which had to be completed around one week before the lab appointment, participants had to respond for the first time to the 10-item multiple-choice and 4-item transfer knowledge test based on previous work [230]. Further, participants worked on questionnaires regarding their achievement motivation [65] and their Web-specific epistemic justification beliefs [29]. At the lab appointment, participants first completed tests assessing their reading comprehension [231] and working memory capacity [51]. The participants were asked to write a first essay (t1) about the topic of the formation of thunderstorms and lightning. Afterward, they were instructed to learn about this specific topic by searching the Web in a self-regulated manner. Participants were informed about the time limit of max. 30 minutes for their web search, and that they could also end the task early. They were encouraged to use every kind of Web resource they would like. After the learning phase, they were asked to write again everything what they now knew about the topic in a free essay (t2) format. Lastly, they were asked to answer the multiple-choice questionnaires (t2) again followed by a questionnaire assessing task engagement [170] and cognitive reflection tasks [77].

Technical environment

All search and learning activities of participants were conducted within a tracking framework consisting of two layers. The first layer was the SMI (SensoMotoric Instruments) ExperimentCenter (3.7) software environment that enabled us to track participants' eye movements as well as their activities during Web search in the formats of screen recordings and navigation log files. ExperimentCenter offered a default version of Mozilla FireFox (ESR 45.6.0) for tracking Web activities, which was started within the ExperimentCenter environment. In addition, by utilizing browser plugins we gathered resources of all visited HTML files as well as tracking of navigation and interaction data (e.g., mouse movements) that we adapted from [255]. For the second layer we utilized browser plugins to gather resources of all visited HTML files and adapted Talibi et al's method [255] to track navigation and interaction data (e.g., mouse movements). For more details we refer to [202].

Data Logging Challenges

As mentioned above, we aimed to collect as much data as possible during the user study to ensure that we would be able to reconstruct the exact information a participant has seen. This objective poses several challenges: First, due to the ever changing nature of online content it is necessary to save a snapshot of every visited website at the time of the actual experiment. Solely saving URLs in order to later revisit the page could lead to different results. It is therefore necessary to save the respective HTML and CSS files for every seen website during the experiment. Second, the given hardware and software determines the way a given website is displayed for the participant, e.g., monitor size and orientation, screen resolution, browser. Therefore, simply reloading the Web page in post-processing will most likely show a different fraction of the website as seen by the participant. To make sure we have the same viewpoint as the participant we created a screen recording over the full duration of the learning session that can be analyzed later. Third, a naive way of logging the browser history will not suffice, since clicking a link does not necessarily mean the respective website is visited from this moment on. Contrary, it is common that learners open multiple links from a SERP via *Open link in new tab* and go through them later, rather than navigating back and forth between browser tabs (SERP and individual websites). This behavior can be tracked by additionally logging detailed information of mouse movements and browser signals (e.g., *new tab got in focus*). Finally, all these problems must be addressed while providing a "natural" search environment that does not distract the users with restrictions regarding their usual browsing behavior or heavy computational load on the browser or computer in general.

4.2.4 Dataset Description

The following section describes the information per user provided by the individual data subsets. Apart from the screen recordings and HTML data, which we cannot

make publicly available due to licensing restriction, all dataset parts are available under https://doi.org/10.25835/0062363.

Resource Data - Screen Recordings

The screen recordings show the entire search process of the participants over the duration of the study and are aligned with the provided logs (Section 4.2.4) and HTML data (Section 4.2.4). The screen recording's video format is MP4, and they have been recorded with a resolution of $1280x720$ at 30 frames per second. The audio track is not included. They are not longer than 30 minutes and start at the point in time the learning session starts. We manually cut the start of the video that showed the participants entering their session IDs.

Resource Data - HTML

Since online content is not persistent, to achieve our goal to enable research on the actual data seen by the participants, we decided to record the content of each visited website, including but not limited to *html, *css, *js, and image files. Due to technical difficulties this process was not entirely successful, forcing us to fill in the gaps at later points in time. In detail, we managed to capture 87.9% of the data at the time of the study, another 4% in March 2020 and finally, another 2.5% in September 2021. For the remaining 5.6% (181 URL) we were not able to record any data. With very few exceptions (a few websites that are not available anymore) these were search engine result pages from Google and YouTube that do not contain any learning relevant information, and when crawled at a later point in time, differ strongly from the original. For these two reasons we decided to exclude them from the dataset. For full transparency we disclosed the date of acquisition in the provided timeline.

Behavioral Data - Browsing Timeline

Each participant's browsing log is represented by one tab separated value (TSV) file as outlined by Table 4.9, chronologically displaying the visited websites with a timestamp in seconds passed since the start of the session. Additionally, we disclose the path to the respective HTML files and their acquisition date as mentioned in Section 4.2.4.

p_id	timestamp	url	html_files	date_of_acquisition

Table 4.9: The fields (columns) in the timeline file associating each displayed web resource with a directory of HTML files and its date of acquisition.

Behavioral Data - Gaze

As mentioned in Section 4.2.3, we used an eye-tracker to record the learner's eye movements over the course of the Web search session. We exported the gaze information from the eye-tracking software as raw data and separated the fixations and saccades with an

I-DT algorithm [227] by marking the entries via the $fixation \in \{0,1\}$ flag. Please note that the y-coordinates are relative to the entire website and not the viewport, i.e., values larger than 1080 are possible. Further, the data can contain incorrect entries originating from tracking errors (e.g., negative values). Similar to the data subsets introduced in the sections before, this data subset contains one TSV file for each participant, chronologically listing the coordinates of the left and right pupil with millisecond precision. Additionally, the URL visible at that point in time is displayed (Table 4.10). This data subset allows for further experiments on website examination behavior and their influence on the learning process.

p_id	timestamp	left_x	left_y	right_x	right_y	url	fixation

TABLE 4.10: The fields (columns) in the gaze data files, chronologically displaying the gaze coordinates for each eye.

Behavioral Data - Browsing Events

The investigation of the Web search behavior requires detailed logs about the learners' interactions with a website, going beyond logging what type of resources they visited. We recorded over 1 million user interaction events of 11 different types. The *focus*, *blur*, and *beforeunload* describe whether a website has come into focus, lost focus, or is about to be closed. The *resize* event tracks if a participant chose to resize the current browser window and captures the resulting window size in the value column encoded as pixel sizes $x|y$. Similarly, if a learner scrolled on a website, a *scroll* event is triggered and we log the scroll distance in vertical and horizontal direction in the value column as $vertical|horizontal$. The *mousemove* event tracks the learners' mouse movements by logging x and y coordinates in the respective columns. Mouse clicks were captured in the *click* event, tracking their location (x and y columns) and the clicked HTML element in the target column as XPath. We recorded keypresses in the *keypressed* event, but omitted recording the key values due to privacy reasons in case the learner chose to login somewhere during the session. However, the chosen queries are available in the URLs of the respective search engines. Lastly, we captured *copy* and *paste* events and, if available, the intended target elements. The structure of the data record per event is displayed in Table 4.11.

p_id	timestamp	track_id	type	value	x	y	target	url

TABLE 4.11: The columns in the event data files for each participant chronologically displaying the browsing interaction events.

Behavioral Data - Browsing Tracks

The browser tracking tool associates events to websites by means of *tracks*. Upon navigating to a website, a track is created and exists until the user navigates somewhere else within the same browser tab or closes it. This setup is geared towards realistic search sessions with multiple concurrent tabs. For each track, our dataset contains the time of creating the

track, URL, and title of the website, as well as the viewport dimensions. Additionally, the data contains the lifetime of the track, as well as the time the track was active, i.e., it was displayed to the user in the active browser tab.

p_id	timestamp	track_id	url	title
viewport width	viewport height	time stay	time active	

TABLE 4.12: The columns in the track data files, capturing information such as URL and active time for a visited website.

Knowledge Data and Questionnaires

As mentioned in Section 4.2.3, we measured the knowledge state of learners at multiple points in time. Additionally, through questionnaires and tests, we captured cognitive abilities and assessments of participants across the study. Thereby, several sub datasets were generated for which we provide the documentations with explanations of measured variables and, if possible, the original German items. This section will give brief explanations of these files, while detailed documentation can be found in the dataset.
demo_knowledge_sum.csv: This file contains demographic information of participants and the summary of the knowledge-related scores (multiple-choice, essay) and cognitive abilities (working memory capacity, reading comprehension, cognitive reflection) (Table 4.13). Reading comprehension was measured through a standardized German screening instrument for adolescents and young adults. Working memory capacity was measured through a reading span task. More detailed information on this topic is provided by Pardi et al. [208].

Feature	Description
p_id	Participant ID
d_sex	1= female; 2 male
d_age	Age of participant
d_field	Field of study
d_no_sem	Number of semesters
d_lang	Mother tongue
k_mc_sum_t1	# of correct mc questions before search
k_mc_sum_t2	# of correct mc questions after search
kg_mc	Knowledge gain multiple choice questions
essay_C1	# of correct concepts before search
essay_C2	# of correct concepts after search
kg_essay	Knowledge gain essays
LGVT_speed	# of words read
LGVT_core	Points for correctly solved sentences
WMC_Recalls	# of correctly recalled sets
WMC_Sentence	# of correctly solved sentences
CRT_sum	# of correctly solved cognitive reflection tasks

TABLE 4.13: The columns in the demo_knowledge_sum data files. One participant per row.

mc_data.csv: This file contains the scores for all multiple-choice questions answered by participants before the lab session (t1) and after the search in the lab (t2). Includes also the confidence rating of participants for each question and the information if the answer was guessed.
essay_data.csv: This file contains the raw essays written by the participants before (t1) and after the search (t2).
internet_specific_epistemic_justification.csv: This file contains the Web-specific epistemic justification measurements based on a translated version of [29].
selfassessment_data.csv: To measure participant's self-assessed performance on the knowledge tests, we used both global self-assessment (estimated numbers of items answered correctly, estimated placement as compared to others and perceived ability to explain the concepts of the learning topic) as well as local on-item confidence rating, indicating how confident participants were that their given answer was correct.
CRT_data.csv: To measure an individual's tendency for cognitive reflection, participants worked on five items of the cognitive reflection task (CRT [77] translated into German. Within the dual-process model of reasoning, there is a distinction between faster responses with little deliberation and slower and more reflective responses. Solving more of the CRT items shows a higher disposition for the latter one, i.e., reflective cognition.
achievement_data.csv: We used the German version of the achievement motives scale [65] to measure hope of success (HS) and fear of failure (FF). This scale contains 10 items that are rated on a scale from 1 to 4, assessing those two dimensions. The achievement motive of an individual describes the general tendency to approach or avoid success in an evaluative situation. The HS score is associated with a range of variables beneficial for learning success, such as, performance in reasoning, persistence, or task enjoyment.
dssq_data.csv: The Dundee Stress State Questionnaire [170] measures subjective states in a performance context. Participants had to indicate their agreement to 7 items with labeled endpoints immediately after the learning phase. The mean score of those items indicates an individual's self-reported task engagement during the learning phase. Differences in task engagement can act as a moderator for task performance.

4.2.5 Framework

In this next section, we describe our *Multimedia Feature Extraction Framework* that works with (parts of) the data described above. Its purpose is to identify the part of the accumulated data that the user actually saw and that is used for learning rather than for navigational or exploratory reasons. This way we ensure that the correlation analysis only considers features that actually may have influenced the participants' KG. We describe two sets of features, namely visual features and textual features. We also explain the challenges related to the acquisition of (potentially) meaningful features as well as the applied feature extraction methods. The output per user of the data logging according to section 4.2.3 is as follows:

1. screen recording (MPEG-4 video format (*.mp4))
2. timeline of visited websites
3. HTML and CSS files of every visited website

To reconstruct the visited websites we decided against utilizing the crawled multimedia information (images and videos) by filtering for the respective extensions, since we found this method is not sufficiently accurate. Instead, we exploit the screen recordings and segment them according to the corrected timeline of the visited Web pages. An overview of the framework is displayed in Figure 4.6.

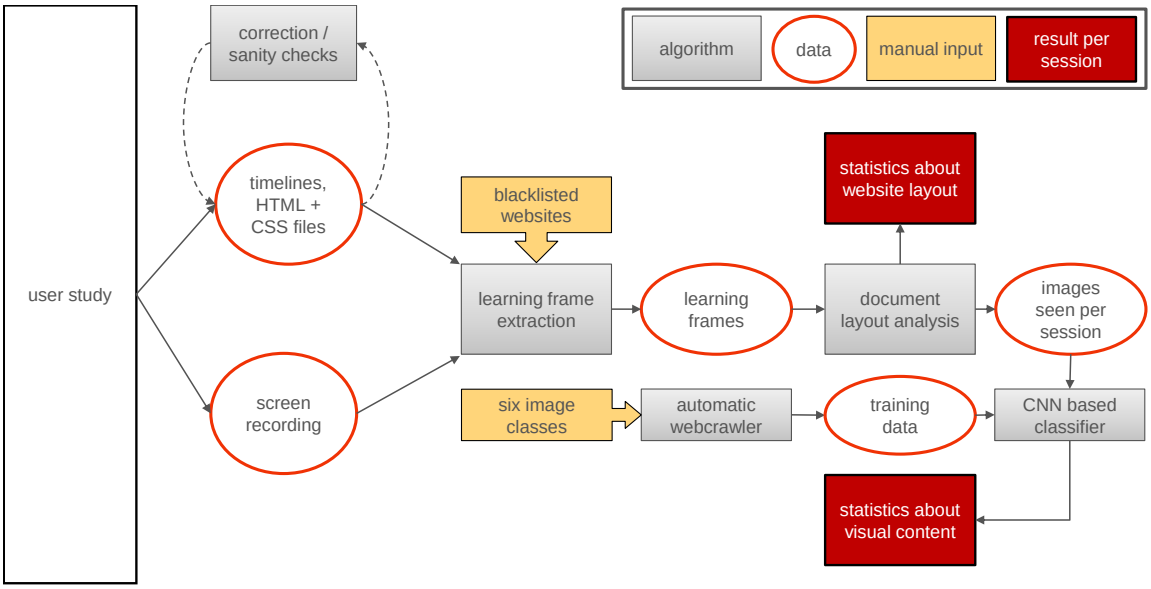


FIGURE 4.6: The overview of the Multimedia Feature Extraction Framework illustrates the process of result generation for *Document Layout Analysis* and *Image Type Classification*. The only manual input is the list of blacklisted websites and the set of image classes. The results per session (red boxes) are the input for our correlation analysis and knowledge state prediction (section 4.2.6).

The *correction* of the timeline of visited websites consists of sorting them by the order they got in focus, rather than sorting the events by the time a browser tab was opened. In this way, we circumvent the problem of participants opening multiple links from the search result page at once in a new tab, described in section 4.2.3. As shown in Figure 4.6, the next step separates the total number of F video frames into L learning relevant and N navigation related frames, with $N + L = F$. We extract a frame every second of the video ($|F| = 173\,787$), but only kept those where the participant spent time on websites related to learning. We are aware that the distinction between relevant and irrelevant is not trivial and hard to generalize, therefore we excluded only three websites. First, we excluded *Google* search page, since it was the only search engine used in all sessions and even though it occasionally provides preview information, we did not want to skew our results towards the design of the Google result page. The second excluded website is *TripAdvisor*,

which was used by one participant to browse for free time activities at the end of his/her session. Third, we blacklisted URLs containing *adblock*, which usually appear when a website asks the user to disable their Adblocker in order to access their content. Very few participants struggled with Adblocker settings in their attempt to access website contents. Obviously, this set of excluded websites differs for other use cases, since websites like TripAdvisor might provide learning relevant information for other topics. This procedure resulted in a total of 119 164 (average: 1268 frames per session) learning relevant frames, that have to be segmented into visual, textual, and background information as described in the next section.

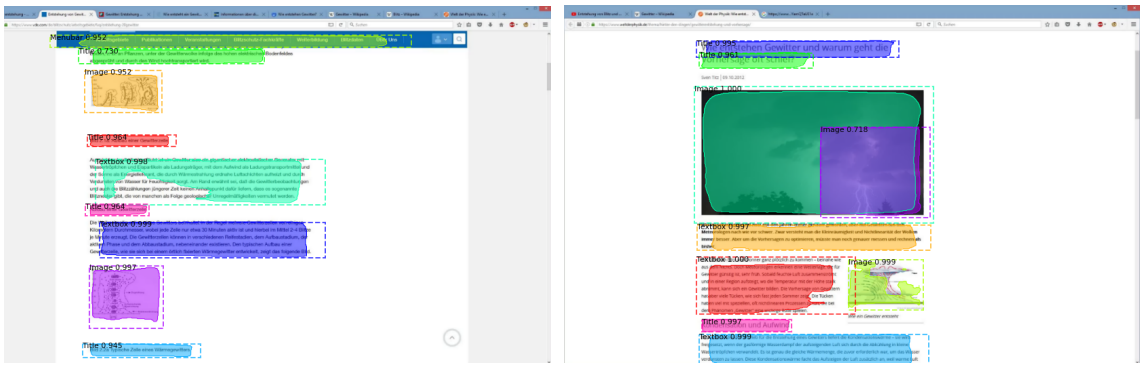
Document Layout Analysis

The goal of this step is to automatically divide each (learning relevant) frame $l \in L$ into coherent regions in order to describe the structure of the page in form of percentages (e.g., 15% images, 25% text, 5% and 55% background). Additionally, the regions should be classified according their content, e.g., image, text, menu etc. This procedure is crucial for the content analysis later but also challenging, since the layout and design of the websites varies heavily. To address this challenge, we rely on a state of the art deep learning approach to segment each frame. We utilize the Mask R-CNN [1] network architecture, originally implemented for instance (object) segmentation, and fine-tune pre-trained weights. Mask R-CNNs have excellent capabilities to adapt the given weights in order to solve related segmentation tasks (for example, color splash or segmentation of aerial images or microscopy imagery). Therefore, we annotated 300 randomly chosen frames extracted in our user study. Our selected tool is the browser-based "VGG Image Annotator", which we fed with a set of six classes, defined as follows:

1. Images/Frames: All types of still images without size constraints, from small thumbnails to fullscreen video frames;
2. Text: Any continuous text paragraph or block that is not part of a headline or button label;
3. Content list: Enumerations that contain content related information, like table of contents or bullet point lists;
4. Heading: Any headlines or titles that divide the page into sections;
5. Menu bar: Buttons or lists of buttons displayed on the page for navigational purposes;
6. Background: Everything that does not fit into the five other classes described above.

These classes are supposed to reflect the core parts of a website. The JSON style output of the manual annotations was then split into 90% training and 10% test data, which we used to fine-tune the fully-connected layers after the pre-trained bounding box detection (i.e., network heads) for 30 epochs with a learning rate of $lr = 0.001$. This option

is predefined by creating the model with parameter *layers* = "heads" and subsequently only retrains the region proposal network (RPN), the classifier and the mask heads. The resulting network is able to segment our screen recording frames appropriately. An example output is depicted in Figure 4.7a.



(A) Example result of the document layout analysis. (B) Another example showing the image-in-image effect.

FIGURE 4.7: Two example outputs of the Document Layout Analysis.

Occasionally, the neural network detects overlapping bounding boxes, which we filter out, if their Intersection over Union (IoU) is larger than 80%, see Figure 4.7b. The performance of the classifier is shown in Table 4.14. The first column denotes the confidence threshold that we assigned to the detector, forcing it to only keep boxes *above* this value. Thus, a lower value lead to a higher recall of the boxes while decreasing the precision, and vice versa. For the following experiments we considered the implementation with the highest threshold (0.9) to prevent an abundance of false positive bounding boxes.

Confidence	mean AP	mean Prec.	mean Rec.	f1-Score
0.4	0.905	0.874	0.937	0.905
0.5	0.905	0.878	0.937	0.907
0.6	0.903	0.894	0.935	0.914
0.7	0.882	0.907	0.920	0.914
0.8	0.866	0.930	0.904	0.917
0.9	0.836	0.945	0.875	0.909

TABLE 4.14: Performance of the CNN-based Document Layout Analysis model comparing different confidence thresholds of the detector.

With this document layout detector, we can collect all the components that comprise the learning content seen by all participants. However, we derive one more feature from the layout analysis: the average size (in pixels) of the seen image $\overline{imgsize}$, see Formulas 4.12 and 4.13. This information is not encoded in the document layout information $dla_i, i \in L$ itself. A website consisting of 10% images might contain five small images or a single large one. Another merit of this feature is its ability to also indirectly measure the watch time of videos, since it is difficult to measure this (simple) feature directly with satisfactory

accuracy. It neglects, for instance, embedded videos on websites other than YouTube and navigational time used on the platform before a video has been found. However, for completeness' sake, we investigate *watchtime* in our experiments as well.

$$\overline{imgsize}_i = \frac{dla_i[\text{'Images/Frames'}]}{n},\ n = \text{num of images in } i,\ i \in L \tag{4.12}$$

Then, these results per frame were added up and divided by the number of learning frames $|L|$ to get results per participant.

$$dla_p = \{\sum_{i=1}^{|L|} \frac{\overline{imgsize}_i}{|L|}, \sum_{i=1}^{|L|} \frac{dla_i}{|L|}\} \tag{4.13}$$

We identified a total of 755 756 bounding boxes that belonged to the class "Images/Frames", which has around five samples per frame on average. This appears to be a lot at first, but has a simple explanation. Every (website) frame that is recorded when watching a (non-maximized) YouTube video contains ten thumbnails of other recommended videos. In order to not skew the results heavily towards this large number of irrelevant images, we filtered them out. A threshold of 100x100 pixels (full image resolution was 1280x800) was applied and the remaining samples will be further examined regarding their shown content.

Image Type Classification

In this section, we demonstrate how the results of document layout analysis can be leveraged to analyze which kind of web page content was seen by the participants. The identification of the displayed content enables us to set the type of information in relation to the knowledge states and KG. In detail, we are interested in the displayed **image type**. To the best of our knowledge, there is no comprehensive and non task-specific taxonomy of image types to be directly applicable here. We focused on covering all topic-relevant types of images in order to learn which type of imagery a learner saw, when searching for the formation of thunderstorms (e.g., weather maps, infographics, real life imagery). As a result, our set of image type classes consisted of: Infographics, Indoor, Maps, Outdoor, Technical Drawings, and Information Visualization.

The definitions of the classes, the related queries to acquire training data through Web images search, and the class distributions are presented in Table 4.15. As shown in Fig. 4.6, this set of classes was the second manual input in our framework. They were fed into our automatic Web crawler via Python's Selenium library that gathered training samples from a Google image search. While this does not always guarantee strong, correct labels, we tried to exploit the fact that image search engines can provide a large diversity of images, which is useful for the training of deep learning classifiers. For this purpose, we added random time intervals to our queries (e.g., for Google "president of the USA after:2008-01-01 before:2015-01-01"), allowing us to gather samples from over 20 years of imagery. By creating a database of image hashes we ensured to not download any

queries	image class	content	#samples
Infographic	Infographic	Workflows or procedures visualized in a single, vivid image. Usually contains arrows.	1 316
Indoor Photography Interior Photography -car	Indoor	Indoor shots, sometimes containing persons that moderate educational content, but also in advertisements.	2 877
Map Weather Map	Map	Normal maps or weather maps.	2 878
Outdoor Photography Nature Photography Outdoor Photography Night Nature Photography Night	Outdoor Imagery	Real life outdoor shots depicting realistic circumstances.	5 913
Technical Drawing Schematic Drawing	Technical Drawing	Information depicted in form of drawings. Related to Infographics, but simpler.	3 060
Diagram Excel Chart Tex	Information Visualization	Any type of visual information that does not really fit any specific type or is a composite of multiple types.	2 729

TABLE 4.15: The automatically crawled training dataset for our image type classifier with a total size of 18 773 samples. Left column shows the set of queries used to crawl them.

duplicates. This however, led to a certain bias in the data since some types of images are less represented in the Web, for instance "Infographic" compared to "Outdoor Images".

The implementation was done in Keras, using a MobileNet [116] architecture. We used Stochastic Gradient Decent optimizer with default settings, categorical cross-entropy loss and trained for 100 epochs. Since we assured that our crawler did not produce duplicates, we omitted data augmentation techniques. The reason is that the dataset already contained visually very diverse samples and also weakly labelled data. Keeping the original class distribution we divided the above mentioned dataset into 90% training and 10% test. The latter one has a size of 1 876 samples.

To evaluate the performance of our automatic web crawler and ensure that we can rely on the outputs of the image type classification we instructed three people (two co-authors and one student assistant) to manually annotate this test set. They were tasked to either assign one of the six image types to a sample or, if an assignment was not possible, label it as a *bad* example. Afterwards, a majority vote decided on the final label for the respective image, or, if two *bad* annotations were given, to remove the image from the test set. This lead to 133 samples being removed resulting in a final test set size of 1 743. The full dataset including training and inference scripts are publicly available at (removed due do double blind submission).

The inter-coder agreement has been evaluated using Krippendorff's alpha [141] and yielded a value of $\alpha = 0.85$ (across all annotators, samples, and classes). Table 4.16 reports recall, precision, and f1 score of the (automatically assigned) labels according to our Web search compared to our ground truth labels based on the human annotations. It stands out that *Information Visualization* scored lowest on precision, which is reasonable considering the fact that it is comprised of visually diverse samples which might also share similarities with the other classes. Similarly, *Infographic* had a low recall which we assume is due to its visual versatility and similarities to the *Information Visualization* class. Finally, the

Class	Information Visualization	Indoor	Infographic	Map	Outdoor Imagery	Technical Drawing
Precision	75.6%	95.6%	96.5%	97.9%	97.5%	96.5%
Recall	94.4%	93.6%	73.5%	97.9%	97.5%	96.5%
f1-Score	84.0%	94.6%	83.5%	97.9%	98.2%	94.7%
#Samples	161	280	151	285	607	259

TABLE 4.16: Comparison of the automatically generated labels with the annotations of the three volunteers, which were used to derive ground-truth data in the experiments, and the resulting number of samples per class in the test set. Results are given in precision, recall, and f1 score.

Class	Information Visualization	Indoor	Infographic	Map	Outdoor Imagery	Technical Drawing
Precision	71.9%	86.9%	80.9%	96.1%	91.3%	82.9%
Recall	79.5%	80.4%	75.5%	86.0%	94.6%	90.0%
f1-Score	75.5%	78.1%	83.5%	90.7%	92.9%	86.3%
#Samples	161	280	151	285	607	259

TABLE 4.17: Performance of the image type classifier according to precision, recall and f1 score. The shown values correspond to an accuracy of 87.15%.

performance of our image type classifier was tested on these images and the classification results are shown in Table 4.17.

The results show that the neural network achieves a very good accuracy on the test set. The classes that achieve the lowest performance are *Infographic* and our fallback class, *Information Visualization*, with a respective precision of 71.9% and 80.9%. This is explainable, since they are the most visually diverse. The accuracy of 87.15% is considered to be sufficiently good as a basis for the correlation analysis. The classification of the images found by the *Document Layout Analysis* provides us with detailed information which image types were seen during a learning session. Specifically, for each image i found in a frame, we report the image type features $types_i$ as the pseudo probabilities provided by the softmax layer of the classifier. This is reasonable, since some images are composites of the different types. It should be noted that the results do not report the number of images seen per class, because a consequence of our frame-wise extraction is that the same image gets extracted multiple times. Instead we analyse the image in every frame again and report the distribution of the image types as a percentage. The idea is to weight the content according to the duration the images have been seen by the learner. The feature vector v_p representing the image types seen per session s is defined as follows:

$$v_s = (\sum_{l=0}^{|L|}\sum_{n=0}^{N_l} p(Info.-Vis.), \ldots, \sum_{l=0}^{|L|}\sum_{n=0}^{N_l} p(Techn.Draw.)) \tag{4.14}$$

Feature vector for the six image types seen per participant. $p(< class >)$ is the pseudo-probability given by the softmax layer. N_l is the number of images detected in frame l.

Textual Information

We introduce 110 features extracted from textual information, taking into consideration the document complexity, the HTML structure and linguistic characteristics. A detailed list of all features can be found in Appendix A.

Document Complexity Features. Based on the assumption that document complexity is correlated with the user's knowledge state on a topic, we have extracted several document complexity-related features. Motivated by previous work [63] and our investigation on the data, we extracted the number of words (c_word), length of words (c_char), and length of sentences (c_sentence) as features. Related work [101] suggests that the syntactic structure of a document, which can be represented by the ratio of the number of nouns, verbs, adjectives, or *other words* (i.e., words that are not verb, noun or adjective) to the total words (c_{noun, verb, adj, oth}), is likely to imply the intention and complexity of its content.

There are several widely used metrics for assessing the readability or complexity of a textual document, which have been studied to be correlated with user's knowledge level [115]. We use Gunning Fog Grade[6] (c_gi), SMOG [174] (c_smog) and Flesch-Kincaid Grade [139] (c_fk) as features. Furthermore, the AoA dictionary proposed by Kuperman et al. [144] contains a listing of more than 30,000 English words along with the age at which native speakers typically learn the term , we compute the age-of-acquisition across all words on Web pages (*c_aoa*), which provides another indicator of document complexity.

HTML Structural Features. A possible explanation of the finding that there is a negative association between the number of hyperlinks embedded in a Web page and the user's KG [59] is that people may not focus on the content in the presence of too many embedded links. Hence we extract the feature h_link by quantifying the number of outbound links (i.e., the <*a*> elements in our case). Furthermore, we extracted more features that might indicate the readability of a webpage based on HTML tags, namely, the average length of each paragraph (h_p), the <*ul*> elements embedded (h_oth_ul), and the number of scripts (h_script).

Linguistic Features. Related work [115] suggests that the amount of words on Web pages that are correlated with different psychological processes and basic sentiment can influence a learner's cognitive state. The writing style could also affect the readability of a learning resource and the engagement of readers. Motivated by the above observations, we use the 2015 Linguistic Inquiry and Word Count (LIWC) dictionaries[7] to compute linguistic features that reflect the psychological processes, sentiment and the writing style of a Web page content. The features of this category are denoted with the prefix l_ in the remainder of this work.

[6] http://gunning-fog-index.com/
[7] http://liwc.wpengine.com/compare-dictionaries/

4.2.6 Experiments and Results

In this section, we report and discuss analysis results for the features described in Section 4.2.5 on our experimental dataset, which contains 113 search sessions. Users have issued 11.1 queries, and browsed 25.4 webpages on average in each session. On average there was a significant increase in learners' knowledge (KG = 2.15 ± 1.84 for full score of 10) after the learning phase. The effect size for KG was large according to Cohen's d = 1.29. Average pre-knowledge score was 5.22 ± 1.76 and post-knowledge was 7.37 ± 1.6.

Textual Information

We computed the Pearson correlation coefficient (denoted as R) between all the features introduced in Sections 4.2.5 and the three knowledge indicators (pre-KS, post-KS, and KG) respectively. We extract resource-content based features and conduct analysis using data collected from these 113 sessions. The high diversity of sessions led to a relatively high *p-value* of the correlation scores. Due to space limitations, out of the 110 features, we only report features showing meaningful correlations in Table 4.18, that is, the features having correlation > 0.1 or < -0.1 with $p < 0.05$ for at least one of the three knowledge indicators. We observe in Table 4.18 that, in total, 9 features fulfill the aforementioned conditions. Among the nine features, seven are linguistic features extracted based on LIWC dictionaries[8] [212] (notations with prefix $l\_$). The linguistic feature *number of body words* (l_{body}) is moderately correlated ($R > 0.2$, $p < 0.05$) with KG. A potential reason could be that l_{body} is representative of the domain knowledge corresponding to the learning task in our study.

HTML based features number of <object> elements (h_{obj}) and number of <img> elements (h_{img}) are weakly correlated with KG ($R > 0.16$, $p < 0.1$). The <object> element is used to embed external resources in a webpage, which is often used for embedding video objects in practice. Although we find positive linear relationship with number of objects and users' KG, we can not draw any conclusion without knowing its relation to the users' prior knowledge state. The impact of video watching behavior on KG is discussed further in the next section. The h_{img} features analyzed in this section provide a meaningful signal for understanding the relationship between images in a webpage and users' knowledge changes. It is worth noting that it is different from the image features analyzed in Section 4.2.6, where the features are calculated based on screen recordings and focused on the visual content that has been seen by the users and includes screenshots of videos. We did not find any document complexity based feature that is linearly correlated with KG in our dataset with $p < 0.1$.

[8] https://liwc.wpengine.com/

	pre-KS	p-val	post-KS	p-val	KG	p-val
l_body	-0.125	0.189	0.098	0.300	**0.205**	**0.030**
l_certain	0.023	0.809	**0.244**	**0.009**	**0.190**	**0.044**
h_img	-0.067	0.481	0.145	0.126	**0.190**	**0.044**
h_obj	0.009	0.927	**0.198**	**0.036**	0.163	0.084
l_informal	0.103	0.275	**0.261**	**0.005**	0.128	0.176
l_netspeak	0.088	0.353	**0.242**	**0.010**	0.126	0.184
l_feel	0.148	0.117	**0.188**	**0.046**	0.022	0.819
l_prep	**0.197**	**0.037**	**0.203**	**0.031**	-0.011	0.906
l_work	-0.053	0.574	**-0.257**	**0.006**	-0.172	0.068

TABLE 4.18: Results of the resource content features correlation. Findings with $|R| > 0.1$ and $p < 0.05$ are highlighted.

	pre-KS	p-val	post-KS	p-val	KG	p-val
$\overline{imgsize}$	-0.174	0.065 (1)	0.006	0.953	0.171	0.070 (2)
watch time in s	-0.177	0.061 (1)	-0.023	0.806	0.149	0.115
Background	0.117	0.217	0.025	0.792	-0.090	0.344
Images/Frames	0.142	0.134	0.041	0.668	-0.100	0.292
Text	-0.164	0.083 (3)	-0.029	0.760	0.131	0.166
Content list	0.072	0.450	0.003	0.973	-0.066	0.489
Heading	-0.060	0.531	-0.114	0.230	-0.042	0.658
Menu bar	0.039	0.682	-0.103	0.277	-0.127	0.180

TABLE 4.19: Results of the document layout analysis. Findings with $|R| > 0.1$ and $p < 0.1$ are underlined. Labels (1) to (3) correspond to the referenced findings in the text.

Document Layout Analysis

Table 4.19 shows how the results of the document layout analysis correlate with the learning performance of the participants. We discuss all findings within the 90% confidence interval, since they might give hints for future research. The findings are numbered from (1) to (6) and can be found in the Tables 4.19 and 4.20.

First, the pre-knowledge state shows a negative correlation with the $\overline{imgsize}$ seen by the user as well as the video watch time in seconds. As mentioned before, these two features describe similar characteristics of a document and correlate positively ($corr_{pearson} = 0.56$). Thus, our first **finding (1)** suggests, that people who knew more about the lightning topic from the beginning searched less for audiovisual content. They presumably search directly for details in more conventional websites to close the gaps of their knowledge instead of watching a video that explains the whole procedure. **Finding (2)** regards that KG has a positive correlation with $\overline{imgsize}$. It can be explained by the fact that prior knowledge and KG are strongly negatively correlated (corr=-0.61, p<0.001). In other words, high KG usually correlates with a low pre-knowledge. In this sense, finding (2) is consistent with finding (1) and suggests that users with a low pre-knowledge consumed more audiovisual material and learned something. To understand **finding (3)**, the negative correlation of text shown on the website with higher pre-knowledge state, we have to take

	pre-KS	p-val	post-KS	p-val	KG	p-val
Information Visualisation	0.011	0.911	0.079	0.408	0.058	0.541
Infographic	0.020	0.831	0.120	0.207	0.085	0.372
Indoor	0.107	0.261	0.002	0.987	-0.100	0.290
Map	0.114	0.231	0.182	0.053 (4)	0.050	0.598
Outdoor	**0.202**	**0.032** (5)	**0.199**	**0.035** (6)	-0.020	0.835
Technical Drawing	0.099	0.295	0.095	0.314	-0.012	0.900

TABLE 4.20: Results of the image type correlation analysis. Findings with $|R| > 0.1$ and $p < 0.1$ are underlined and $p < 0.05$ marked as bold. Labels (4) to (6) correspond to the references in the text.

into account another weak correlation (Images/Frames). Our results suggest that websites visited by participants with *higher pre-knowledge* have the following characteristics: they were comprised of more images (85% confidence interval) (as for low knowledge users) , but following finding (1), these images were smaller than the average. Thus, users with high pre-knowledge viewed less videos, but web pages with some image content. Now, finding (3) indicates that from these websites, people preferred those that had a lower share of textual information and thus, more visual information. Or, in other words, larger images that were directly embedded into the text rather than simple, optional thumbnails.

Image Type Classification

The second set of results, as shown in Table 4.20, presents the results of our content analysis. We discuss the correlations within the 95% confidence interval. The **finding (4)** indicates that *(weather-)maps* and post-knowledge are positively correlated. Content-wise, especially weather maps may help understand the conditions necessary to form thunderclouds (areas of high and low pressure) and lead to lightning. **Findings (5) and (6)** are related to the correlation between *Outdoor Images* and pre- as well as post-knowledge state, that is the more knowledgeable the participants were the more outdoor images they saw. It has to be further investigated whether these were outdoor images that also had a explaining function beyond simple decorations.

The results presented in the previous two subsections give some hints how the different strategies of learners with different knowledge states can be related to search session behavior.

Classification-Based Feature Analysis

Since one of the most important applications for our work is to allow search engines to predict users' knowledge states (and subsequently gain) depending on a set of given learning resources automatically, we follow the classification approach for KG prediction to further investigate the relation between user KG and the features extracted from multimedia resources. We aim for a fair comparison with the state of the art in users' KG

prediction in Web search. Thus, we follow the same experimental setup as used by Yu, Gadiraju, Holtz, Rokicki, Kemkes, and Dietze [293], in particular for the assignment of labels, the applied classifier, and its parameter tuning, unless other settings are denoted.

Ground Truth. We group search sessions in their experimental dataset into three classes based on the *Standard Deviation Classification* approach. We used statistically defined intervals ($X < -0.5\,SD + \overline{X} = low; -0.5\,SD + \overline{X} < X < 0.5\,SD + \overline{X} = moderate; 0.5\,SD + \overline{X} < X = high$) for the classification of the sessions with low, moderate, or high KG. By following the same approach, we label the 113 sessions in our dataset which we can extract both category of features based on the amount of KG exhibited in the session and result in 44 low, 42 moderate and 27 high KG sessions.

Classifier. Random forest has shown to be the most effective classifier for KG prediction by Yu et al. [293] and it assesses feature importance. Hence, in our study, we adopt a random forest classifier, and tune the hyperparameters ($max\_depth, max\_features, n\_estimators$) for accuracy using grid search with the scoring metric *accuracy* and 10-fold cross-validation. For our experiments, we used the *scikit-learn* library for Python [9].

Metrics. After tuning the hyper-parameters of each classifier, we run 10 repetitions of 10-folds cross-validation and evaluate the classification results of each classifier according to the following metrics:

- **Accuracy (*Accu*) across all classes:** percentage of search sessions that were classified with the correct class label.
- **Precision (*P*), Recall (*R*), F1 (*F*1) score of class *i*:** the standard precision, recall and F1 score on the prediction result of each class *i*.
- **Macro average of precision (*P*), recall (*R*), and F1 (*F*1):** the average of the corresponding score across three classes.

Classification Results

The average performance of 10 repetitions of the tuned random forest classifier is shown in Table 4.21. Compared to the classifiers reported in related work [293], we have achieved comparable performance with less training data (113 sessions versus 468 sessions) and more unbalanced classes. We also present the performance of the random forest classifier using behavior features (*BE*, i.e., the approach used in [293]) on our ground truth dataset in the last row of Table 6 for reference. When comparing between model performances on our dataset (Table 4.21), the feature combination visual information (VI) & textual information (TI) outperforms using behavior features at 85% confidence level in terms of accuracy. However, it is worth noting that, we did not manage to extract all the behavior features introduced in the related work, in particular, the features relevant to the clicking on SERPs, as the necessary data is not available in our dataset. Since we focus on understanding the influence of textual and multimedia resource content on users' KG during search, the

[9] https://scikit-learn.org/

classification model and the analysis of user behavior features are out of the scope of this work. We list the results of the classifier trained on user behavior features as evidence that our classification has reached satisfying performance and can provide evidence for the following analysis.

The classifier using features from both categories (VI&TI) has achieved the best performance with respect to overall accuracy. This indicates that by analyzing the textual content and multimedia features of user viewed Web resource, we could collect evidence for predicting users' KG. Compared between different classes, when using features from both categories, the classifier performs better on low and moderate KG classes, a potential reason is that the high KG class has the lease amount of training data in our ground truth dataset.

	Low			Moderate			High			Macro Avg			All
Feature	P	R	F1	P	R	F1	P	R	F1	P	R	F1	Accu
VI&TI	**41.5**	**52.0**	**46.1**	**39.1**	**40.0**	**39.5**	28.4	14.8	19.1	36.4	35.6	34.9	**38.7**
TI	39.9	**52.0**	45.0	36.6	33.8	35.0	28.9	17.4	21.5	35.1	34.4	33.8	37.0
VI	38.0	38.0	37.9	38.0	38.1	38.0	30.8	**31.1**	**30.8**	35.6	35.7	35.6	36.4
BE [293]	39.7	47.0	43.0	37.4	39.5	38.4	**34.9**	21.1	26.0	**37.3**	**35.9**	**35.8**	38.1

TABLE 4.21: Result of KG classification showing our results regarding multimedia (VI), textual (TI) features, and their combination and comparing them with the state-of-the-art based on BE.

Feature Importance

To analyze the usefulness of individual features, we make use of the *Mean Decrease in Impurity (MDI)* metric computed based on the Random Forest model. MDI is defined as the total decrease in node impurity (weighted by the probability of reaching that node) averaged over all trees of the ensemble [30]. We list and discuss the 20 features (Table 4.22) with the highest and lowest MDI values.

We observe that 6 out of 10 features with highest importance are textual features, which is intuitive as 1) there are more textual content features (110) than multimedia features (13), and 2) with recent advances of natural language processing techniques, we were able to design more sophisticated textual features such as the complexity of language and emotions behind words, while it is still challenging to analyze the semantics behind multimedia data. Nevertheless, results indicate that the 13 multimedia features have shown promising importance for the classification, with *Heading*, $\overline{imgsize}$, *Menu Bar*, *Infographic*, *Technical Drawing* and *Outdoor* rank at 4, 5, 8, 9, 13, 15, respectively, among the 123 features in total. None of the multimedia features falls into the 10 least important features according to MDI. Among the six textual features with highest importance, five are linguistic based features extracted based LIWC text analysis, and the rest one is document complexity feature computed based on the SMOG Readability Formula.

	Highest		Lowest	
Rank	**feature**	**MDI**	**feature**	**MDI**
1	l_home	0.039	l_affect	0.004
2	l_relig	0.030	l_Tone	0.004
3	l_certain	0.018	l_power	0.004
4	Heading	0.018	l_AllPunc	0.003
5	*imgsize*	0.016	h_vid	0.003
6	c_smog	0.015	l_filler	0.003
7	l_focuspresent	0.015	l_sad	0.003
8	Menubar	0.015	h_aud	0.003
9	Infographic	0.014	l_Authentic	0.002
10	l_netspeak	0.014	h_obj	0.001

TABLE 4.22: Features having highest and lowest feature importance according to MDI values.

4.2.7 Summary

In this section, we have investigated whether features describing multimedia resource content can help predict users' KG in a SAL task. Our results are based on a large lab study with N=113 participants, where we recorded the individuals' behavior and the accessed Web resources. We extracted the textual and multimedia features to classify the KG of the participants. Finally, we provided a comprehensive analysis of feature importance. It was shown that the combination of our feature categories can serve for KG prediction based on viewed resource content, which potentially can help improve a learning-oriented search result ranking (if content features are used accordingly). Although the classification accuracy is on a moderate level in terms of recall and precision, they suggest that KG is predictable. Particularly image and video features improved the classification notably when used jointly with text-based features.

4.3 Summary

Research Question 2

To what extent can we extract textual, multimedia, and cross-modal features and utilize them for knowledge gain prediction?

In this chapter we investigated how different uni- and cross-modal modalities influence the success of knowledge gain prediction. The results showed that it is challenging to determine a *best* type of feature for this task in general. As discussed in Section 1.3.2, results in this research area depend on several circumstances, such as the learner's cognitive abilities, their pre-knowledge state, the topic, and the presentation itself. Results in Section 4.1.6 indicated a slight advantage for text-based features, generated from semantic sentence embeddings. Conversely, Section 4.2.6 showed that a combination of multimedia and resource features leads to the best prediction accuracy.

This inconclusiveness suggests that we have not yet reached a level of automatic understanding that allows us to reliably forecast the knowledge gain for educational resources of the variety we have analyzed in Chapter 3 and 4.

◇ ◇ ◇

To better understand how information, in general, is perceived, Chapter 5 investigates semantic image-text relations that describe the process of meaning-making between content and learner.

5 Semantic Image-Text Relations

Chapters 3 and 4 suggest that we have not yet reached a sufficient understanding of the influence of multimodal information on the learning outcome. In this chapter, we investigate multimodal meaning-making from a more general perspective. For this, we explore possible combinations of visual and textual information backed up by research from communication science with the goal of establishing a neural network-based classification model, that is applicable to arbitrary content in the image-text domain. This goal is summarized in the third and final research question:

Research Question 3

Based on insights from linguistics and visual communications, how can we derive computational models that describe the relationship between image and text?

Sections 5.1–5.3.8 outline how we approach this research question, and Section 5.4 examines the applicability of the proposed system to other cross-modal domains.

5.1 Motivation

In our digitized world, we face multimodal information on a daily basis in various situations: consumption of news, entertainment, everyday learning or learning in formal education, social media, advertisements, etc. Different modalities help to convey information in an optimal manner that is facilitating effective and efficient communication. For instance, imagine to describe the exact shape of a leaf in textual form or, on the contrary, a specific date such as a birthday in solely visual form [103, 104]. Neither of them is possible in a straightforward and comprehensible way, and in general, it is not possible to translate every kind of information from one modality to another. Although a quote says that "a picture is worth a thousand words", it is normally very difficult or even impossible to denote these words. Thus, to appropriately make use of a single modality or two modalities is a key element for effective and efficient communication.

In a similar context, bridging the *semantic gap* has been identified as one of the key challenges in image retrieval (and multimedia) research, defined as "the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation" by Smeulders et al. [243]. One challenge at this point in time was that information extraction from images was limited to low-level features. As a consequence, most multimedia and computer vision approaches

FIGURE 5.1: An example of a complex message portrayed by an image-text pair elucidating the gap between the textual information and the image content. (CC0 by https://pixabay.com/service/license/)

aim to solve the (perceptual) problem of object and scene recognition, considering visual concepts as semantic, high-level features. In fact, impressive progress has been reported for tasks such as object and visual concept recognition [99, 142], or image captioning [135, 7] in recent years. However, these approaches mostly address *only one possible interpretation* of visual content focusing on, for instance, objects and persons, but lack capabilities of human scene interpretation *going beyond the visible scene content*, i.e., interpreting symbols, gestures, and other contextual information. The complexity increases when we consider *multimodal information or cross-modal references* instead of *solely visual information*. The semantic gap is often caused (or enlarged) by a *modality gap*, since there is no direct translation between different modalities in general, as outlined above. In this chapter, we focus on the interplay of visual and textual information. An example is depicted in Fig. 5.1, which illustrates the interplay of interdependent textual and visual information. Today's state-of-the-art approaches normally do not contribute to answer intricate questions like "How much context or meaning is shared between text and image independent of the amount of shared concepts?" or "Does the type of information (or *image-text class*) match the current user query or retrieval scenario?". A deeper understanding of the multimodal interplay of image and text and the resulting message is necessary to answer such questions. A challenge is that textual and related visual information are often not directly aligned. Moreover, their interplay is typically complex, and there is a large number of roles image and text can take on. In communication sciences and linguistics, this fact is often denoted as the "visual/verbal divide", which, for example, is well observable in comics or audiovisual data and examined in detail by Bateman [16].

Recently, this research topic has gained some attention from some computer science researchers, who, either intentionally or unintentionally, assimilated ideas from communication sciences. Zhang et al. [298] investigate image-text relations in advertisements and distinguish between equivalent and non-equivalent parallel information transfer. They propose a method that automatically detects if the ad's slogan and pictorial component convey the same message independently or if there is a bigger, mutual message. While

this distinction is useful, it has been actually proposed before but was termed differently (e.g., *additive* and *parallel* [140], *independent* and *complementary* [167], and in a more general manner in own previous work [103, 104]). Kruk et al. [143] tailor Marsh and White's taxonomy [166] to measure the author's intent of Instagram posts and two kinds of image-text relations, namely the *contextual relation* between the literal meanings of the image and caption, and the *semiotic relationship* between the meanings of the image and caption. To address Instagram posts, they suggest additions to established definitions, thus making their system less able to generalize to other domains. Henning and Ewerth [103, 104] presented a more general approach by introducing two metrics to describe image-text relations: *cross-modal mutual information (CMI)* and *semantic correlation (SC)*. The metrics are based on the assumptions that visual and textual information can relate to each other a) based on their depicted or mentioned content or b) based on their semantic context.

In this chapter, we follow this paradigm and present the following contributions: After presenting related work in the fields of Multimedia Retrieval and computable image-text relations (Section 5.2), we first extend this set of two metrics by introducing a third metric called "Status" based on insights from linguistics and communication sciences in Section 5.3.1. Second, in Section 5.3.3, we show how this set of metrics can be used to derive a set of eight semantic image-text classes, which are also coherent with studies and taxonomies from linguistics and communication science. Third, we demonstrate how to automatically gather samples from various Web resources in order to create a large (training) dataset in Section 5.3.5, which we make publicly available. Finally, we present two baselines in the form of deep learning systems to predict either the three metrics or the eight image-text classes directly in Section 5.3.6. We evaluate them in Section 5.3.7. Lastly, in Section 5.4, we investigate how well these models are able to generalize the learned concepts by applying them to unseen content.

5.2 Related Work

Multimedia Retrieval

Numerous publications in recent years deal with multimodal information in retrieval tasks. The general problem of reducing or bridging the semantic gap [243] between images and text is the main issue in cross-media retrieval [217, 12, 185, 184, 289]. Fan et al. [73] tackle this problem by modeling humans' visual senses with a multi-sensory fusion network. They handle the *cognitive and semantic gap* by improving the comparability of heterogeneous media features and obtain good results for image-to-text and text-to-image retrieval. Liang et al. [150] propose a self-paced cross-modal subspace matching method by constructing a multimodal graph that preserves both the intra-modality and inter-modality similarity. Another application is targeted by Mazloom et al. [173], who extract a set of engagement parameters to predict the popularity of social media posts. While the confidence in predicting basic emotions like happiness or sadness can be improved by

multimodal features [288], even more, complex semantic concepts like sarcasm [229] or metaphors [241] can be predicted. This is enabled by evaluating the textual cues in the context of the image, providing a new level of semantic richness. The attention-based text embeddings introduced by Bahdanau et al. [11] analyze textual information under the consideration of previously generated image embeddings and improve tasks like document classification [291] and image caption generation [126, 7, 147].

A prerequisite to using heterogeneous modalities is the encoding in a joint feature space, which depends on the type of modality to encode, the number of training samples available, the type of classification to perform and the desired interpretability of the models [14]. One type of algorithms utilizes *Multiple Kernel Learning* [36, 81]. Application areas are multimodal affect recognition [215, 124], event detection [292], and Alzheimer's disease classification [153]. Deep neural networks can also be utilized to model multimodal embeddings. For instance, these systems can be used for the generation of image captions [135]; Ramanishka et al. [218] exploit audiovisual data and metadata, i.e., a video's domain, to generate coherent video descriptions "in the wild", using a Convolutional Neural Networks (CNN) [99]) to encode visual data. Alternative network architectures are GoogleNet [253] or DenseNet [118].

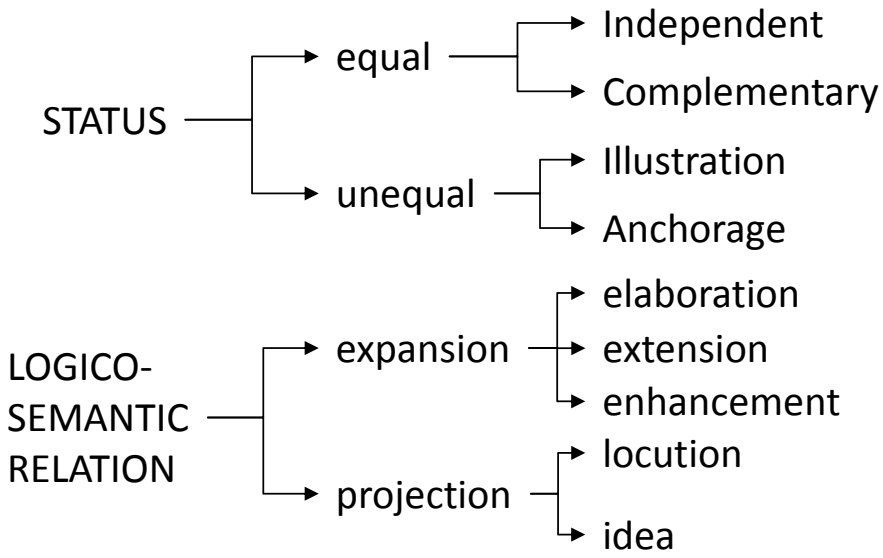


FIGURE 5.2: Part of Martinec and Salway's taxonomy [167] that distinguishes image-text relation based on status (simplified).

Image-Text Relations

The interpretation of multimodal information and the "visual/verbal divide" have been investigated in the field of visual communication and applied linguistics for many years [16]. One research direction in recent decades has dealt with assigning image-text pairs to distinct image-text classes. In a pioneering work, Barthes [15] discusses the respective

roles and functions of text and images. He proposes a first taxonomy, which introduces different types of (hierarchical) status relations between the modalities. If status is unequal, the classes *Illustration* and *Anchorage* are distinguished, otherwise their relation is denoted as *Relay*.

Martinec and Salway [167] extend Barthes' taxonomy and further divide the image-text pairs of *equal* rank into a *Complementary* and *Independent* class, indicating that the information content is either intertwined or equivalent in both modalities. They combine it with Halliday's [94] logicosemantics relations, originally developed to distinguish text clauses. Martinec and Salway revised these grammatical categories to capture the specific logical relationships between text and image regardless of their *status*. McCloud [176] focuses on comic books, whose characteristic is that image and text typically do not share information by means of depicted or mentioned concepts, albeit they have a strong semantic connection. McCloud denotes this category as *Interdependent* and argues that "pictures and words go hand in hand to convey an idea that neither could convey alone". Other authors mention the case of negative correlations between the mentioned or visually depicted concepts (for instance, Nöth [192] or van Leeuwen [268]), denoting them *Contradiction* or *Contrast*, respectively. Van Leeuwen also states that they can be used intentionally, e.g., in magazine advertisements by choosing opposite colors or other formal features to draw attention to certain objects.

Computable Image-Text Relations

Henning and Ewerth [103, 104] propose two metrics to characterize image-text relations in a general manner: *cross-modal mutual information* and *semantic correlation*. They suggest an autoencoder with multimodal embeddings to learn these relations while minimizing the need for annotated training data. Zhang et al. [298] investigate image-text relations in advertisements and distinguish, for instance, between equivalent parallel and non-equivalent parallel information transfer. However, they disregard previous work, e.g., in the field of communication science, and instead of utilizing previous definitions, define their own set of relations. Kruk et al. [143] utilize Marsh and White's taxonomy [166] to model the author's intent of Instagram posts. Two kinds of image-text relations are suggested: the *contextual relation* between the literal meanings of the image and caption and the *semiotic relationship* between the image and the caption.

More recently, Vempala and Preotiuc-Pietro [271] collected a dataset of tweets and labeled them according to two rules: 1) does the image add information to the semantics of the tweet or not, and 2) whether the literal information given in the text can be found in the image or not. While this work helps describe the nature of tweet-based information, it is not based on previous research and is also not comprehensive to all possible ways image and text can interact. Next, Sharma et al. [238] extend the simple relations present in typical image captioning datasets (e.g., MS COCO [151]) by, after applying an intricate filtering mechanism, automatically pairing each image with their alt-texts from the Web.

Class	Uncorrelated CMI=0 SC=0 STAT=0	Interdependent CMI=0 SC=1 STAT=0	Complementary CMI=1 SC=1 STAT=0	Illustration CMI=1 SC=1 STAT=T	Anchorage CMI=1 SC=1 STAT=I	Contrasting CMI=1 SC=-1 STAT=0	Bad Illustration CMI=1 SC=-1 STAT=T	Bad Anchorage CMI=1 SC=-1 STAT=I
What is captured?	No shared concepts or semantic background	No shared concepts, but joint message on a higher semantic level	Modalities complement each other	Text is supplemented with exchangable image	Image is supplemented with a caption describing visual concepts	Modalities complement each other, but contain contradicting details	Given visual example is ill composed, unusual or ambiguous	A given caption describes details of displayed information wrong
Possible Usecases	• Filter for retrieval tasks • Adblocker	• AdBlocker • Marketing Retrieval Tasks	• Recommender systems • Cross-modal retrieval • Web search	• Search tasks in eductional settings • Text books	• Search tasks in educational settings, e.g.: definitions or explanations	• Quality check • Filter for retrieval tasks or recommender systems	• Quality check • Filter for retrieval tasks or recommender systems	• Quality check • Filter for retrieval tasks or recommender systems

FIGURE 5.3: Overview of the proposed image-text classes and their potential use cases.

This provides a greater variety in image-text pairs and results, due to a large number of samples, to better results on popular cross-modal retrieval benchmarks. However, specific metrics are again not labeled and, thus, not comprehensively present.

Alikhani et al. [3] take another research direction and describe the joint message of both modalities according to the discourse coherence theory by Hobbs [109] and Phillips [213]. According to [103], these coherences assume a positive semantic correlation and classify the type of coherence in more detail. For example *Subjective* describes that the text gives the author's reaction of evaluation of what can be seen in the image. Alternatively, *Action*, which entails that the text describes an extended process with the image being a snapshot of that process. Yet, this coherence describes only one aspect of how image and text relate to each other and is not holistic. Image-Text pairs that do not fit a category are labeled *Irrelevant*. A more general approach is needed if we want to answer research question 4.

5.3 Characterization and Classification of Semantic Image-Text Relations

5.3.1 Analysis and Discussion of Related Work

The discussion of related work reveals that the complex cross-modal interplay of image and text has not been systematically modeled and investigated yet from a computer science perspective. In this section, we derive a categorization of classes of semantic image-text relations which can be used for multimedia information retrieval and Web search. This categorization is based on previous work in the fields of visual communication (sciences) and information retrieval. However, one drawback of taxonomies in communication sciences is that their level of detail makes it sometimes difficult to assign image-text pairs to a particular class, as criticized by Bateman [16].

First, we evaluate the image-text classes described in communication science literature. As a point of departure, we consider Martinec and Salway's taxonomy (Fig. 5.2), which yields the classes *Illustration*, *Anchorage*, *Complementary*, and *Independent*. We disregard the class *Independent* since it is very uncommon that both modalities describe exactly the same information. Next, we introduce the class *Interdependent* suggested by McCloud [176], which

in contrast to *Complementary* consists of image-text pairs where the intended meaning cannot be gathered from either of them exclusively. While a number of categorizations does not consider negative semantic correlations at all, Nöth [192], van Leeuwen [268], and Henning and Ewerth [103] consider this aspect. We believe that it is important for information retrieval tasks to consider negative correlations as well, for instance, in order to identify less useful multimodal information, contradictions, mistakes, etc. Consequently, we introduce the classes *Contrasting*, *Bad Illustration*, and *Bad Anchorage*, which are the negative counterparts for *Complementary*, *Illustration*, and *Anchorage*. Finally, we consider the case when text and image are *uncorrelated*.

While one objective of our work is to derive meaningful, distinctive, and comprehensible image-text classes, another contribution is their systematic characterization. For this purpose, we leverage the metrics cross-modal mutual information (CMI) and SC [103]. However, these two metrics are not sufficient to model a wide range of image-text classes. It is apparent that the *status* relation, originally introduced by Barthes [15], is adopted by the majority of taxonomies established in the last four decades (e.g., [167, 264]), implying that this relation is essential to describe an image-text pair. It portrays how two modalities can relate to one another in a hierarchical way reflecting their relative importance. Either the text supports the image (*Anchorage*), or the image supports the text (*Illustration*), or both modalities contribute equally to the overall meaning (e.g., *Complementary*. This encourages us to extend the two-dimensional feature space of CMI and SC with the *status* dimension (*status (STAT)*). In the next section, we provide definitions for the three metrics and subsequently infer a categorization of semantic image-text classes from them. Our goal is to reformulate and clarify the interrelations between visual and textual content in order to make them applicable for multimodal indexing and retrieval. An overview of the image-text classes and their mapping to the metrics, as well as possible use cases is given in Figure 5.3.

5.3.2 Deduction of Semantic Image-Text Metrics

Concepts and entities

The following definitions are related to concepts and entities in images and text. Generally, plenty of concepts and entities can be found in images ranging from the main focus of interest (e.g., a person, a certain object, an event, a diagram) to barely visible or background details (e.g., a leaf of grass, a bird in the sky). Normally, the meaning of an image is related to the main objects in the foreground. When assessing relevant information in images, it is reasonable to regard these concepts and entities, which, however, adds a certain level of subjectivity in some cases. But most of the time the important entities can be easily determined.

Cross-modal mutual information (CMI)

Depending on the (fraction of) mutual presence of concepts and entities in both image and text, the cross-modal mutual information ranges from 0 (no overlap of depicted concepts) to 1 (concepts in image and text overlap entirely). It is important to point out that CMI ignores a deeper semantic meaning, in contrast to *semantic correlation*. If, for example, a small man with a blue shirt is shown in the image, while the text talks about a tall man with a red sweater, the CMI would still be positive due to the mutual concept "man". But since the description is confusing and hinders interpretation of the multimodal information, SC of this image-text pair would be negative. Image-text pairs with high CMI can be found in image captioning datasets, for instance. The images and their corresponding captions have a descriptive nature, which is why they have explicit representations in both modalities. In contrast, news articles or advertisements often have a loose connection to their associated images by means of mutual entities or concepts. The range of CMI is $[0, 1]$.

Semantic correlation (SC)

The (intended) meaning of image and text can range from coherent (SC=1), over uncorrelated (SC=0) to contradictory (SC=-1). This refers to concepts, descriptions and interpretation of symbols, metaphors, as well as to their relations to one another. Typically, an interpretation requires contextual information, knowledge, or experience and it cannot be derived exclusively from the entities in the text and the objects depicted in the image. The range of possible values is $[-1, 1]$, where a negative value indicates that the co-occurrence of an image and a text is contradicting and disturbs the comprehension of the multimodal content. This is the case if a text refers to an object in an image and cannot be found there, or has different attributes as described in the text. An observer might notice a contradiction and ask herself "Do image and text belong together at all, or were they placed jointly by mistake?". A positive score on the contrary suggests that both modalities share a semantic context or meaning. The third possible option is that there is no semantic correlation between entities in the image and the text, yielding $SC = 0$.

Status (STAT)

Status describes the hierarchical relation between an image and text with respect to their relative importance. Either the image is "subordinate to the text" ($stat = T$), implying an exchangeable image which plays the minor role in conveying the overall message of the image-text pair, or the text is "subordinate to the image" ($stat = I$), usually characterizing text with additional information (e.g., a caption) for an image that is the center of attention. An *equal status* ($stat = 0$) describes the situation where image and text are equally important to convey the overall message.

Images which are "subordinate to text" (class *Illustration*) 'elucidate' or 'realize' the text. This is the case, if a text describes a general concept and the associated image shows a concrete example of that concept. Examples for the class *Illustration* can be found in textbooks and encyclopedias. On the contrary, in the class *Anchorage* the text is "subordinate to the image". This is the case, if the text answers the question "What can be seen in this image?". It is common that direct references to objects in the image can be found and the readers are informed what they are looking at. This type of image-text pair can be found in newspapers or scientific documents, but also in image captioning data sets. The third possible state of a *status relation* is "equal", which describes an image-text pair where both modalities contribute individually to the conveyed information. Also, either part contains details that the other one does not. According to Barthes [15], this class describes the situation where the information depicted in either modality is part of a more general message and together they elucidate information on a higher level that neither could do alone.

5.3.3 Categorization of Image-Text Classes

In this section, we show how the combination of our three metrics can be naturally mapped to distinctive image-text classes (see also Fig. 5.3). For this purpose, we simplify the data value space for each dimension. The level of semantic correlation can be represented by the interval $[-1, 1]$. Henning and Ewerth [103, 104] distinguish five levels of CMI and SC. We, however, omit these intermediate levels since the general idea of positive, negative, and uncorrelated pairs is sufficient for the task of assigning image-text pairs to distinct classes. Therefore, the possible states of SC are: $sc \in \{-1, 0, 1\}$. For a similar reason, finer levels for CMI are omitted, resulting in two possible states for $cmi \in \{0, 1\}$, which correspond to *no overlap* and *overlap*. Possible states of status are $stat \in \{T, 0, I\}$: *image subordinate to text* ($stat = T$), *equal status* ($stat = 0$), and *text subordinate to image* ($stat = I$).

If approached naively, there are $2 \times 3 \times 3 = 18$ possible combinations of SC, CMI , and STAT . A closer inspection reveals that (only) eight of these classes match with taxonomies in communication science, confirming the coherence of our analysis. The remaining ten classes can be discarded since they cannot occur in practice or do not make sense. The reasoning is given after we have defined the eight classes that form the categorization.

Uncorrelated ($cmi = 0, sc = 0, stat = 0$)
This class contains image-text pairs that do not belong together in an obvious way. They neither share entities and concepts nor there is an interpretation for a semantic correlation (e.g., see Fig. 5.5, left).

Complementary ($cmi = 1, sc = 1, stat = 0$)
The class *Complementary* comprises the classic interplay between visual and textual information, i.e., both modalities share information but also provide information that the other one does not. Neither of them is dependent on the other one and their status is equal. It

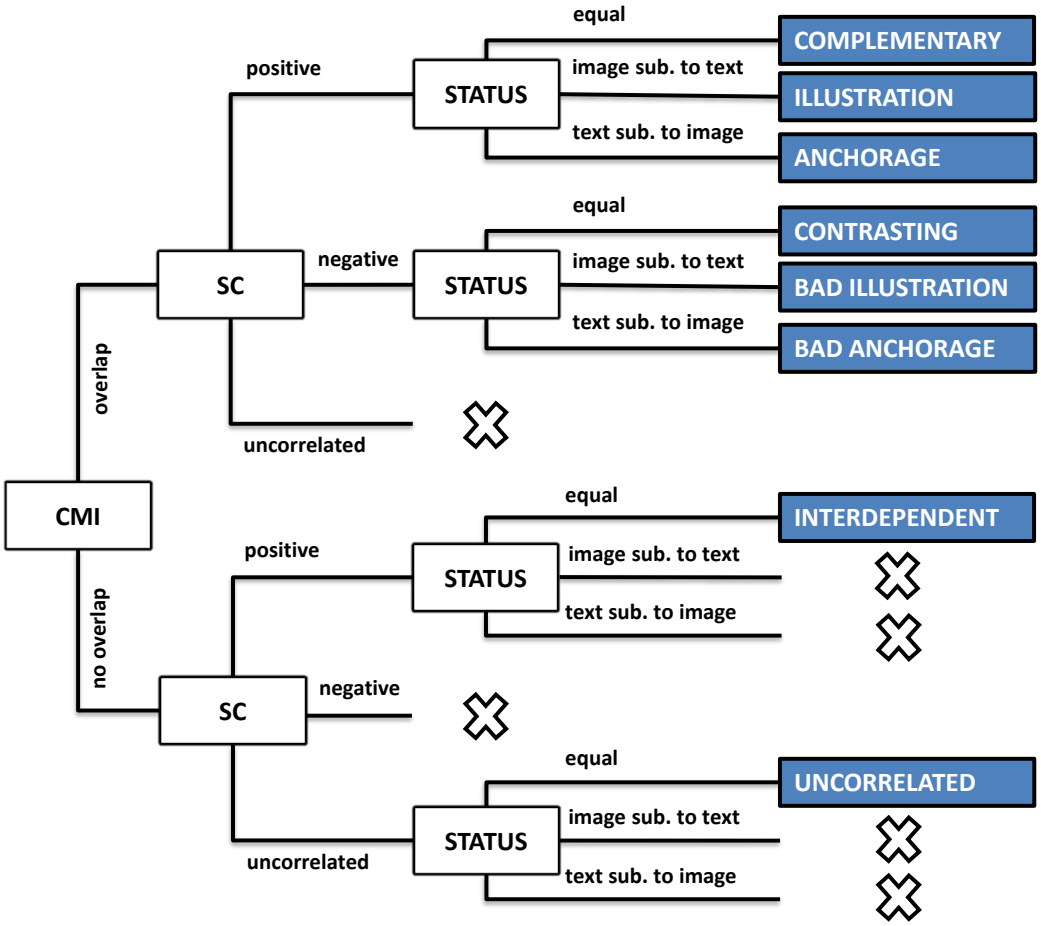


FIGURE 5.4: Our categorization of image-text relations. Discarded subtrees or leaves are marked by an **X** for clarity. Please note that there are no hierarchical relations implied.

is important to note that the amount of information is not necessarily the same in both modalities. The most significant factor is that an observer is still able to understand the key information provided by either of the modalities alone (Fig. 5.5, right). The definitions of the next two classes will clarify that further.

Interdependent ($cmi = 0, sc = 1, stat = 0$)
This class includes image-text pairs that do not share entities or concepts by means of mutual information, but are related by a semantic context. As a result, their combination conveys a new meaning or interpretation which neither of the modalities could have achieved on its own. Such image-text pairs are prevalent in advertisements where companies combine eye-catching images with funny slogans supported by metaphors or puns, without actually naming their product (Fig. 5.5, middle). Another genre that relies heavily on these *interdependent* examples are comics or graphic novels, where speech bubbles and accompanying drawings are used to tell a story. Interdependent information is also prevalent in movies and TV material in the auditory and visual modalities.

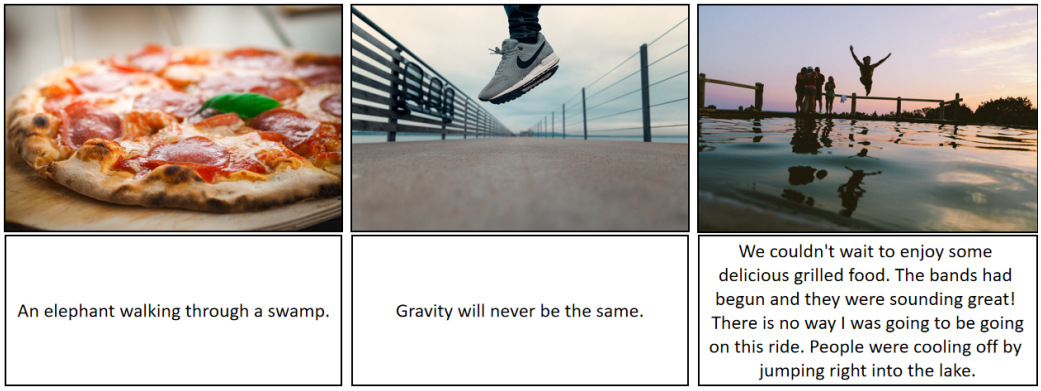


FIGURE 5.5: Examples for the *Uncorrelated* (left), *Interdependent* (middle) and *Complementary* (right) classes. (🄯 by https://pixabay.com/service/license/)

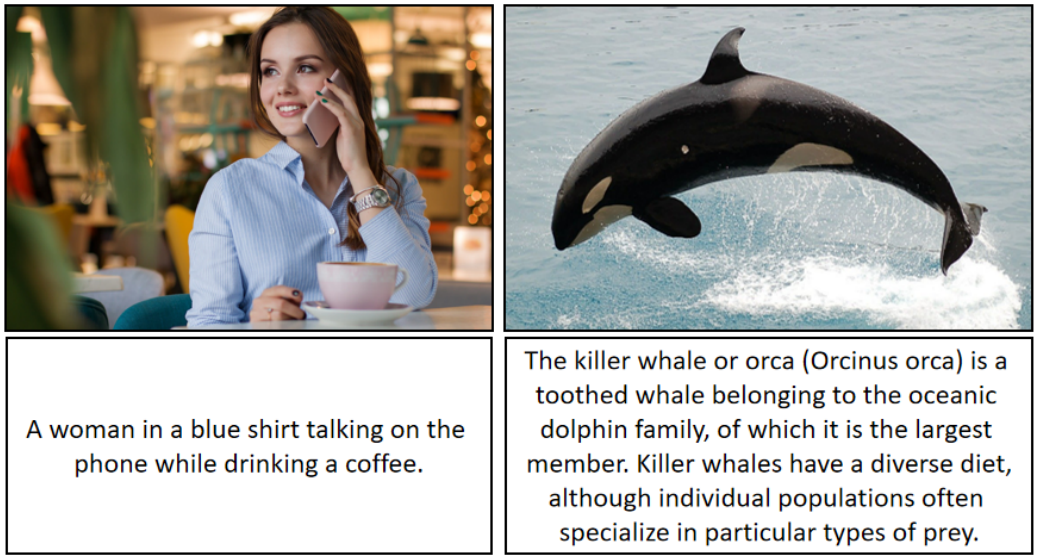


FIGURE 5.6: Examples for the *Anchorage* (left) and *Illustration* (right) classes. (🄯 by https://pixabay.com/service/license/)

Anchorage ($cmi = 1, sc = 1, stat = I$)
On the contrary, the *Anchorage* class is an image description and acts as a supplement for an image. Barthes states that the role of the text in this class is to fix the interpretation of the visual information as intended by the author of the image-text pair [15]. It answers the question "What is it?" in a more or less detailed manner. This is often necessary since the possible meaning or interpretation of an image can noticeably vary and the caption is provided to pinpoint the author's intention. Therefore, an *Anchorage* can be a simple image caption, but also a longer text that elucidates the hidden meaning of a painting. It is similar to *Complementary*, but the main difference is that the text is subordinate to image in *Anchorage* (see Fig. 5.6).

Illustration ($cmi = 1, sc = 1, stat = T$)
The class *Illustration* contains image-text pairs where the visual information is subordinate to the text and has therefore a lower *status*. An instance of this class could be, for example,

a text that describes a general concept and the accompanying image depicts a specific example (Fig. 5.6). A distinctive feature of this class is that the image is replaceable by a very different image without rendering the constellation invalid. If the text is a definition of the term "mammal", it does not matter if the image shows an elephant, a mouse, or a dolphin. Each of these examples would be valid in this scenario. In general, the text is not dependent on the image to provide the intended information.

Contrasting ($cmi = 1, sc = -1, stat = 0$)
Bad Illustration ($cmi = 1, sc = -1, stat = T$)
Bad Anchorage ($cmi = 1, sc = -1, stat = I$)
These three classes are the counterparts to *Complementary*, *Illustration*, and *Anchorage*: they share their primary features, but have a **negative SC** (see Fig. 5.7). In other words, the transfer of knowledge is impaired due to inconsistencies or contradictions when jointly viewing image and text [103]. In contrast to *uncorrelated* image-text pairs, these classes share information and obviously they belong together in a certain way, but particular details or characteristics are contradicting. For instance, a *Bad Illustration* pair could consist of a textual description of a bird, whose most prominent feature is its colorful plumage, but the bird in the image is actually a grey pigeon. This can be confusing and an observer might be unsure if she is looking at the right image. Similarly, contradicting textual counterparts exist for each of these classes. In section 5.3.5, we describe how we generate training samples for these classes.

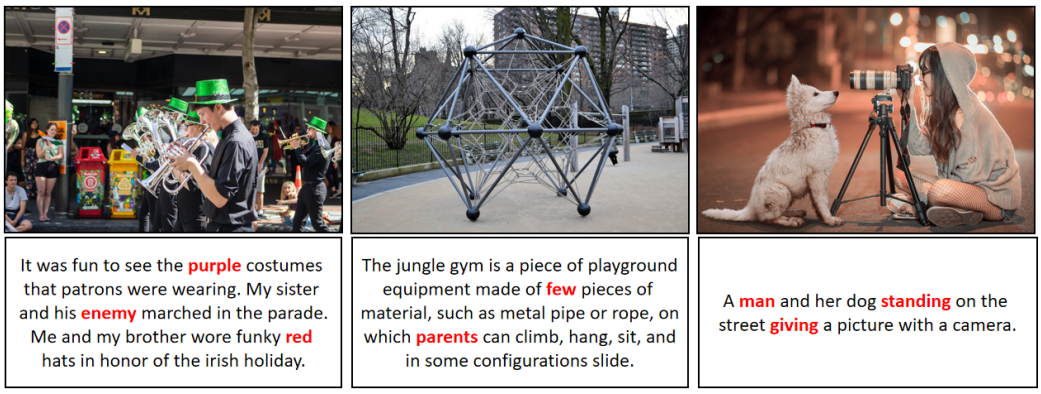


FIGURE 5.7: Examples for the *Contrasting* (left), *Bad Illustration* (middle), and *Bad Anchorage* (right) classes. (CC BY https://pixabay.com/service/license/)

Contradictory Image-Text Relations

The eight classes described above form the categorization as shown in Figure 5.4. The following ten combinations of metrics were discarded, since they do not yield meaningful image-text pairs.

Cases A: $cmi = 0, sc = -1, stat = T, 0, I$

These three classes cannot exist: If the shared information is zero, then there is nothing that can contradict one another. As soon as a textual description relates to a visual concept in the image, there is cross-modal mutual information and $CMI > 0$.

Cases B: $cmi = 0, sc = 0, stat = T, I$
The metric combination $cmi = 0, sc = 0, stat = 0$ describes the class *Uncorrelated* of image-text pairs which are neither in contextual nor visual relation to one another. Since it is not intuitive that a text is subordinate to an uncorrelated image or vice versa, these two classes are discarded.

Cases C: $cmi = 0, sc = 1, stat = T, I$
Image-text pairs in the class *Interdependent* ($cmi = 0, sc = 1, stat = 0$) are characterized by the fact, that even though they do not share any information they still complement each other by conveying additional or new meaning. Due to the nature of this class a subordination of one modality to the other one is not plausible: Neither of the conditions for the states *image subordinate to text* and *text subordinate to image* is fulfilled due to lack of shared concepts and entities. Therefore, these two classes are discarded.

Cases D: $cmi = 1, sc = 0, stat = T, 0, I$
As soon as there is an overlap of essential depicted concepts there has to be a minimum of semantic overlap. We consider entities as essential, if they contribute to the overall information or meaning of the image-text pair. This excludes trivial background information such as the type of hat a person wears in an audience behind a politician giving a speech. The semantic correlation can be minor, but it would still correspond to $SC = 1$ according to the definition above. Therefore, the combination $cmi = 1, sc = 0$ and the involved possible combinations of *STAT* are discarded.

5.3.4 Automatic Prediction of Semantic Image-Text Classes

In this section, we present our approach to automatically predict the introduced image-text metrics and classes. We propose a deep learning architecture that realizes a multimodal embedding for textual and pictorial data. Deep neural networks achieve better results, when they are trained with a large amount of data. However, for the addressed task no such dataset exists. Crowdsourcing is an alternative to avoid the time-consuming task of manually annotating training data on our own, but requires significant efforts to maintain the quality of annotations obtained in this way. Therefore, we follow two strategies to create a sufficiently large training set. First, we automatically collect image-text pairs from different open access Web sources. Second, we suggest a method for training data augmentation (Section 5.3.5) that allows us to also generate samples for the image-text classes that rarely occur on the Web, for instance, *Bad Illustration*. We suggest two classifiers, a **"classic"** approach, which simply outputs the most likely image-text class, as well as a cascaded approach based on classifiers for the three metrics. The motivation for the latter

is to divide the problem into three easier classification tasks. Their subsequent **"cascaded"** execution will still lead us to the desired output of image-text classes according to Fig. 5.4. The deep learning architecture is explained in section 5.3.4.

5.3.5 Training Data Augmentation

The objective is to acquire a large training dataset of high quality image-text pairs with a minimum effort in manual labor. On the one hand, there are classes like *Complementary* or *Anchorage* available from a multitude of sources and can therefore be easily crawled. Other classes like *Uncorrelated* do not naturally occur in the Web, but can be generated with little effort. On the other hand, there are rare classes like *Contrasting* or *Bad Anchorage*. While they do exist and it is desirable to detect these image-text pairs as well (see Fig. 5.3), there is no abundant source of such examples that could be used to train a robust classifier.

Only few datasets are publicly available that contain images and corresponding textual information, which are not simply based on tags and keywords but also use cohesive sentences. Two examples are the image captioning dataset MSCOCO [151] as well as the Visual Storytelling dataset VIST [119]. A large number of examples can be easily taken from these datasets, namely for the classes *Uncorrelated*, *Complementary*, and *Anchorage*. Specifically, the underlying hierarchy of MSCOCO is exploited to ensure that two randomly picked examples are not semantically related to one another, and then join the caption of one sample with the image of the other one to form *Uncorrelated* samples. In this way, we gathered 60 000 ***uncorrelated*** training samples.

The VIST dataset has three types of captions for their five-image-stories. The first one "Desc-in-Isolation" resembles the generic image-caption dataset and can be used to generate examples for the class ***Anchorage***. These short descriptions are similar to MSCOCO captions, but slightly longer, so we decided to use them. Around 62 000 examples have been generated this way. The pairs represent this class well, since they include textual descriptions of the visually depicted concepts without any low-level visual concepts or added interpretations. More examples could have been generated similarly, but we have to restrict the level of class imbalance. The second type of VIST captions "Story-in-Sequence" is used to create ***Complementary*** samples by concatenating the five captions of a story and pairing them randomly with one of the images of the same story. Using this procedure, we generated 33 088 examples.

While there are certainly much more possible constellations of *complementary* content from a variety of sources, the various types of stories of this dataset give a solid basis. The same argumentation holds for the ***Interdependent*** class. Admittedly, we had to manually label a set of about 1 007 entries of Hussain et al.'s Internet Advertisements data set [121] to generate these image-text pairs. While they exhibit the right type of image-text relations, the accompanied slogans (in the image) are not annotated separately and optical character recognition did not achieve high accuracy due to ornate fonts etc. Furthermore, some image-text pairs had to be removed, since some slogans specifically mention the product

Class	Num. of Samples
Uncorrelated	60 000
Interdependent	1 007
Complementary	33 088
Illustration	5 447
Anchorage	62 637
Contrasting	31 368
Bad Illustration	4 099
Bad Anchorage	27 210

TABLE 5.1: Distribution of class labels in the generated dataset.

Class	Num. of Samples
STAT 0	125 463
STAT T	9 546
STAT I	89 847
SC -1	62 677
SC 0	60 000
SC 1	102 179
CMI 0	61 007
CMI 1	163 849

TABLE 5.2: Distribution of metric labels in the generated dataset.

name. This contradicts the condition that there is no overlap between depicted concepts and textual description, i.e., $cmi = 0$.

The ***Illustration*** class is established by combining one random image for each concept of the ImageNet dataset [226] with the summary of the corresponding article of the English Wikipedia, in case it exists. This nicely fits the nature of the class since the Wikipedia summary often provides a definition including a short overview of a concept. An image of the ImageNet class with the same name as the article should be a replaceable example image of that concept.

The three remaining classes ***Contrasting***, ***Bad Illustration*** and ***Bad Anchorage*** occur rarely and are hard to detect automatically. Therefore, it is not possible to automatically crawl a sufficient amount of samples. To circumvent this problem, we suggest to transform the respective positive counterparts by replacing 530 keywords [152] (adjectives, directional words, colors) by antonyms and opposites in the textual description of the positive examples to make them less comprehensible. For instance, "tall man standing in front of a green car" is transformed into a "small woman standing behind a red car". While this does not absolutely break the semantic connection between image and text, it surely describes certain attributes incorrectly which impairs the accurate understanding and subsequently justifies the label of $sc = -1$. This strategy allows us to transform a substantial amount of the "positive" image-text pairs into their negative counterparts. Finally, for all classes we truncated the text if it exceeded 10 sentences. In total the dataset consists of 224 856 image-text pairs. Table 5.1 and 5.2 give an overview about the data distribution, first sorted by class and the second one according to the distribution of the three metrics, which were also used in our experiments.

5.3.6 Design of Multimodal Deep Classifiers

As mentioned above, we introduce two classification approaches: "classic" and "cascade". The advantage of the latter is that it is easier to maintain a better class balance of samples, while it is also the easier classification problem. For instance, example data of the classes *Contrasting*, *Bad Illustration*, and *Bad Anchorage* are used to train the neural network how

negative semantic correlation looks like. This should make the training process more robust against overfitting and underfitting, but naturally also increases the training and evaluation time by a factor of three.

Both methods follow the architecture shown in Figure 5.8, but for "cascade" three networks have to be trained and subsequently applied to predict an image-text class. To encode the input image, the deep residual network "Inception-ResNet-v2" [253] is used, which is pre-trained on the dataset of the ImageNet challenge [226]. To embed this model in our system, we remove all fully-connected layers and extract the feature maps with an embedding size of 2048 from the last convolutional layer.

The text is encoded by a pre-trained model of the word2vec [183] successor fastText [130], which has the remarkable ability to produce semantically rich feature vectors even for unknown words. This is due to its skip-gram technique, which does not observe words as a whole but as n-grams, that is a sum of word parts. For instance, the word *library* would be decomposed into the following tri-grams: <li, lib, ibr, bra, rar, ary, ry>. Thus, it enables the system to recognize a word or derived phrasings despite of typing errors. FastText utilizes an embedding size of 300 for each word and we feed them into a bidirectional GRU inspired by Yang et al. [291], which reads the sentence(s) forwards and backwards before subsequently concatenating the resulting feature vectors. In addition, an attention mechanism is incorporated through another convolutional layer, which reduces the image encoding to 300 dimensions, matching the dimensionality of the word representation set by fastText. In this way it is ensured that the neural network reads the textual information under the consideration of the visual features, which enforces it to interpret the features in unison. The final text embedding has a dimension of 1024. After concatenating image (to get a global feature representation from the image, we apply average pooling to the aforementioned last convolutional layer) and text features, four consecutive fully connected layers (dimensions: 1024, 512, 256, 128) comprise the classification layer. This layer has two outputs for *CMI*, three outputs for *SC* and *STAT*, or eight outputs for the "classic" classifier, respectively. For the actual classification process in the cascade approach, the resulting three models have to be applied sequentially in an arbitrary order. We select the order $CMI \Rightarrow SC \Rightarrow STAT$, the evaluations of the three classifiers yield the final assignment to one of the eight image-text classes (Fig. 5.4).

5.3.7 Experiments and Results

The dataset was split into a training set and a manually verified test set to ensure high quality labels. It initially contained 800 image-text pairs, where for each of the eight classes 100 examples were taken out of the automatically crawled and augmented data. The remaining 239 307 examples were used to train the four different models (three for the "cascade" classifier and one for the "classic" approach) for 100 000 iterations each with the TensorFlow framework. The *Adam optimizer* was used with its standard learning rate and a dropout rate of 0.3 for the image embedding layer and 0.4 for the text embedding

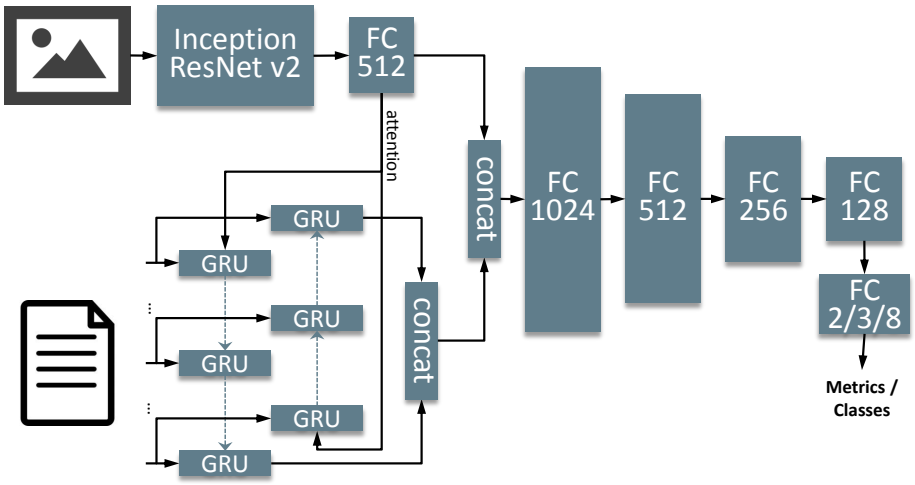


FIGURE 5.8: General structure of the deep learning system with multimodal embedding. The last fully connected layer (FC) has 2, 3, or 8 outputs depending on whether CMI (two levels), SC/STAT (three levels), or all eight image-text classes ("classic" approach) are classified.

layer. Also a softmax cross entropy loss was used and a batch size of 12 on a NVIDIA Titan X. All images were rescaled to a size of 299 × 299 and Szegedy et al.'s [254] image preprocessing techniques were applied. This includes random cropping of the image as well as random brightness, saturation, hue and contrast distortion to avoid overfitting. In addition, we limit the length of the textual information to 50 words per sentence and 30 sentences per image-text pair. All "Inception-ResNet-v2" layers were pre-trained with the ILSVRC (ImageNet Large Scale Visual Recognition Competition) 2010 [226] dataset to reduce the training effort. The training and test datasets are publicly available at https://doi.org/10.25835/0010577.

To assure highly accurate ground-truth data for our test set, we asked three persons of our group (one of them is a co-author) to manually annotate the 800 image-text pairs.

Class	Uncorr.	Interdep.	Compl.	Illustration	Anchorage	Contrasting	Bad Illu.	Bad Anch.
Recall	69.2%	97.6%	83.8%	83.7%	90.3%	89.0%	98.6%	91.9%
Precision	98.7%	96.3%	88.0%	80.7%	87.3%	78.3%	69.0%	87.0%
#Samples	149	100	106	95	95	87	71	95

TABLE 5.3: Comparison of the automatically generated labels with the annotations of the three volunteers (i.e., ground-truth data) and the resulting number of samples per class in the test set.

Each annotator received an instruction document that contained short definitions of the three metrics (Section 5.3.2), the categorization in Fig. 5.4, and one example per image-text class (similar to Figures 5.5-5.7). The inter-coder agreement has been evaluated using Krippendorff's alpha [141] and yielded a value of $\alpha = 0.847$ (across all annotators, samples, and classes). A class label was assigned, if the majority of annotators agreed on it for a sample. Besides the eight image-text classes, the annotators could also mark a sample as *Unsure* which denotes that an assignment was not possible. If *Unsure* was the majority

Class	Undef.	Uncorrelated	Interdep.	Compl.	Illustration	Anchorage	Contrasting	Bad Illust.	Bad Anch.	Sum
Undefined	**0**	0	0	0	0	0	0	0	0	0
Uncorrelated	2	**96**	4	7	21	1	4	13	1	149
Interdependent	3	3	**92**	1	0	1	0	0	0	100
Complementary	1	0	1	**93**	0	2	9	0	0	106
Illustration	1	0	0	0	**82**	0	0	12	0	95
Anchorage	11	4	5	25	1	**41**	2	1	5	95
Contrasting	0	0	0	2	0	0	**85**	0	0	87
Bad Illustration	0	0	0	0	8	0	0	**63**	0	71
Bad Anchorage	9	2	0	4	0	6	33	0	**41**	95
Precision	-	91.4%	90.2%	70.5%	73.2%	80.4%	63.9%	70.8%	87.2%	-
Recall	-	64.4%	92.00%	87.7%	86.3%	43.2%	97.7%	88.7%	43.1%	-

TABLE 5.4: Confusion matrix for the "cascade" classifier on the testset of 798 image-text pairs. The rows depict true positives (bold) and false negatives, i.e., the distribution of the ground-truth samples over the eight classes (+ Undefined). The columns show the true positives (bold) and false positives and thus, the samples that the model actually identified as the respective class.

Class	Undef.	Uncorrelated	Interdep.	Compl.	Illustration	Anchorage	Contrasting	Bad Illust.	Bad Anch.	Sum
Uncorrelated	-	**67**	3	5	23	34	5	11	1	149
Interdependent	-	0	**94**	0	0	5	0	0	1	100
Complementary	-	0	0	**93**	0	4	9	0	0	106
Illustration	-	0	0	0	**84**	0	0	11	0	95
Anchorage	-	2	2	0	2	**83**	0	0	6	95
Contrasting	-	0	0	3	0	0	**84**	0	0	87
Bad Illustration	-	0	0	0	2	0	0	**69**	0	71
Bad Anchorage	-	2	0	0	0	21	1	0	**71**	95
Precision	-	94.4%	94.9%	92.1%	75.7%	56.5%	84.8%	75.8%	89.9%	-
Recall	-	45.0%	94.0%	87.7%	88.4%	87.4%	96.5%	97.2%	74.7%	-

TABLE 5.5: Confusion matrix for the "classic" classifier on the testset of 798 image-text pairs. The rows depict true positives (bold) and false negatives, i.e., the distribution of the ground-truth samples over the eight classes. The columns show the true positives (bold) and false positives and thus, the samples that the model actually identified as the respective class. (Undefined column was added for better comparability with Table 5.4.)

of votes, the sample was not considered for the test set. This only applied for two pairs, which reduced the size of the final test set to 798.

Comparing the human labels with the automatically generated labels allowed us to evaluate the quality of the data acquisition process. Therefore we computed how good the automatic labels matched with the human ground truth labels (Table 5.3). The low recall for the class *Uncorrelated* indicates that there were uncorrelated samples in the other data sources that we exploited. The *Bad Illustration* class has the lowest precision and was mostly confused with *Illustration* and *Uncorrelated*, that is the human annotators considered the automatically "augmented" samples either as still valid or uncorrelated.

-	**CMI 0**	**CMI 1**	**SC 0**	**SC 1**	**SC -1**	**STAT 0**	**STAT T**	**STAT I**
Precision	87.7%	91.4%	81.8%	84.2%	86.6%	82.5%	82.2%	92.8%
Recall	80.3%	94.9%	90.5%	64.4%	88.4%	90.5%	100.0%	54.2%

TABLE 5.6: Performance of the single metric classifiers.

The results for predicting image-text classes using both the "classic" (Table 5.5) and "cascade" approach (Table 5.4) are presented in confusion matrices by means of precision and recall. For a better comparison, Fig. 5.9 shows the individual performance for each image-text class. The overall results for our classifiers in predicting CMI, SC, STAT as

Classifier	CMI	SC	STAT	Cascade	Classic
Ours	90.3%	84.6%	83.8%	**74.3%**	**80.8%**
Henning & Ewerth [103]	68.8%	49.6%	-	-	-

TABLE 5.7: Test set accuracy of the metric-specific classifiers and the two final classifiers after 75 000 iterations.

well as the image-text classes are presented in Table 5.7. The accuracy of the classifiers for CMI , SC and STAT ranges from 83.8% to 90.3%, while the two classification variations for the image-text classes achieved an accuracy of 74.3% (*cascade*) and 80.8% (*classic*). We also compared our method with our previous approach [103, 104] by mapping their intermediate steps for CMI =0, 1, 2 to 0, CMI =3,4 to 1, and SC=$\pm$0.5 to $\pm$1.

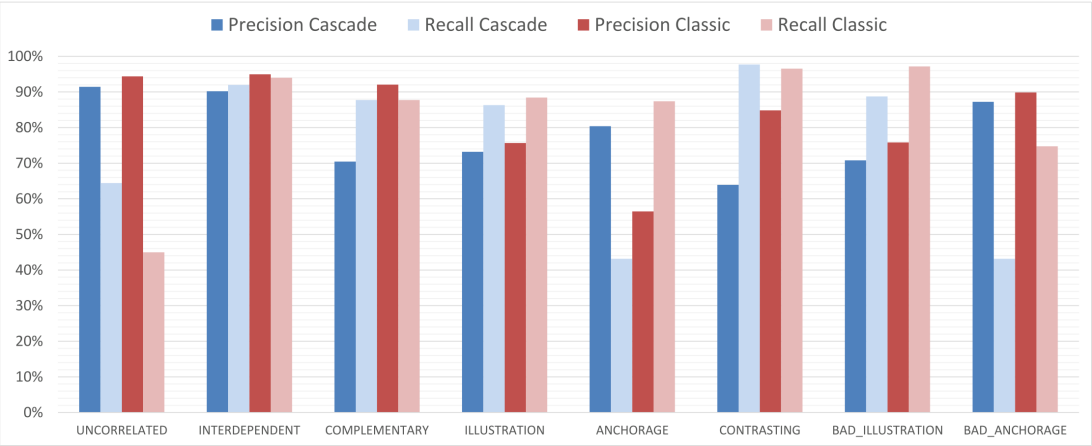


FIGURE 5.9: Results for both classifiers.

5.3.8 Discussion

As shown by Tables 5.4 and 5.5, the *classic* approach outperformed the *cascade* method by about 6% in terms of accuracy, indicating that a direct prediction of the image-text class is to be preferred over a combination of three separate classifiers. A reason might be that an overall judgment of the connection between image and text is probably more accurate than combining the independent ones, because all aspects of the multimodal message are regarded. This is also pleasant since an application would only need to train one classifier instead of three. Nonetheless, as can be seen in Table 5.7, results of the single metric classifiers suggest that they are still useful for applications that require just a single dimension, e.g., CMI for image captioning tasks. Regarding the image-text classes *Uncorrelated* achieved the lowest recall indicating that both classifiers often detected a connection (either in the SC dimension or CMI), even though there was none. This might be due to the concept detector contained in InceptionResnetV2 focusing on negligible background elements that a human would not consider to be of importance (cf. Section 5.3.2). However, the high precision indicates that if it was detected it was almost always correct, in particular for the cascade classifier. The classes with positive SC are mainly confused with their negative counterparts, which is understandable since

the difference between a positive and a negative SC is often caused by a few keywords in the text. But the performance is still impressive when considering that positive and negative instances differ only in a few keywords, while image content, sentence length and structure are identical.

The "cascaded" classifier struggled the most with both *Anchorage* classes, confusing them with *Complementary* and *Contrasting*. This is an indicator indicates that the Status classifier failed to identify that the text is subordinate and as can be seen in Table 5.6, it has indeed the lowest recall of the three dimensions. Another interesting observation can be reported regarding the cascade approach: the rejection class *Undefined*, which is predicted if an invalid leaf of the categorization (the crosses in Fig. 5.4) is reached, can be used to judge the quality of our categorization. In total, 10 out of 18 leaves represent such an invalid case, but only 27 image-text pairs (3.4%) of all test samples were assigned to it. Thus, the distinction seems to be of high quality which is due to the good results of the classifiers for the individual metrics (Table 5.7).

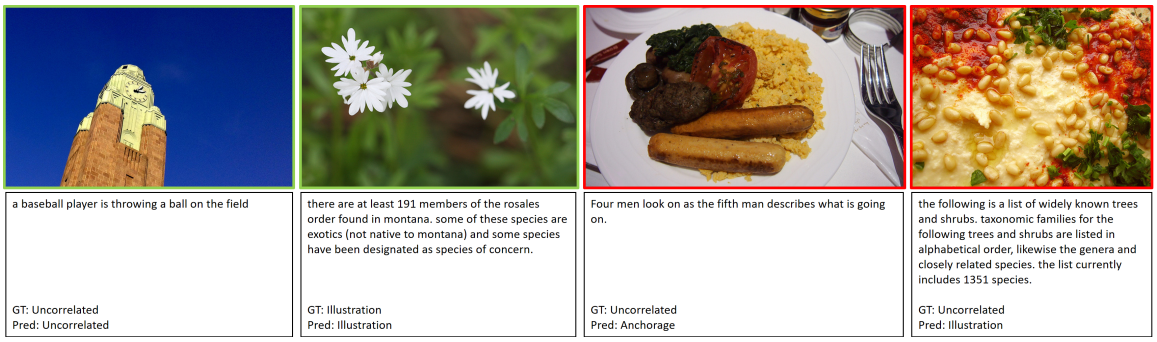


FIGURE 5.10: Example predictions of the "classic" classifier. Green box: correct prediction; Red box: false prediction.

Fig. 5.10 shows some examples for correctly and incorrectly predicted image-text pairs. The third column in this Figure shows a false prediction of an uncorrelated pair as anchorage. There were some errors of false positives for anchorage (or illustrations), which seem to be partially caused by the typically corresponding shorter (or longer) text length. But the overall results indicate that the system does not solely rely on this feature, of course, otherwise a distinction of eight classes of this quality would not have been achievable. This is supported also by the correctly predicted example in Fig. 5.10, left, where despite the short text the image-text pair is classified as uncorrelated (and not as anchor).

5.4 Applicability of the proposed Image-Text Metrics

5.4.1 Motivation

In the following section we assess the applicability of the proposed image-text relations and classes in other cross-modal domains. We choose two recent works that provide datasets for their cross-modal applications and assess how the trained models from Section 5.3.7 are able to predict semantic image-text relations on unseen data. This task can be considered zero-shot since we do not fine-tune our models on the new data and directly predict our image-text metrics and classes on data from two domains unknown to these models.

The first dataset, *Conceptual Captions* by Sharma et al. [238], goes beyond simple image captioning (cf., MS COCO [151]). It processes billions of images and their associated alt-texts from the Web. Their framework makes a specific effort only to consider alt-texts of high value. The authors ensure to check for a high unique word ratio that covers various word types, low token repetition, capitalization, and the application of pornography and profanity filters. It also scans for common sequences (e.g., "click to enlarge", "stock photo"), and thus allows only 3% of candidates to pass their system. They also filter out candidates for which none of the text tokens can be mapped to parts of the image. Thus, all image-text classes with $CMI = 0$ are dismissed depending on the chosen concept classification algorithm. Nonetheless, we hope that our predicted image-text combinations go beyond the *Anchorage* class and present a certain variety in semantic image-text classes. There are no further annotations provided.

Second, Vempala and Preotiuc-Pietro [271] published a Tweet-based dataset (*Twitter Dataset* from here on). Their goal is to determine the cross-modal relationship in terms of image-text classes similar to our the approach in Section 5.3. The main difference between our categorization and theirs is that we attempt a general approach, while they focus solely on the prediction of possible image-text relations with positive SC that appear in Twitter posts, see Figure 5.11.

(a) Image adds to the tweet meaning & Text is represented in image

(b) Image adds to the tweet meaning & Text is not represented in image

(c) Image does not add to meaning & Text is represented in image

(d) Image does not add to meaning & Text is not represented in image

FIGURE 5.11: Examples of the four classes presented in Vempala and Preotiuc-Pietro's work. According to the categorization in Section 5.3 example (a) is *Anchorage*, (c) is *Complementary*, (b) and (d) are *Interdependent*. Source: [271].

We are not able to use the labels provided by the authors directly since their definitions differ from the metrics discussed in Section 5.3. As shown in Figure 5.11, Vempala and Preotiuc-Pietro [271] distinguish between different relations with regards to 1) whether the text is represented in the image (similar to our *CMI* metric) and 2) the image adds to the meaning of the text or not. The second metric is inspired by whether image and text convey their information in parallel, meaning they generally portray similar information and are thus, independent or not (cf. *equal* Status in Figure 2.14). If not, they either complement each other (*Complementary*) or their meaning is multiplied (*Interdependent*). However, even though the first metric captures the difference between $CMI = 0$ and $CMI = 1$, the different combinations associated with the other metric combinations can not be expressed. Conversely, examples (b) and (d) in Figure 5.11 are of the *Interdependent* class by our definition, but the authors again distinguish between whether the image adds to the text meaning or not. This distinction is, however, only uni-directional ("image adds to text") and subjective to an extent. For example, Figure 5.11 (d) is labeled as "image does not add to the text", while one could argue that the emotions on the person's face convey the author's emotion about the finalization of their bachelor's studies.

Consequently, we manually label 1000 samples of the *Twitter Dataset* and the *Conceptual Caption* dataset to evaluate the performance of our proposed classifiers. Our goal is to assess how well the proposed deep neural networks are able to generalize the intricate cross-modal relations. Since the following experiments took place two years after the publication of the work in Section 5.3, we decided to replace some of the components of the multimodal embedding with newer or other components. Specifically, we replaced the image encoder with a *ResNet-101* [99] and the text encoder with *Bert_base* [60]. The improvements are documented in Table 5.8. We achieve an overall better performance in all three metrics, from 3.31% for CMI, 3.5% for SC, and 10.56% for STAT. Surprisingly, the approach without the attention mechanism achieved the best overall score.

Image-Encoder	Text-Encoder	Attention	CMI	SC	Stat	IT Class
ResNet-101	BERT-base	no	93.61%	88.1%	94.36%	87.59%
ResNet-101	BERT-base	yes	93.23%	87.47%	93.86%	86.97%
InceptionResnetV2	fastText	yes	90.3%	84.6%	83.8%	80.8%

Table 5.8: Performance of the improved versions of the multimodal embedding approach proposed in Section 5.3.7. The last line is the old model for comparison. The highlighted model in line one was used for the following experiments.

5.4.2 Experiments

The following Table 5.9 shows the distribution of image-text classes in the two manually annotated subsets.

The annotation shows that *Conceptual Captions*, as expected, consists of around 60% *Anchorage* samples, but also a fair amount of *Interdependent* (16.9%) and *Complementary* (16.7%) image-text pairs in addition to samples with negative semantic correlation. As

Class	Conceptual Captions	Twitter dataset
Uncorrelated	8	5
Interdependent	169	737
Complementary	167	162
Illustration	11	14
Anchorage	597	74
Contrasting	3	2
Bad Illustration	26	0
Bad Anchorage	19	6
Sum	1000	1000

TABLE 5.9: The distribution of semantic image-text classes after manual annotation of 1000 samples of the *Twitter dataset* and *Conceptual Captions* dataset.

for the *Twitter Dataset*, the majority (73.7%) of samples represent the *Interdependent* class, presumably due to the way information is portrayed in social media in the form of memes and loose references to the visual content (compare again Figure 5.11. The remainder of the *Twitter Dataset* is comprised of *Complementary* and *Anchorage* samples, and also a few *Illustrations*. Following Section 5.3 we conduct the classification once directly ("classic") and in a cascaded manner.

Twitter Dataset Discussion

Class	Uncorr.	Interdep.	Compl.	Illustr.	Anchorage	Contr.	Bad Illustr.	Bad Anch.	Sum
Uncorrelated	**0**	1	3	0	1	0	0	0	5
Interdependent	16	**173**	140	1	407	0	0	0	737
Complementary	4	24	**36**	0	97	0	0	1	162
Illustration	0	3	3	**0**	8	0	0	0	14
Anchorage	2	9	15	0	**48**	0	0	0	74
Contrasting	0	0	0	0	2	**0**	0	0	2
Bad Illustration	0	0	0	0	0	0	**0**	0	0
Bad Anchorage	0	1	0	0	5	0	0	**0**	6
Precision	0.0%	81.99%	18.27%	0.0%	8.45%	-	-	0.0%	1000
Recall	0.0%	23.47%	22.22%	0.0%	64.86%	-	-	0.0%	1000
Section 5.3.7									
Precision	0.00%	79.17%	0.00%	2.23%	5.71%	0.00%	0.00%	0.00%	1000
Recall	0.00%	43.83%	0.00%	85.71%	2.70%	0.00%	0.00%	0.00%	1000

TABLE 5.10: The results of the *Twitter dataset* examination using the direct classic approach. It achieved an accuracy of 25.70% while the model from Section 5.3.7 achieved 33.70%. The rows depict true positives (bold) and false negatives, i.e., the distribution of the ground-truth samples over the eight classes. The columns show the true positives (bold) and false positives and thus, the samples that the model actually identified as the respective class.

As can be seen in Table 5.10 the classifier achieved an overall accuracy of 25.7%, which is lower than the performance of the model from Section 5.3.7 as it achieved an approximately 20% higher recall on the majority class *Interdependent*. Still, we can consider both a mild success because 73.7% of the dataset consists of this intricate class, whereas the underlying dataset gathered in Section 5.3.5, provides merely 1007 samples for the training process. Further, their positive semantic correlation is often based on

humor, irony, or sarcasm, which is difficult to detect. The model was able to identify 173 samples (23.47%) and mistook the rest for the *Complementary* and *Anchorage* class, which is, however, still the correct guess for the semantic correlation metric. It is also positive that only 16 image-text pairs were labeled as *Uncorrelated*, even though there was no overlap in visual/textual concepts and entities in the 737 *Interdependent* samples. Also, only a few classes were falsely labeled as *Interdependent* as the class achieved a high precision of 81.99%. Besides the *Anchorage* class, which yielded the highest recall of 64.86%, possibly due to the high number of samples in the training data, the other results are (as expected for a full domain transfer) unsatisfactory. The model detected the second most common class, *Complementary*, only 36 out of 162 times and falsely assigned it to *Anchorage* 97 times. However, the distinction between these two classes is difficult as the model has to identify whether one modality provides additional information to the overall message or the caption is just a description of the shown visual content. Eight qualitative examples of correct and incorrect classified samples can be seen in Figure 5.12.

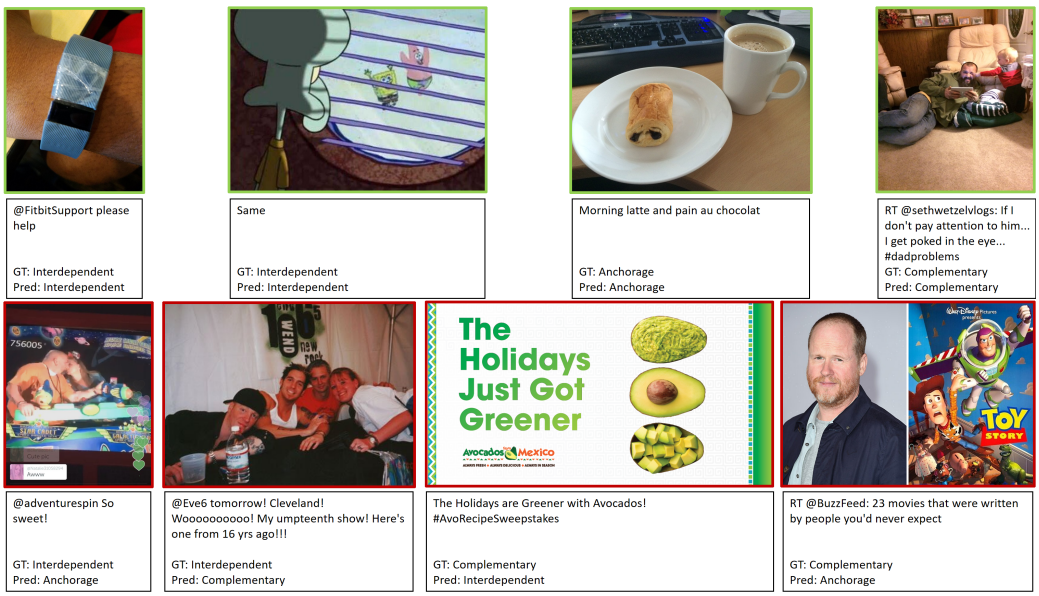


FIGURE 5.12: Four correctly classified and four misclassified examples of the Twitter dataset predicted by the the "classic" approach.

As can be seen, starting in the top row, our model correctly identified two non-trivial *Interdependent* samples with no visual overlap. Again, this is especially surprising since samples of this kind were not present in the training data. As a reminder, all training examples of this class are based on advertisements. Sample three and four depict correctly identified, non-trivial *Anchorage* and *Complementary* pairs.

The analysis of the second sample row reveals possible future extensions of the proposed categorization. Our model identified the first example as *Anchorage*, perhaps due to the referencing nature of the caption that denotes the image content as "sweet". However, there is no overlap in shown concepts or entities, which is why this image falls into the *Interdependent* category. Nonetheless, maybe the *Status* relation is not always equal for

Interdependent samples. We can further strengthen this argument by examining the second pair that, again, has no content overlap. The part of the caption saying "[..]Here's one [..]", however, implies that the author references the image and adds additional information to it, but without actually naming what can be seen. The two image-text pairs on the bottom were, presumably, misclassified due to missing context information, e.g., how the inside of an avocado looks like.

Class	Undef.	Uncorr.	Interdep.	Compl.	Illustr.	Anchorage	Contr.	Bad Illustr.	Bad Anch.	Sum
Undefined	0	0	0	0	0	0	0	0	0	0
Uncorrelated	0	**0**	1	4	0	0	0	0	0	5
Interdependent	12	5	**165**	507	14	31	2	0	1	737
Complementary	0	0	31	**108**	4	15	3	0	1	162
Illustration	0	0	0	10	**2**	2	0	0	0	14
Anchorage	3	1	12	45	3	**10**	0	0	0	74
Contrasting	0	0	1	1	0	0	**0**	0	0	2
Bad Anchorage	1	0	2	3	0	0	0	**0**	0	6
Bad Illustration	0	0	0	0	0	0	0	0	0	0
Precision	-	0.0%	77.83%	15.93%	8.70%	17.24%	0.0%	0.0%	0.0%	1000
Recall	-	0.0%	22.39%	66.67%	14.29%	13.51%	0.0%	0.0%	0.0%	1000
Section 5.3.7										
Precision	0.00%	0.52%	81.30%	15.65%	4.88%	6.17%	0.00%	0.00%	0.00%	1000
Recall	0.00%	20.00%	42.47%	11.11%	14.29%	6.76%	0.00%	0.00%	0.00%	1000

TABLE 5.11: The results of the *Twitter dataset* examination predicted by the cascaded approach where invalid combinations of the three metrics CMI, SC, and STAT are denoted as Undefined. It achieved an accuracy of 28.5% while the model from Section 5.3.7 achived 33.90%. The rows depict true positives (bold) and false negatives, i.e., the distribution of the ground-truth samples over the eight classes (+ Undefined). The columns show the true positives (bold) and false positives and thus, the samples that the model actually identified as the respective class.

For comparison, Table 5.11 shows the results of the cascaded approach that achieved a slightly higher accuracy of 28.5%, similar to the model from Section 5.3.7. The most noticeable difference between the two methods is that the majority of the *Interdependent* samples were confused with *Complementary* rather than *Anchorage*. While this is closer to the right decision since the *Status* metric is identified as equal, the *CMI* model falsely picked up an overlap for the majority of pairs. Again though, this method was able to achieve a high precision (77.83%) for the identified *Interdependent* samples and confuse only five of them with *Uncorrelated*.

Conceptual Captions Dataset Discussion

The classification of the image-text pairs from the Conceptual Captions dataset yielded an overall better result when compared to the Twitter Dataset by means of accuracy (54.3% for classic, 55.9% for cascaded). Presumably, the less challenging *Anchorage* class, which constitutes 59.7% of the samples, is the reason for that. However, the original model from Section 5.3.7 performed significantly worse than that, with accuracies of 12.80% for classic and 10.80% for cascaded. It classified the majority of *Anchorage* samples as either *Uncorrelated* (305) or *Interdependent* (179). It is challenging to determine the cause for this performance drop with certainty. An educated guess would be that the original model overfitted on a certain structural feature of the *Anchorage* samples, for example, text length.

Since this feature was most likely different in the Conceptual Captions dataset, it could not identify this class correctly. We will discuss these and other limitations in more detail in Section 6.2.

Class	Uncorr.	Interdep.	Compl.	Illustr.	Anchorage	Contr.	Bad Illustr.	Bad Anch.	Sum
Uncorrelated	**0**	1	1	0	6	0	0	0	8
Interdependent	10	**22**	4	0	132	0	0	1	169
Complementary	14	9	**5**	2	136	0	0	1	167
Illustration	1	2	0	**1**	7	0	0	0	0
Anchorage	55	11	4	0	**514**	0	0	13	597
Contrasting	0	0	0	0	3	**0**	0	0	0
Bad Illustration	2	0	2	0	21	0	**0**	1	0
Bad Anchorage	2	0	0	0	16	0	0	**1**	19
Precision	0.0%	48.89%	31.25%	33.33%	61.56%	-	-	5.88%	1000
Recall	0.0%	13.02%	2.99%	9.09%	86.10%	-	-	5.26%	1000
Section 5.3.7									
Precision	0.00%	19.56%	0.00%	1.73%	75.00%	0.00%	0.00%	0.00%	1000
Recall	0.00%	57.99%	0.00%	72.73%	3.52%	0.00%	0.00%	0.00%	1000

TABLE 5.12: The results of the *Conceptual Captions* examination predicted by the classic approach. It achieved an accuracy of 54.3% while the model from Section 5.3.7 achived 12.70%. The rows depict true positives (bold) and false negatives, i.e., the distribution of the ground-truth samples over the eight classes. The columns show the true positives (bold) and false positives and thus, the samples that the model actually identified as the respective class.

For the newer implementation, as can be seen in Tables 5.12 and 5.13, both models were able to identify a promising number of *Anchorage* pairs with recalls of 86.1% for classic and 79.23% for cascaded. The corresponding precision values are 61.56% and 71.99%. Besides that, both approaches also achieved moderate precision scores (48.89% for classic and 41.12% for cascade) for the 169 *Interdependent* samples but paired with low recall (13.02% and 26.04%). For this discussion, we want to take a closer look at the more intricate image-text classes found in the dataset.

Figure 5.13 shows four correct and four incorrectly classified samples. They start with two *Interdependent* pairs that successfully picked up the semantic connection between the shown visual and textual information. This is followed up by a *Bad Anchorage* where the model was able to identify that the shown animal is not a crane but a stork. The fourth sample shows a typical *Complementary* sample whose text provides additional information to the image. The second row of results shows instances where the model failed. On the left, it can be assumed that the model associated the word "photography" with the camera and therefore concluded on *Complementary* rather than *Interdependent*. Similarly, the second example shows no overlap of entities and concepts. It also provides another example of an *Interdependent* image-text pair with a reference. The third example shows another possible drawback of the underlying dataset: objects are only represented in a realistic setting. As a result, the skyscrapers in this digital image were not detected, and the model decided on the wrong class. However, it is common to find imagery from drawings, comics, video games, or virtual reality in modern media. The final example was a mix-up between *Anchorage* and *Bad Anchorage*. We can not see a tour boat in the

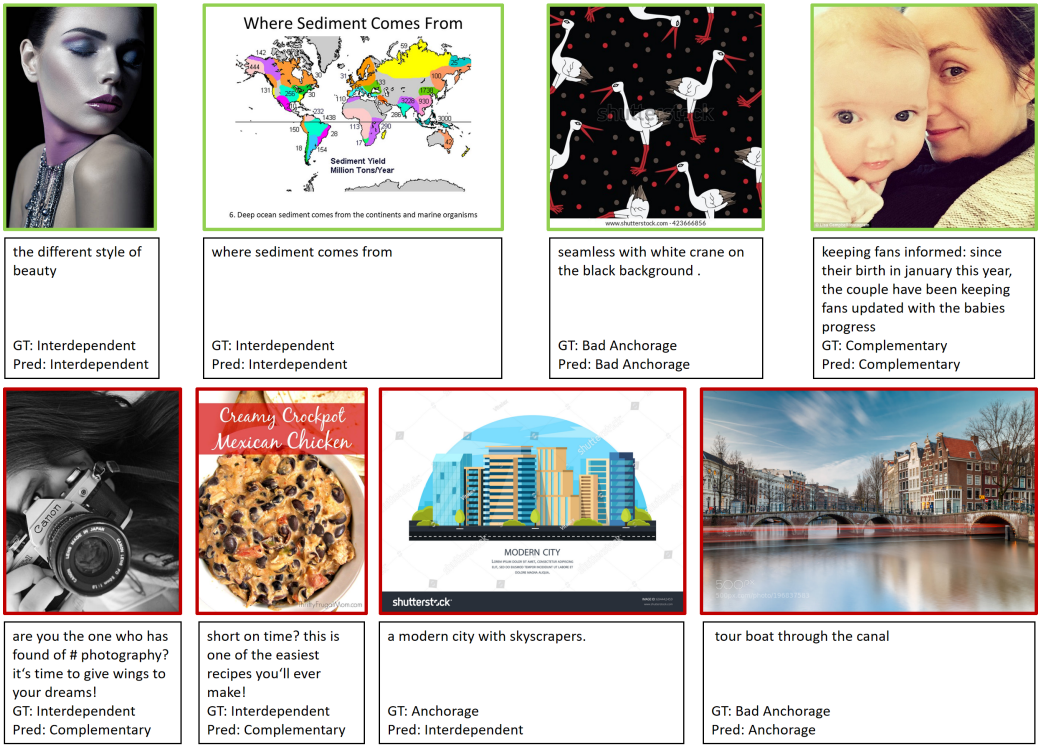


FIGURE 5.13: Four correctly classified and four misclassified examples of the Conceptual Captions dataset predicted by the "classic" approach.

image, just a light trail. Again, this type of negative semantic correlation was not present in the training data.

Class	Undefined	Uncorr.	Interdep.	Compl.	Illustr.	Anchorage	Contr.	Bad Illustr.	Bad Anch.	Sum
Undefined	**0**	0	0	0	0	0	0	0	0	0
Uncorrelated	2	**0**	3	1	0	2	0	0	0	8
Interdependent	12	5	**44**	34	1	72	0	1	0	169
Complementary	10	7	27	**40**	4	75	1	0	3	167
Illustration	0	0	3	3	**1**	4	0	0	0	0
Anchorage	34	31	23	23	0	**437**	1	0	12	597
Contrasting	0	0	0	0	0	3	**0**	0	0	0
Bad Illustration	2	2	3	2	0	17	0	**0**	0	0
Bad Anchorage	1	0	4	2	0	11	0	0	**1**	19
Precision	-	0.0%	41.12%	38.10%	16.67%	71.99%	0.0%	0.0%	6.25%	1000
Recall	-	0.0%	26.04%	23.95%	9.09%	79.23%	0.0%	0.0%	5.26%	1000
Section 5.3.7										
Precision	0.00%	0.42%	21.70%	18.31%	0.00%	52.17%	0.00%	0.00%	0.00%	1000
Recall	0.00%	25.00%	40.83%	7.78%	0.00%	4.02%	0.00%	0.00%	0.00%	1000

TABLE 5.13: The results of the *Conceptual Captions* examination predicted by the cascaded approach where invalid combinations of the three metrics CMI, SC, and STAT are denoted as Undefined. It achieved an accuracy of 55.9% while the model from Section 5.3.7 achived 10.80%. The rows depict true positives (bold) and false negatives, i.e., the distribution of the ground-truth samples over the eight classes (+ Undefined). The columns show the true positives (bold) and false positives and thus, the samples that the model actually identified as the respective class.

5.4.3 Conclusion

This section investigated whether the samples in the dataset curated in Section 5.3.5 suffice to model the proposed image-text relations in a way that allows the trained models to detect them reliably in unseen data as well. The results give us confidence that the model somewhat generalized the task. However, the achieved performance is not yet satisfactory. We can assume that the reasons for this are the lack of diversity of the underlying data sources, the class imbalance, and the basic approach to model the shared multimodal embeddings. More recent works in related tasks such as fake news detection [242] incorporate real-world information with the help of, e.g., knowledge graphs [302], and put more effort into modeling the inter- and intramodal relationship and the feature fusion process [222]. However, this goes beyond the scope of this thesis and is topic for future work.

5.5 Summary

Research Question 3

Based on insights from linguistics and visual communications, how can we derive computational models that describe the relationship between image and text?

In this chapter, we made an effort to understand the fundamentals of meaning-making in the domain of visual and textual information. With regards to research question 3, we showed how we can transfer image-text relations from linguistics to computer science. In particular, we proposed definitions for three image-text metrics (CMI, SC, STAT), explained their origins and showcased examples. We derived eight semantic image-text classes based on these metrics and consolidated them in a categorization. Afterward, multiple data augmentation techniques were utilized to gather a dataset of image-text pairs from various Web sources and cross-modal datasets. With this, we trained two CNN-based classifiers to predict 1) our metrics individually and 2) the image-text classes directly.

Finally, to determine our models' applicability and ability to generalize to unseen content, we evaluated them on two unseen datasets. The experiments revealed that our models can predict image-text classes "in-the-wild" with an accuracy of 54.4% for *classic* and 55.9% for *cascaded*. Regarding generalizability, the results indicate that the model is able to distinguish between uncorrelated content and samples that appear uncorrelated (*Interdependent*) because they share no concepts or entities (CMI = 0), but convey their message through, for instance, irony, humor, or metaphors.

◇ ◇ ◇

The next chapter summarizes the thesis, outlines the findings and contributions, discusses their limitations, and details areas for future work.

6 Conclusions

This chapter summarizes the thesis, gives an overview of the contributions (Section 6.1), and relates the findings to the research questions described in Chapter 1. In Section 6.2 we discuss the limitations of the proposed methods, while Section 6.3 outlines possible avenues for future work on the respective research tasks.

6.1 Summary and Contributions

Research Question 1

How can we utilize textual metadata associated with learning content to improve exploratory search in video search portals?

The experiments in this thesis shed light on various parts of the Web-based learning process, starting in Chapter 3. Concerning **Research Question 1**, we investigated how the TIB AV-Portal, a scientific video Web platform, can be improved by means of machine learning methods, specifically NLP-based methods. We proposed two algorithmic solutions to improve the explorative search capabilities. First, we implemented a recommender system that considers content similarities derived via semantic word embeddings from textual metadata. We showed how to supplement the video metadata with external resources to strengthen these recommendations. A usability test revealed that this approach can return satisfactory results and that we can utilize external resources to improve the quality of the rather superficial keywords (automatically generated by the AV-Portal system). Second, we implemented a visualization technique to summarize the content of the longer videos in the learning domain. With the help of semantic word embeddings we could consolidate the topics of a video in an interactive, explorable visualization. Evaluation of the correctness of the visualizations with the help of a user study showed that our approach produced accurate representations of the video content.

Research Question 2

To what extent can we extract textual, multimedia, and cross-modal features and utilize them for knowledge gain prediction?

Chapter 4 has gone beyond considering only text-based information. Our goal was to obtain a richer representation of the modalities present in a learning resource, that is visual and auditory features. Considering **Research Question 2**, we estimate how useful the individual modalities are with respect to the task of KG prediction. This

assessment entailed design aspects that describe the individual compositions of slide content, visual content features like the image type, acoustic features that describe the lecturer's presentation, and cross-modal features that combine some of these aspects. Also, we extended the list of textual features by adding syntactic, structural, lexical, and readability metrics. In Section 4.1.6, we have presented a broad set of experiments that evaluate the performance of four different classifiers on all possible combinations of multimedia, text, and sentence embedding feature-sets. The results suggest that depending on the two tasks, namely 1) KG prediction for a specific video for an arbitrary learner or 2) predicting the individual KG of a learner on a given video, different feature sets have to be applied for better results. We have found indications that multimodal and cross-modal features that describe the visual layout of the content and the auditory presentation quality are a better indicator for predicting the user-independent KG for an individual video. Conversely, textural features that describe, for instance, the readability, complexity, and structure of the content are more suited to predict the KG for a video when the learner's personal performance is considered. In summary, these results hint at which modality to focus on, depending on the task. They also imply that the consideration of multimodal features, besides features representing text content, might be beneficial, and dismissing cues could be suboptimal. As shown in Section 4.1.6, a feature importance analysis can help identify the optimal subset for the task at hand.

Next, our work in Section 4.2 has considered another set of multimedia features in conjunction with textual features. Therefore, we utilized a dataset of log data from a larger user study. We proposed an almost automatic *Multimedia Extraction Framework* to gather statistics about the Web-based content each learner saw. These statistics contained information about the document layout and image content. Our correlation analysis in Section 4.2.6 indicated that (1) learners with low pre-knowledge preferred audio-visual content more than textual. Presumably, video content is more suited to give a first impression of a topic, while learners can explore textual content more efficiently to close gaps in existing knowledge. Finding (2) showed that a low pre-knowledge state correlated with a larger observed image size (i.e., video frames), which supports this hypothesis. Nonetheless, finding (3) showed that if learners with higher pre-knowledge utilized websites with images, a large and embedded layout is favorable compared to thumbnail-type images. Further, we showed that pre- and post-knowledge states correlate with some of the observed image types.

Research Question 3

Based on insights from linguistics and visual communications, how can we derive computational models that describe the relationship between image and text?

In Chapter 5, we have explored the subareas of linguistics and visual communications that have studied the fundamentals of meaning-making through different modalities for decades. Our goal, as outlined by **Research Question 3**, was to exploit interdisciplinary

knowledge regarding interpretable, computable metrics that describe the relationship between images and their associated text. We have considered research from computer science and communication science and suggested to extend a set of two image-text metrics (CMI and SC), as introduced by previous work, by a third metric (STAT). From there we derived a categorization of eight semantic image-text classes. These metrics describe fundamentally the different ways how information in visual and textual information are connected in order to convey an intended message. We have collected an extensive dataset with image-text pairs by augmenting and combining multimedia datasets and other Web resources. Further, we have proposed a multimodal neural network architecture for classification that predicts the individual metrics as well as image-text classes with satisfactory accuracy. Finally, we have investigated how well these models generalize this task by applying the trained models to challenging, unseen data in form of 1) a Tweet-based dataset and 2) a dataset of images from the Web and their alt-texts. On the one hand, the experiments achieved mediocre results (54.4% and 55.9% accuracy) for the task of image-text class prediction, which requires all three underlying metrics to be correct (cf. Section 5.4.3). However, the models' ability to detect a positive semantic correlation for unseen samples of the *Interdependent* class was surprisingly good. These results indicate that, by increasing the variance and sample size of the underlying dataset, the models can detect the proposed metrics and classes in "in-the-wild"-scenarios.

6.2 Limitations

This section outlines the limitations to consider when interpreting our findings of Chapters 3, 4, and 5.

6.2.1 Chapter 3: Improving Video Learning Platforms with Text-Based Features

The method presented in Section 3.1 is limited by the fact, that its implementation requires a preceding extraction of the semantic keywords. Furthermore, the scale of the conducted user study in the evaluation was small considering the fact that we employed non-experts of the respective video topics. Presumably, an expert review of the recommendations would have provided more reliable results. These two limitations also apply to the work in Section 3.2, however, the usefulness questionnaire we conducted in this experiment also revealed another limitation. In order to find out what led participants to give the rating "uncorrelated", we reviewed these six videos and found that their content revolved around the subject of engineering and had very application-specific content, which might be a limitation of the system. One video, for instance, discusses the cause, consequences, and solutions of driftwood accumulation on bridges leading to overflowing rivers (https://av.tib.eu/media/11442). A lot of technical terms, switching contexts from the real world to model testing to technical considerations, paired with topic-specific phrases yielded a visualization that was only marginally helpful. Finally, more results

present a ”good match“ instead of an exact match due to the nature of the entities extracted from the speech transcript. For example, videos and tutorials from the field of mathematics contain many important terms when explaining a concept that are rather general and not closely related to the topic itself. That includes words like ”square“, ”point“ and ”integral“. Yet, these words are captured by the automatic annotation tool of the TIB AV-Portal and are present in the dataset, but they contribute only marginally to the comprehension of the video even though they appear very frequently. A respective less useful summary result is exemplarily presented in Figure 3.7. This limitation of our proposed model is also reflected in the results of Task II, where our participants agreed that the visualization would be more helpful if we omitted the redundant keywords.

6.2.2 Chapter 4: Prediction of Knowledge Gain with Multimodal Features

The main limitation for Section 4.1 is the size of the user study that was set in the MOOC learning evironment (13 participants). Larger studies with more users and a variety of videos have a better chance to produce reliable results. Regarding Section 4.2, one limitation of our approach is the generalizability to other learning scenarios since we focused on just one task. Our approach was to conduct one extensive, well-structured learning task rather than multiple small ones. As we have seen in the experiment, the Internet provides learning resources of various layouts for this particular topic. However, we are aware that the characteristics of a task influence the appearance and quantity of the respective results (e.g., text, pictures, videos) and their characteristics (e.g., quality, complexity, visualizability). Especially the underlying knowledge type of a task could influence the distribution of findable and suitable multimedia content for learning. However, considering the findings by Gadiraju et al. [78], who found relationships between learners' prior topic familiarity and their navigation and query behavior across different topics, we think that our findings could be generalized to other SAL tasks that are comparable in complexity and structure. One example would be the topic ”photosynthesis“, since it is classified as learning about causal concepts [267], similar to our topic about the formation of thunderstorms and lightning. Learning about causal concepts requires the understanding of 1) different concepts (e.g. cloud, electricity, lightning) and 2) causalities (e.g., why is the cloud getting charged through particles). This knowledge is obtainable through text, videos, and images considered in our experiments.

6.2.3 Chapter 5: Semantic Image-Text Relations

The categorization proposed in Section 5.3 is built upon work from computer science and linguistics. Bateman [16] critically discusses the advantages and disadvantages of the most popular image-text taxonomies and highlights that the interpretation of content represented in the visual and textual modality can be subjective. We aimed at alleviating this effect: first, we have examined different approaches to model image-text relations and focused on their similarities rather than dissimilarities under the assumption that they

would agree on the most fundamental metrics. Second, we have represented our semantic image-text classes with simple samples from Web sources as described in Section 5.3.5. As the inter-coder agreement highlighted (cf. Figure 5.3), this resulted in a dataset with low subjectivity. However, still some of the assigned labels in the dataset presented in Section 5.3 might be subject to discussion. Also, for broad classes like *Complementary* and *Interdependent*, it has to be stated that they are not representative in their current state, as was also shown in Section 5.4. As the mediocre results in this section have shown, this limits their ability to generalize to unseen content.

6.3 Future Work

While this thesis has presented multiple insights in the research field *Web-based Learning*, various interesting research questions remain open. The sample applications in Chapter 3, that are built upon textual features to improve the TIB AV-Portal, could be adapted for different video domains, considering more or other metrics and utilizing additional modalities. Also, their impact should be evaluated more thoroughly by implementing them into a productive system. The examination of a real learner's search behavior and success might give more realistic results than the independent evaluation of, for instance, the visualization in Section 3.2.

The user studies in Chapter 4, concerning the knowledge gain prediction in informal learning settings, should be repeated on additional topics, more diverse participant groups, and learning tasks to see whether the findings can be generalized. Also, the set of content-related features, textual and visual, used in these approaches can be extended by, for example, transformer-based deep features.

Finally, the ideas to describe the interplay of images and text to convey meaning presented in Chapter 5 offer further opportunities for future work. First, additional metrics might need to be considered for a holistic view derived from either previous research or real-world image-text pairs whose relationship can not be expressed properly, yet. Second, following Henning and Ewerth [104], the proposed metrics could be subdivided into multiple levels to get a precise description of, for instance, the overlap of mutual information between image and text. Third, to improve the "in-the-wild" performance of the trained models, as shown in Section 5.4, it might be beneficial to infuse real-world knowledge into the models. This might empower these models to improve the results of applications like Fake News detection because they could detect inaccuracies on person, event and location level similar to Müller-Budack et al. [190]. Fourth, better automatic solutions for dataset generation, as proposed in Section 5.3.5 would support all of the above topics. This entails more diverse data sources to get a more unbiased representation of what, for example, positive SC might look like. Also, digital-born images need to be considered and not only photos. The most important relationships this refers to are 1) negative semantic correlation because there are more types of contradictions between

image and text than shown in this thesis. For example, recent work by Parcalabescu [206] shows a handful of methods to falsify image captions automatically. Moreover 2), the definition of the *Interdependent* class goes way beyond the currently available advertisement category in our dataset. As the qualitative examples in Section 5.4 highlighted, the multimodal expression on audiovisual platforms like Twitter or in movies presents diverse combinations of *Interdependent* image-text pairs that are not advertisements.

Appendices

A Resource Features

The following table contains the entire list of resource features extracted in [202].

	notation	description
complexity	c_adj	Ratio of adjectives to total number of words
	c_char	Average number of characters per term
	c_fk	Flesh-Kincaid Grad Readability Index
	c_gi	Gunning Fog Grade Readability Index
	c_noun	Ratio nouns to total number of words
	c_oth	Ratio other words to total number of words
	c_sentence	Average number of words per sentence
	c_smog	SMOG Readability Index
	c_uniq_word	Ratio of unique words to total number of words
	c_verb	Ratio of verbs to total number of words
	c_word	Number of words in each web page
HTML structure	h_aud	Number of <audio>element
	h_img	Number of <img>elements
	h_obj	Number of <object>element
	h_nav_ul	Number of elements in<nav>elements
	h_oth_ul	Number of elements not in <nav>elements
	h_p	Avg. length of paragraphs in <p>elements
	h_script	Number of <script>elements \hline
	h_vid	Number of <video>elements
Linguistic	l_achieve	Number of achievement words
	l_adj	Number of adjectives
	l_adverb	Number of common adverbs
	l_affect	number of affect words
	l_affiliation	Number of affiliation words
	l_AllPunc	Number of punctuation
	l_Analytic	Number of analytic words
	l_anger	Number of anger words

	notation	**description**
Linguistic	l_anx	Number of anxiety words
	l_Apostro	Number of apostrophes
	l_article	Number of articles
	l_assent	Number of assent words
	l_Authentic	Number of authentic words
	l_auxverb	Number of auxiliary verbs
	l_bio	Number of biological process words
	l_body	Number of body words
	l_cause	Number of causal words
	l_certain	Number of certainty words
	l_Clout	Number of clout words
	l_cogproc	Number of cognitive process words
	l_Colon	Number of colons
	l_Comma	Number of commas
	l_compare	Number of comparatives
	l_conj	Number of conjunctions
	l_Dash	Number of dashes
	l_death	Number of death words
	l_Dic	Number of dictionary words
	l_differ	Number of differentiation words
	l_discrep	Number of discrepancy words
	l_drives	Number of core drives and needs
	l_Exclam	Number of exclamation marks
	l_family	Number of family words
	l_feel	Number of feeling words
	l_female	Number of female referents
	l_filler	Number of filters
	l_focusfuture	Number of future focus words
	l_focuspast	Number of past focus words
	l_focuspresent	Number of present focus words
	l_friend	Number of friend words
	l_function	Number of function words
	l_health	Number of health words
	l_hear	Number of hearing words
	l_home	Number of home words
	l_i	Number of I pronouns
	l_informal	Number of informal speech words
	l_ingest	Number of ingesting words
	l_insight	Number of insightful words
	l_interrog	Number of interrogatives
	l_ipron	Number of impersonal pronouns
	l_leisure	Number of leisure words
	l_male	Number of male referents
	l_money	Number of money words
	l_motion	Number of motion words
	l_negate	Number of negations

	notation	**description**
Linguistic	l_negemo	number of negative emotional words
	l_netspeak	Number of netspeak words
	l_nonflu	Number of nonfluencies
	l_number	Number of numbers
	l_OtherP	Number of other punctuation
	l_Parenth	Number of parentheses (pairs)
	l_percept	Number of perceptual processes
	l_Period	Number of periods
	l_posemo	Number of positive emotion words
	l_power	Number of power words
	l_ppron	Number of personal pronouns
	l_prep	Number of prepositions
	l_pronoun	Number of total pronouns
	l_QMark	Number of question marks
	l_quant	Number of quantifiers
	l_Quote	Number of quotation marks
	l_relativ	Number of relativity words
	l_relig	Number of religion words
	l_reward	Number of reward focus words
	l_risk	Number of risk words
	l_sad	Number of sadness words
	l_see	Number of seeing words
	l_SemiC	Number of semicolons
	l_sexual	Number of sexual words
	l_shehe	Number of she or he pronouns
	l_Sixltr	Number of words with more then 6 letters
	l_social	Number of social words
	l_space	Number of space words
	l_swear	Number of swear words
	l_tentat	Number of tentative words
	l_they	Number of they pronouns
	l_time	Number of time words
	l_Tone	number of emotional tone words
	l_verb	Number of regular verbs
	l_we	Number of we pronouns
	l_work	Number of work words
	l_you	Number of you pronouns

B Full Textual Feature List

The following table contains the entire list of textual features extracted in [201].

Feature	Description	Category
amount_adj_sli	Num. of adjectives in the slides	Syntactic (word types)
avg_adj_sli	Average Num. of adjectives per line in the slides	Syntactic (word types)
ratio_adj_sli	Relation of the Num. of adjectives to unfiltered tokens in the slides	Syntactic (word types)
amount_adpos_sli	Num. of adpositions in the slides	Syntactic (word types)
avg_adpos_sli	Average Num. of adpositions per line in the slides	Syntactic (word types)
ratio_adpos_sli	Relation of the Num. of adpositions to unfiltered tokens in the slides	Syntactic (word types)
amount_noun_sli	Num. of nouns in the slides	Syntactic (word types)
avg_noun_sli	Average Num. of nouns per line in the slides	Syntactic (word types)
ratio_noun_sli	Relation of the Num. of nouns to unfiltered tokens in the slides	Syntactic (word types)
amount_pronoun_sli	Num. of pronouns in the slides	Syntactic (word types)
avg_pronoun_sli	Average Num. of pronouns per line in the slides	Syntactic (word types)
ratio_pronoun_sli	Relation of the Num. of pronouns to unfiltered tokens in the slides	Syntactic (word types)
ratio_pronoun_noun_sli	Relation of the Num. of pronouns to nouns in the slides	Syntactic (word types)
amount_verb_sli	Num. of verbs in the slides	Syntactic (word types)
avg_verb_sli	Average Num. of verbs per line in the slides	Syntactic (word types)
ratio_verb_sli	Relation of the Num. of verbs to unfiltered tokens in the slides	Syntactic (word types)
amount_main_verb_sli	Num. of main verbs in the slides	Syntactic (word types)
avg_main_verb_sli	Average Num. of main verbs per line in the slides	Syntactic (word types)
ratio_main_verb_sli	Relation of the Num. of main verbs to unfiltered tokens in the slides	Syntactic (word types)
amount_aux_sli	Num. of auxiliaries in the slides	Syntactic (word types)
avg_aux_sli	Average Num. of auxiliaries per line in the slides	Syntactic (word types)
ratio_aux_sli	Relation of the Num. of auxiliaries to unfiltered tokens in the slides	Syntactic (word types)
amount_adverb_sli	Num. of adverbs in the slides	Syntactic (word types)
avg_adverb_sli	Average Num. of adverbs per line in the slides	Syntactic (word types)

Continuation on the next page

Table B.1 – Continuation

Feature	Description	Category
ratio_adverb_sli	Relation of the Num. of adverbs to unfiltered tokens in the slides	Syntactic (word types)
amount_coord_conj_sli	Num. of coordinate conjunctions in the slides	Syntactic (word types)
avg_coord_conj_sli	Average Num. of coordinate conjunctions per line in the slides	Syntactic (word types)
ratio_coord_conj_sli	Relation of the Num. of coordinate conjunctions to unfiltered tokens in the slides	Syntactic (word types)
amount_determiner_sli	Num. of determiner in the slides	Syntactic (word types)
avg_determiner_sli	Average Num. of determiner per line in the slides	Syntactic (word types)
ratio_determiner_sli	Relation of the Num. of determiner to unfiltered tokens in the slides	Syntactic (word types)
amount_interj_sli	Num. of interjections in the slides	Syntactic (word types)
avg_interj_sli	Average Num. of interjections per line in the slides	Syntactic (word types)
ratio_interj_sli	Relation of the Num. of interjections to unfiltered tokens in the slides	Syntactic (word types)
amount_num_sli	Num. of Num.s in the slides	Syntactic (word types)
avg_num_sli	Average Num. of Num.s per line in the slides	Syntactic (word types)
ratio_num_sli	Relation of the Num. of Num.s to unfiltered tokens in the slides	Syntactic (word types)
amount_particle_sli	Num. of particles in the slides	Syntactic (word types)
avg_particle_sli	Average Num. of particles per line in the slides	Syntactic (word types)
ratio_particle_sli	Relation of the Num. of particles to unfiltered tokens in the slides	Syntactic (word types)
amount_subord_conj_sli	Num. of particles in the slides	Syntactic (word types)
avg_subord_conj_sli	Average Num. of particles per line in the slides	Syntactic (word types)
ratio_subord_conj_sli	Relation of the Num. of particles to unfiltered tokens in the slides	Syntactic (word types)
amount_foreign_sli	Num. of foreign words in the slides	Syntactic (word types)
avg_foreign_sli	Average Num. of foreign words per line in the slides	Syntactic (word types)
ratio_foreign_sli	Relation of the Num. of foreign words to unfiltered tokens in the slides	Syntactic (word types)
amount_content_word_sli	Num. of content words in the slides	Syntactic (word types)
avg_content_word_sli	Average Num. of content words per line in the slides	Syntactic (word types)
ratio_content_word_sli	Relation of the Num. of content words to unfiltered tokens in the slides	Syntactic (word types)
amount_function_word_sli	Num. of function words in the slides	Syntactic (word types)
avg_function_word_sli	Average Num. of function words per line in the slides	Syntactic (word types)

Continuation on the next page

Table B.1 – Continuation

Feature	Description	Category
ratio_function_word_sli	Relation of the Num. of function words to unfiltered tokens in the slides	Syntactic (word types)
amount_filtered_sli	Num. of filtered words in the slides	Syntactic (word types)
avg_filtered_sli	Average Num. of filtered words per line in the slides	Syntactic (word types)
ratio_filtered_sli	Relation of the Num. of filtered words to unfiltered tokens in the slides	Syntactic (word types)
amount_adj_tra	Num. of adjectives in the srt	Syntactic (word types)
avg_adj_tra	Average Num. of adjectives per sentence in the srt	Syntactic (word types)
ratio_adj_tra	Relation of the Num. of adjectives to unfiltered tokens in the srt	Syntactic (word types)
amount_adpos_tra	Num. of adpositions in the srt	Syntactic (word types)
avg_adpos_tra	Average Num. of adpositions per sentence in the srt	Syntactic (word types)
ratio_adpos_tra	Relation of the Num. of adpositions to unfiltered tokens in the srt	Syntactic (word types)
amount_noun_tra	Num. of nouns in the srt	Syntactic (word types)
avg_noun_tra	Average Num. of nouns per sentence in the srt	Syntactic (word types)
ratio_noun_tra	Relation of the Num. of nouns to unfiltered tokens in the srt	Syntactic (word types)
amount_pronoun_tra	Num. of pronouns in the srt	Syntactic (word types)
avg_pronoun_tra	Average Num. of pronouns per sentence in the srt	Syntactic (word types)
ratio_pronoun_tra	Relation of the Num. of pronouns to unfiltered tokens in the srt	Syntactic (word types)
ratio_pronoun_noun_tra	Relation of the Num. of pronouns to nouns in the srt	Syntactic (word types)
amount_verb_tra	Num. of verbs in the srt	Syntactic (word types)
avg_verb_tra	Average Num. of verbs per sentence in the srt	Syntactic (word types)
ratio_verb_tra	Relation of the Num. of verbs to unfiltered tokens in the srt	Syntactic (word types)
amount_main_verb_tra	Num. of main verbs in the srt	Syntactic (word types)
avg_main_verb_tra	Average Num. of main verbs per sentence in the srt	Syntactic (word types)
ratio_main_verb_tra	Relation of the Num. of main verbs to unfiltered tokens in the srt	Syntactic (word types)
amount_aux_tra	Num. of auxiliaries in the srt	Syntactic (word types)
avg_aux_tra	Average Num. of auxiliaries per sentence in the srt	Syntactic (word types)
ratio_aux_tra	Relation of the Num. of auxiliaries to unfiltered tokens in the srt	Syntactic (word types)
amount_adverb_tra	Num. of adverbs in the srt	Syntactic (word types)
avg_adverb_tra	Average Num. of adverbs per sentence in the srt	Syntactic (word types)

Continuation on the next page

Table B.1 – Continuation

Feature	Description	Category
ratio_adverb_tra	Relation of the Num. of adverbs to unfiltered tokens in the srt	Syntactic (word types)
amount_coord_conj_tra	Num. of coordinate conjunctions in the srt	Syntactic (word types)
avg_coord_conj_tra	Average Num. of coordinate conjunctions per sentence in the srt	Syntactic (word types)
ratio_coord_conj_tra	Relation of the Num. of coordinate conjunctions to unfiltered tokens in the srt	Syntactic (word types)
amount_determiner_tra	Num. of determiner in the srt	Syntactic (word types)
avg_determiner_tra	Average Num. of determiner per sentence in the srt	Syntactic (word types)
ratio_determiner_tra	Relation of the Num. of determiner to unfiltered tokens in the srt	Syntactic (word types)
amount_interj_tra	Num. of interjections in the srt	Syntactic (word types)
avg_interj_tra	Average Num. of interjections per sentence in the srt	Syntactic (word types)
ratio_interj_tra	Relation of the Num. of interjections to unfiltered tokens in the srt	Syntactic (word types)
amount_num_tra	Num. of Num.s in the srt	Syntactic (word types)
avg_num_tra	Average Num. of Num.s per sentence in the srt	Syntactic (word types)
ratio_num_tra	Relation of the Num. of Num.s to unfiltered tokens in the srt	Syntactic (word types)
amount_particle_tra	Num. of particles in the srt	Syntactic (word types)
avg_particle_tra	Average Num. of particles per sentence in the srt	Syntactic (word types)
ratio_particle_tra	Relation of the Num. of particles to unfiltered tokens in the srt	Syntactic (word types)
amount_subord_conj_tra	Num. of particles in the srt	Syntactic (word types)
avg_subord_conj_tra	Average Num. of particles per sentence in the srt	Syntactic (word types)
ratio_subord_conj_tra	Relation of the Num. of particles to unfiltered tokens in the srt	Syntactic (word types)
amount_foreign_tra	Num. of foreign words in the srt	Syntactic (word types)
avg_foreign_tra	Average Num. of foreign words per sentence in the srt	Syntactic (word types)
ratio_foreign_tra	Relation of the Num. of foreign words to unfiltered tokens in the srt	Syntactic (word types)
amount_content_word_tra	Num. of content words in the srt	Syntactic (word types)
avg_content_word_tra	Average Num. of content words per sentence in the srt	Syntactic (word types)
ratio_content_word_tra	Relation of the Num. of content words to unfiltered tokens in the srt	Syntactic (word types)
amount_function_word_tra	Num. of function words in the srt	Syntactic (word types)
avg_function_word_tra	Average Num. of function words per sentence in the srt	Syntactic (word types)
ratio_function_word_tra	Relation of the Num. of function words to unfiltered tokens in the srt	Syntactic (word types)

Continuation on the next page

Table B.1 – Continuation

Feature	Description	Category
amount_filtered_tra	Num. of filtered words in the srt	Syntactic (word types)
avg_filtered_tra	Average Num. of filtered words per sentence in the srt	Syntactic (word types)
ratio_filtered_tra	Relation of the Num. of filtered words to unfiltered tokens in the srt	Syntactic (word types)
sim_pres_sli	Num. of Simple Present tenses in the slides	Syntactic (temporal)
ratio_sim_pres_sli	Relation of Num. of Simple Present tenses to all phrases in the slides	Syntactic (temporal)
pres_prog_sli	Num. of Present Progressive tenses in the slides	Syntactic (temporal)
ratio_pres_prog_sli	Relation of Num. of Present Progressive tenses to all phrases in the slides	Syntactic (temporal)
pres_perf_sli	Num. of Present Perfect tenses in the slides	Syntactic (temporal)
ratio_pres_perf_sli	Relation of Num. of Present Perfect tenses to all phrases in the slides	Syntactic (temporal)
pres_perf_prog_sli	Num. of Present Perfect Progressive tenses in the slides	Syntactic (temporal)
ratio_pres_perf_prog_sli	Relation of Num. of Present Perfect Progressive tenses to all phrases in the slides	Syntactic (temporal)
sim_pas_sli	Num. of Simple Past tenses in the slides	Syntactic (temporal)
ratio_sim_pas_sli	Relation of Num. of Simple Past tenses to all phrases in the slides	Syntactic (temporal)
pas_prog_sli	Num. of Past Progressive tenses in the slides	Syntactic (temporal)
ratio_pas_prog_sli	Relation of Num. of Past Progressive tenses to all phrases in the slides	Syntactic (temporal)
pas_perf_sli	Num. of Past Perfect tenses in the slides	Syntactic (temporal)
ratio_pas_perf_sli	Relation of Num. of Past Perfect tenses to all phrases in the slides	Syntactic (temporal)
pas_perf_prog_sli	Num. of Past Perfect Progressive tenses in the slides	Syntactic (temporal)
ratio_pas_perf_prog_sli	Relation of Num. of Past Perfect Progressive tenses to all phrases in the slides	Syntactic (temporal)
will_sli	Num. of Will-Future tenses in the slides	Syntactic (temporal)
ratio_will_sli	Relation of Num. of Will-Future tenses to all phrases in the slides	Syntactic (temporal)
fu_prog_sli	Num. of Future Progressive tenses in the slides	Syntactic (temporal)
ratio_fu_prog_sli	Relation of Num. of Future Progressive tenses to all phrases in the slides	Syntactic (temporal)
fu_perf_sli	Num. of Future Perfect tenses in the slides	Syntactic (temporal)
ratio_fu_perf_sli	Relation of Num. of Future Perfect tenses to all phrases in the slides	Syntactic (temporal)
fu_perf_prog_sli	Num. of Future Perfect Progressive tenses in the slides	Syntactic (temporal)
ratio_fu_perf_prog_sli	Relation of Num. of Future Perfect Progressive tenses to all phrases in the slides	Syntactic (temporal)

Continuation on the next page

Table B.1 – Continuation

Feature	Description	Category
cond_sim_sli	Num. of Conditional Simple tenses in the slides	Syntactic (temporal)
ratio_cond_sim_sli	Relation of Num. of Conditional Simple tenses to all phrases in the slides	Syntactic (temporal)
cond_prog_sli	Num. of Conditional Progressive tenses in the slides	Syntactic (temporal)
ratio_cond_prog_sli	Relation of Num. of Conditional Progressive tenses to all phrases in the slides	Syntactic (temporal)
cond_perf_sli	Num. of Conditional Perfect tenses in the slides	Syntactic (temporal)
ratio_cond_perf_sli	Relation of Num. of Conditional Perfect tenses to all phrases in the slides	Syntactic (temporal)
cond_perf_prog_sli	Num. of Conditional Perfect Progressive tenses in the slides	Syntactic (temporal)
ratio_cond_perf_prog_sli	Relation of Num. of Conditional Perfect Progressive tenses to all phrases in the slides	Syntactic (temporal)
gerund_sli	Num. of Gerund/Present Participle tenses in the slides	Syntactic (temporal)
ratio_gerund_sli	Relation of Num. of Gerund/Present Participle tenses to all phrases in the slides	Syntactic (temporal)
perf_part_sli	Num. of Perfect Participle tenses in the slides	Syntactic (temporal)
ratio_perf_part_sli	Relation of Num. of Perfect Participle tenses to all phrases in the slides	Syntactic (temporal)
inf_sli	Num. of Present Infinitive tenses in the slides	Syntactic (temporal)
ratio_inf_sli	Relation of Num. of Present Infinitive tenses to all phrases in the slides	Syntactic (temporal)
perf_inf_sli	Num. of Perfect Infinitive tenses in the slides	Syntactic (temporal)
ratio_perf_inf_sli	Relation of Num. of Perfect Infinitive tenses to all phrases in the slides	Syntactic (temporal)
active_sli	Num. of active verb forms in the slides	Syntactic (temporal)
ratio_active_sli	Relation of Num. of active verb forms to all phrases in the slides	Syntactic (temporal)
passive_sli	Num. of passive verb forms in the slides	Syntactic (temporal)
ratio_passive_sli	Relation of Num. of passive verb forms to all phrases in the slides	Syntactic (temporal)
sim_pres_tra	Num. of Simple Present tenses in the srt	Syntactic (temporal)
ratio_sim_pres_tra	Relation of Num. of Simple Present tenses to all phrases in the srt	Syntactic (temporal)
pres_prog_tra	Num. of Present Progressive tenses in the srt	Syntactic (temporal)
ratio_pres_prog_tra	Relation of Num. of Present Progressive tenses to all phrases in the srt	Syntactic (temporal)
pres_perf_tra	Num. of Present Perfect tenses in the srt	Syntactic (temporal)

Continuation on the next page

Table B.1 – Continuation

Feature	Description	Category
ratio_pres_perf_tra	Relation of Num. of Present Perfect tenses to all phrases in the srt	Syntactic (temporal)
pres_perf_prog_tra	Num. of Present Perfect Progressive tenses in the srt	Syntactic (temporal)
ratio_pres_perf_prog_tra	Relation of Num. of Present Perfect Progressive tenses to all phrases in the srt	Syntactic (temporal)
sim_pas_tra	Num. of Simple Past tenses in the srt	Syntactic (temporal)
ratio_sim_pas_tra	Relation of Num. of Simple Past tenses to all phrases in the srt	Syntactic (temporal)
pas_prog_tra	Num. of Past Progressive tenses in the srt	Syntactic (temporal)
ratio_pas_prog_tra	Relation of Num. of Past Progressive tenses to all phrases in the srt	Syntactic (temporal)
pas_perf_tra	Num. of Past Perfect tenses in the srt	Syntactic (temporal)
ratio_pas_perf_tra	Relation of Num. of Past Perfect tenses to all phrases in the srt	Syntactic (temporal)
pas_perf_prog_tra	Num. of Past Perfect Progressive tenses in the srt	Syntactic (temporal)
ratio_pas_perf_prog_tra	Relation of Num. of Past Perfect Progressive tenses to all phrases in the srt	Syntactic (temporal)
will_tra	Num. of Will-Future tenses in the srt	Syntactic (temporal)
ratio_will_tra	Relation of Num. of Will-Future tenses to all phrases in the srt	Syntactic (temporal)
fu_prog_tra	Num. of Future Progressive tenses in the srt	Syntactic (temporal)
ratio_fu_prog_tra	Relation of Num. of Future Progressive tenses to all phrases in the srt	Syntactic (temporal)
fu_perf_tra	Num. of Future Perfect tenses in the srt	Syntactic (temporal)
ratio_fu_perf_tra	Relation of Num. of Future Perfect tenses to all phrases in the srt	Syntactic (temporal)
fu_perf_prog_tra	Num. of Future Perfect Progressive tenses in the srt	Syntactic (temporal)
ratio_fu_perf_prog_tra	Relation of Num. of Future Perfect Progressive tenses to all phrases in the srt	Syntactic (temporal)
cond_sim_tra	Num. of Conditional Simple tenses in the srt	Syntactic (temporal)
ratio_cond_sim_tra	Relation of Num. of Conditional Simple tenses to all phrases in the srt	Syntactic (temporal)
cond_prog_tra	Num. of Conditional Progressive tenses in the srt	Syntactic (temporal)
ratio_cond_prog_tra	Relation of Num. of Conditional Progressive tenses to all phrases in the srt	Syntactic (temporal)
cond_perf_tra	Num. of Conditional Perfect tenses in the srt	Syntactic (temporal)
ratio_cond_perf_tra	Relation of Num. of Conditional Perfect tenses to all phrases in the srt	Syntactic (temporal)
cond_perf_prog_tra	Num. of Conditional Perfect Progressive tenses in the srt	Syntactic (temporal)

Continuation on the next page

Table B.1 – Continuation

Feature	Description	Category
ratio_cond_perf_prog_tra	Relation of Num. of Conditional Perfect Progressive tenses to all phrases in the srt	Syntactic (temporal)
gerund_tra	Num. of Gerund/Present Participle tenses in the srt	Syntactic (temporal)
ratio_gerund_tra	Relation of Num. of Gerund/Present Participle tenses to all phrases in the srt	Syntactic (temporal)
perf_part_tra	Num. of Perfect Participle tenses in the srt	Syntactic (temporal)
ratio_perf_part_tra	Relation of Num. of Perfect Participle tenses to all phrases in the srt	Syntactic (temporal)
inf_tra	Num. of Present Infinitive tenses in the srt	Syntactic (temporal)
ratio_inf_tra	Relation of Num. of Present Infinitive tenses to all phrases in the srt	Syntactic (temporal)
perf_inf_tra	Num. of Perfect Infinitive tenses in the srt	Syntactic (temporal)
ratio_perf_inf_tra	Relation of Num. of Perfect Infinitive tenses to all phrases in the srt	Syntactic (temporal)
active_tra	Num. of active verb forms in the srt	Syntactic (temporal)
ratio_active_tra	Relation of Num. of active verb forms to all phrases in the srt	Syntactic (temporal)
passive_tra	Num. of passive verb forms in the srt	Syntactic (temporal)
ratio_passive_tra	Relation of Num. of passive verb forms to all phrases in the srt	Syntactic (temporal)
ADJP_sli	Num. of adjective phrases in the slides	Syntactic (phrases)
ratio_ADJP_sli	Relation of Num. of adjective phrases to all phrases in the slides	Syntactic (phrases)
avg_ADJP_sli	Average Num. of adjective phrases per line in the slides	Syntactic (phrases)
ADVP_sli	Num. of adverb phrases in the slides	Syntactic (phrases)
ratio_ADVP_sli	Relation of Num. of adverb phrases to all phrases in the slides	Syntactic (phrases)
avg_ADVP_sli	Average Num. of adverb phrases per line in the slides	Syntactic (phrases)
NP_sli	Num. of noun phrases in the slides	Syntactic (phrases)
ratio_NP_sli	Relation of Num. of noun phrases to all phrases in the slides	Syntactic (phrases)
avg_NP_sli	Average Num. of noun phrases per line in the slides	Syntactic (phrases)
PP_sli	Num. of prepositional phrases in the slides	Syntactic (phrases)
ratio_PP_sli	Relation of Num. of prepositional phrases to all phrases in the slides	Syntactic (phrases)
avg_PP_sli	Average Num. of prepositional phrases per line in the slides	Syntactic (phrases)
S_sli	Num. of simple declarative clauses in the slides	Syntactic (phrases)
ratio_S_sli	Relation of Num. of simple declarative clauses to all phrases in the slides	Syntactic (phrases)
avg_S_sli	Average Num. of simple declarative clauses per line in the slides	Syntactic (phrases)

Continuation on the next page

Table B.1 – Continuation

Feature	Description	Category
FRAG_sli	Num. of fragments in the slides	Syntactic (phrases)
ratio_FRAG_sli	Relation of Num. of fragments to all phrases in the slides	Syntactic (phrases)
avg_FRAG_sli	Average Num. of fragments per line in the slides	Syntactic (phrases)
SBAR_sli	Num. of subordinate clauses in the slides	Syntactic (phrases)
ratio_SBAR_sli	Relation of Num. of subordinate clauses to all phrases in the slides	Syntactic (phrases)
avg_SBAR_sli	Average Num. of subordinate clauses per line in the slides	Syntactic (phrases)
SBARQ_sli	Num. of direct questions with wh-element in the slides	Syntactic (phrases)
ratio_SBARQ_sli	Relation of Num. of direct questions with wh-element to all phrases in the slides	Syntactic (phrases)
avg_SBARQ_sli	Average Num. of direct questions with wh-element per line in the slides	Syntactic (phrases)
SINV_sli	Num. of declarative sentence with subject-aux inversion in the slides	Syntactic (phrases)
ratio_SINV_sli	Relation of Num. of declarative sentence with subject-aux inversion to all phrases in the slides	Syntactic (phrases)
avg_SINV_sli	Average Num. of declarative sentence with subject-aux inversion per line in the slides	Syntactic (phrases)
SQ_sli	Num. of yes/no questions and subconstituent of SBARQ excluding wh-element in the slides	Syntactic (phrases)
ratio_SQ_sli	Relation of Num. of yes/no questions and subconstituent of SBARQ excluding wh-element to all phrases in the slides	Syntactic (phrases)
avg_SQ_sli	Average Num. of yes/no questions and subconstituent of SBARQ excluding wh-element per line in the slides	Syntactic (phrases)
VP_sli	Num. of verb phrases in the slides	Syntactic (phrases)
ratio_VP_sli	Relation of Num. of verb phrases to all phrases in the slides	Syntactic (phrases)
avg_VP_sli	Average Num. of verb phrases per line in the slides	Syntactic (phrases)
WHADVP_sli	Num. of wh-adverb phrases in the slides	Syntactic (phrases)
ratio_WHADVP_sli	Relation of Num. of wh-adverb phrases to all phrases in the slides	Syntactic (phrases)
avg_WHADVP_sli	Average Num. of wh-adverb phrases per line in the slides	Syntactic (phrases)
WHNP_sli	Num. of wh-noun phrases in the slides	Syntactic (phrases)
ratio_WHNP_sli	Relation of Num. of wh-noun phrases to all phrases in the slides	Syntactic (phrases)
avg_WHNP_sli	Average Num. of wh-noun phrases per line in the slides	Syntactic (phrases)

Continuation on the next page

Table B.1 – Continuation

Feature	Description	Category
WHPP_sli	Num. of wh-prepositional phrases in the slides	Syntactic (phrases)
ratio_WHPP_sli	Relation of Num. of wh-prepositional phrases to all phrases in the slides	Syntactic (phrases)
avg_WHPP_sli	Average Num. of wh-prepositional phrases per line in the slides	Syntactic (phrases)
avg_phrases_sli	Average Num. of phrases per line in the slides	Syntactic (phrases)
ADJP_tra	Num. of adjective phrases in the srt	Syntactic (phrases)
ratio_ADJP_tra	Relation of Num. of adjective phrases to all phrases in the srt	Syntactic (phrases)
avg_ADJP_tra	Average Num. of adjective phrases per sentence in the srt	Syntactic (phrases)
ADVP_tra	Num. of adverb phrases in the srt	Syntactic (phrases)
ratio_ADVP_tra	Relation of Num. of adverb phrases to all phrases in the srt	Syntactic (phrases)
avg_ADVP_tra	Average Num. of adverb phrases per sentence in the srt	Syntactic (phrases)
NP_tra	Num. of noun phrases in the srt	Syntactic (phrases)
ratio_NP_tra	Relation of Num. of noun phrases to all phrases in the srt	Syntactic (phrases)
avg_NP_tra	Average Num. of noun phrases per sentence in the srt	Syntactic (phrases)
PP_tra	Num. of prepositional phrases in the srt	Syntactic (phrases)
ratio_PP_tra	Relation of Num. of prepositional phrases to all phrases in the srt	Syntactic (phrases)
avg_PP_tra	Average Num. of prepositional phrases per sentence in the srt	Syntactic (phrases)
S_tra	Num. of simple declarative clauses in the srt	Syntactic (phrases)
ratio_S_tra	Relation of Num. of simple declarative clauses to all phrases in the srt	Syntactic (phrases)
avg_S_tra	Average Num. of simple declarative clauses per sentence in the srt	Syntactic (phrases)
FRAG_tra	Num. of fragments in the srt	Syntactic (phrases)
ratio_FRAG_tra	Relation of Num. of fragments to all phrases in the srt	Syntactic (phrases)
avg_FRAG_tra, SBAR_tra	Average Num. of fragments per sentence in the srt	Syntactic (phrases)
SBAR_tra	Num. of subordinate clauses in the srt	Syntactic (phrases)
ratio_SBAR_tra	Relation of Num. of subordinate clauses to all phrases in the srt	Syntactic (phrases)
avg_SBAR_tra	Average Num. of subordinate clauses per sentence in the srt	Syntactic (phrases)
SBARQ_tra	Num. of direct questions with wh-element in the srt	Syntactic (phrases)

Continuation on the next page

Table B.1 – Continuation

Feature	Description	Category
ratio_SBARQ_tra	Relation of Num. of direct questions with wh-element to all phrases in the srt	Syntactic (phrases)
avg_SBARQ_tra	Average Num. of direct questions with wh-element per sentence in the srt	Syntactic (phrases)
SINV_tra	Num. of declarative sentence with subject-aux inversion in the srt	Syntactic (phrases)
ratio_SINV_tra	Relation of Num. of declarative sentence with subject-aux inversion to all phrases in the srt	Syntactic (phrases)
avg_SINV_tra	Average Num. of declarative sentence with subject-aux inversion per sentence in the srt	Syntactic (phrases)
SQ_tra	Num. of yes/no questions and subconstituent of SBARQ excluding wh-element in the srt	Syntactic (phrases)
ratio_SQ_tra	Relation of Num. of yes/no questions and subconstituent of SBARQ excluding wh-element to all phrases in the srt	Syntactic (phrases)
avg_SQ_tra	Average Num. of yes/no questions and subconstituent of SBARQ excluding wh-element per sentence in the srt	Syntactic (phrases)
VP_tra	Num. of verb phrases in the srt	Syntactic (phrases)
ratio_VP_tra	Relation of Num. of verb phrases to all phrases in the srt	Syntactic (phrases)
avg_VP_tra	Average Num. of verb phrases per sentence in the srt	Syntactic (phrases)
WHADVP_tra	Num. of wh-adverb phrases in the srt	Syntactic (phrases)
ratio_WHADVP_tra	Relation of Num. of wh-adverb phrases to all phrases in the srt	Syntactic (phrases)
avg_WHADVP_tra	Average Num. of wh-adverb phrases per sentence in the srt	Syntactic (phrases)
WHNP_tra	Num. of wh-noun phrases in the srt	Syntactic (phrases)
ratio_WHNP_tra	Relation of Num. of wh-noun phrases to all phrases in the srt	Syntactic (phrases)
avg_WHNP_tra	Average Num. of wh-noun phrases per sentence in the srt	Syntactic (phrases)
WHPP_tra	Num. of wh-prepositional phrases in the srt	Syntactic (phrases)
ratio_WHPP_tra	Relation of Num. of wh-prepositional phrases to all phrases in the srt	Syntactic (phrases)
avg_WHPP_tra	Average Num. of wh-prepositional phrases per sentence in the srt	Syntactic (phrases)
avg_phrases_tra	Average Num. of phrases per sentence in the srt	Syntactic (phrases)
avg_trigram_sli	Average Num. of trigrams per line in the slides	Syntactic (other)
avg_tetragram_sli	Average Num. of tetragrams per line in the slides	Syntactic (other)
amount_statement_sli	Num. of statements in the slides	Syntactic (other)

Continuation on the next page

Table B.1 – Continuation

Feature	Description	Category
ratio_statement_sli	Relation of Num. of statements to all lines in the slides	Syntactic (other)
amount_question_sli	Num. of questions in the slides	Syntactic (other)
ratio_question_sli	Relation of Num. of statements to all lines in the slides	Syntactic (other)
amount_tok_sli	Num. of words in the slides	Syntactic (other)
sum_tok_len_sli	Num. of characters in the slides	Syntactic (other)
min_tok_len_sli	Size of the shortest word in the slides	Syntactic (other)
avg_tok_len_sli	Average size of the words in the slides	Syntactic (other)
max_tok_len_sli	Size of the longest word in the slides	Syntactic (other)
avg_trigram_tra	Average Num. of trigrams per sentence in the srt	Syntactic (other)
avg_tetragram_tra	Average Num. of tetragrams per sentence in the srt	Syntactic (other)
amount_statement_tra	Num. of statements in the srt	Syntactic (other)
ratio_statement_tra	Relation of Num. of statements to all sentences in the srt	Syntactic (other)
amount_question_tra	Num. of questions in the srt	Syntactic (other)
ratio_question_tra	Relation of Num. of statements to all sentences in the srt	Syntactic (other)
amount_tok_tra	Num. of words in the srt	Syntactic (other)
sum_tok_len_tra	Num. of characters in the srt	Syntactic (other)
min_tok_len_tra	Size of the shortest word in the srt	Syntactic (other)
avg_tok_len_tra	Average size of the words in the srt	Syntactic (other)
max_tok_len_tra	Size of the longest word in the srt	Syntactic (other)
flesch_ease_sli	Flesch-Reading-Ease result of the slides	Readability
flesch_kin_sli	Flesch-Kincaid result of the slides	Readability
gunning_fog_sli	Gunning-Fog Index result of the slides	Readability
smog_sli	SMOG result of the slides	Readability
ari_sli	Automated Readability Index result of the slides	Readability
coleman_sli	Coleman-Liau Index result of the slides	Readability
read_time_sli	Reading time of the text of the slides	Readability
flesch_ease_tra	Flesch-Reading-Ease result of the srt	Readability
flesch_kin_tra	Flesch-Kincaid result of the srt	Readability
gunning_fog_tra	Gunning-Fog Index result of the srt	Readability
smog_tra	SMOG result of the srt	Readability
ari_tra	Automated Readability Index result of the srt	Readability
coleman_tra	Coleman-Liau Index result of the srt	Readability
read_time_tra	Reading time of the text of the srt	Readability
avg_freq_tok_sli	Average frequency of a word in the slides	Lexical
min_age_sli	Minimum age of acquisition of a word in the slides	Lexical
avg_age_sli	Average age of acquisition of a word in the slides	Lexical

Continuation on the next page

Table B.1 – Continuation

Feature	Description	Category
max_age_sli	Maximum age of acquisition of a word in the slides	Lexical
amount_syl_sli	Num. of syllables in the slides	Lexical
amount_one_syl_sli	Num. of syllables in the slides	Lexical
amount_two_syl_sli	Num. of syllables in the slides	Lexical
amount_psyl_sli	Num. of syllables in the slides	Lexical
amount_hard_sli	Num. of syllables in the slides	Lexical
avg_syl_sli	Average Num. of syllables per word in the slides	Lexical
ratio_one_syl_sli	Relation of Num. of monosyllabic words to all unfiltered tokens in the slides	Lexical
ratio_two_syl_sli	Relation of Num. of two-syllable words to all unfiltered tokens in the slides	Lexical
ratio_psyl_sli	Relation of Num. of polysyllabic words to all unfiltered tokens in the slides	Lexical
ratio_hard_sli	Relation of Num. of hard words to all unfiltered tokens in the slides	Lexical
amount_uni_tok_sli	Num. of unique words in the slides	Lexical
ratio_uni_tok_sli	Relation of Num. of unique words to all words in the slides	Lexical
amount_uni_lemma_sli	Num. of unique lemmas of the words in the slides	Lexical
ratio_uni_lemma_sli	Relation of Num. of unique lemmas of the words to all words in the slides	Lexical
avg_freq_tok_tra	Average frequency of a word in the srt	Lexical
min_age_tra	Minimum age of acquisition of a word in the srt	Lexical
avg_age_tra	Average age of acquisition of a word in the srt	Lexical
max_age_tra	Maximum age of acquisition of a word in the srt	Lexical
amount_syl_tra	Num. of syllables in the srt	Lexical
amount_one_syl_tra	Num. of syllables in the srt	Lexical
amount_two_syl_tra	Num. of syllables in the srt	Lexical
amount_psyl_tra	Num. of syllables in the srt	Lexical
amount_hard_tra	Num. of syllables in the srt	Lexical
avg_syl_tra	Average Num. of syllables per word in the srt	Lexical
ratio_one_syl_tra	Relation of Num. of monosyllabic words to all unfiltered tokens in the srt	Lexical
ratio_two_syl_tra	Relation of Num. of two-syllable words to all unfiltered tokens in the srt	Lexical
ratio_psyl_tra	Relation of Num. of polysyllabic words to all unfiltered tokens in the srt	Lexical
ratio_hard_tra	Relation of Num. of hard words to all unfiltered tokens in the srt	Lexical
amount_uni_tok_tra	Num. of unique words in the srt	Lexical

Continuation on the next page

Table B.1 – Continuation

Feature	Description	Category
ratio_uni_tok_tra	Relation of Num. of unique words to all words in the srt	Lexical
amount_uni_lemma_tra	Num. of unique lemmas of the words in the srt	Lexical
ratio_uni_lemma_tra	Relation of Num. of unique lemmas of the words to all words in the srt	Lexical
sum_lines	Num. of lines in the slides	Structural (slides)
min_line_len	Minimum Num. of words per line in the slides	Structural (slides)
avg_line_len	Average Num. of words per line in the slides	Structural (slides)
max_line_len	Maximum Num. of words per line in the slides	Structural (slides)
min_lines	Minimum Num. of lines per slide	Structural (slides)
avg_lines	Average Num. of lines per slide	Structural (slides)
max_lines	Maximum Num. of lines per slide	Structural (slides)
min_words_slide	Minimum Num. of words per slide	Structural (slides)
avg_words_slide	Average Num. of words per slide	Structural (slides)
max_words_slide	Maximum Num. of words per slide	Structural (slides)
min_line_chars	Minimum Num. of characters per line in the slides	Structural (slides)
avg_line_chars	Average Num. of characters per line in the slides	Structural (slides)
max_line_chars	Maximum Num. of characters per line in the slides	Structural (slides)
amount_slides	Amount of slides of a presentation	Structural (slides)
amount_subtitles	Num. of subtitles in the srt	Structural (srt)
amount_sentences	Num. of sentences in the srt	Structural (srt)
speak_time	Subtitle display time in the srt	Structural (srt)
speak_difference	Difference between display time and reading time (180 WPM)	Structural (srt)
min_sen_len	Minimum Num. of words per sentence in the srt	Structural (srt)
avg_sen_len	Average Num. of words per sentence in the srt	Structural (srt)
max_sen_len	Maximum Num. of words per sentence in the srt	Structural (srt)
min_sen_chars	Minimum Num. of characters per sentence in the srt	Structural (srt)
avg_sen_chars	Average Num. of characters per sentence in the srt	Structural (srt)
max_sen_chars	Maximum Num. of characters per sentence in the srt	Structural (srt)
embed_slide	Average sentence embedding of the slides	Semantic
similarity_sli	Average similarity between the lines in the slides	Semantic
embed_srt	Average sentence embedding of the srt	Semantic

Continuation on the next page

Table B.1 – Continuation

Feature	**Description**	**Category**
similarity_tra	Average similarity between the sentences in the srt	Semantic
diff_similarity	Average similarity between the sentences in the srt	Semantic
similarity_vectors	similarity of the two embeddings	Semantic
Person_ID	The ID of a user	User specific

Own Publications

[70] Ralph Ewerth, Christian Otto, and Eric Müller-Budack. "Computational Approaches for the Interpretation of Image-Text Relations". In: Oct. 2021, pp. 109–138. ISBN: 9783110725001. DOI: 10.1515/9783110725001-005.

[117] Johannes von Hoyer, Anett Hoppe, Yvonne Kammerer, Christian Otto, Georg Pardi, Markus Rokicki, Ran Yu, Stefan Dietze, Ralph Ewerth, and Peter Holtz. "The Search as Learning Spaceship: Toward a Comprehensive Model of Psychological and Technological Facets of Search as Learning". In: *Frontiers in Psychology* 13 (Mar. 2022). DOI: 10.3389/fpsyg.2022.827748.

[177] Justyna Medrek, Christian Otto, and Ralph Ewerth. "Recommending Scientific Videos Based on Metadata Enrichment Using Linked Open Data". In: *Digital Libraries for Open Knowledge, 22nd International Conference on Theory and Practice of Digital Libraries, TPDL 2018, Porto, Portugal, September 10-13, 2018, Proceedings*. Vol. 11057. Lecture Notes in Computer Science. Springer, 2018, pp. 286–292. DOI: 10.1007/978-3-030-00066-0_25.

[188] Markus Mühling, Nikolaus Korfhage, Eric Müller, Christian Otto, Matthias Springstein, Thomas Langelage, Uli Veith, Ralph Ewerth, and Bernd Freisleben. "Deep learning for content-based video retrieval in film and television production". In: vol. 76. 21. 2017, pp. 22169–22194. DOI: 10.1007/s11042-017-4962-9.

[189] Eric Müller, Christian Otto, and Ralph Ewerth. "Semi-supervised Identification of Rarely Appearing Persons in Video by Correcting Weak Labels". In: *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, ICMR 2016, New York, New York, USA, June 6-9, 2016*. ACM, 2016, pp. 381–384. DOI: 10.1145/2911996.2912073.

[197] Christian Otto, Sebastian Holzki, and Ralph Ewerth. "Is This an Example Image? - Predicting the Relative Abstractness Level of Image and Text". In: *Advances in Information Retrieval - 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14-18, 2019, Proceedings, Part I*. Vol. 11437. Lecture Notes in Computer Science. Springer, 2019, pp. 711–725. DOI: 10.1007/978-3-030-15712-8_46.

[198] Christian Otto, Markus Rokicki, Georg Pardi, Wolfgang Gritz, Daniel Hienert, Ran Yu, Johannes von Hoyer, Anett Hoppe, Stefan Dietze, Peter Holtz, Yvonne Kammerer, and Ralph Ewerth. "SaL-Lightning Dataset: Search and Eye Gaze Behavior, Resource Interactions and Knowledge Gain during Web Search". In: *CHIIR '22: ACM SIGIR Conference on Human Information Interaction and Retrieval, Regensburg, Germany, March 14 - 18, 2022*. Ed. by David Elsweiler. ACM, 2022, pp. 347–352. DOI: 10.1145/3498366.3505835.

[199] Christian Otto, Matthias Springstein, Avishek Anand, and Ralph Ewerth. "Characterization and classification of semantic image-text relations". In: *International Journal of Multimedia Information Retrieval* 9.1 (2020), pp. 31–45. DOI: 10.1007/s13735-019-00187-6.

[200] Christian Otto, Matthias Springstein, Avishek Anand, and Ralph Ewerth. "Understanding, Categorizing and Predicting Semantic Image-Text Relations". In: *Proceedings of the 2019 on International Conference on Multimedia Retrieval, ICMR 2019, Ottawa, ON, Canada, June 10-13, 2019*. Ed. by Abdulmotaleb El-Saddik, Alberto Del Bimbo, Zhongfei Zhang, Alexander G. Hauptmann, K. Selçuk Candan, Marco Bertini, Lexing Xie, and Xiao-Yong Wei. ACM, 2019, pp. 168–176. DOI: 10.1145/3323873.3325049.

[201] Christian Otto, Markos Stamatakis, Anett Hoppe, and Ralph Ewerth. "Predicting Knowledge Gain for MOOC Video Consumption". In: *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners' and Doctoral Consortium - 23rd International Conference, AIED 2022, Durham, UK, July 27-31, 2022, Proceedings, Part II*. Ed. by Maria Mercedes T. Rodrigo, Noburu Matsuda, Alexandra I. Cristea, and Vania Dimitrova. Vol. 13356. Lecture Notes in Computer Science. Springer, 2022, pp. 458–462. DOI: 10.1007/978-3-031-11647-6_92.

[202] Christian Otto, Ran Yu, Georg Pardi, Johannes von Hoyer, Markus Rokicki, Anett Hoppe, Peter Holtz, Yvonne Kammerer, Stefan Dietze, and Ralph Ewerth. "Predicting Knowledge Gain During Web Search Based on Multimedia Resource Consumption". In: *Artificial Intelligence in Education - 22nd International Conference, AIED 2021, Utrecht, The Netherlands, June 14-18, 2021, Proceedings, Part I*. Vol. 12748. Lecture Notes in Computer Science. Springer, 2021, pp. 318–330. DOI: 10.1007/978-3-030-78292-4_26.

[240] Jianwei Shi, Christian Otto, Anett Hoppe, Peter Holtz, and Ralph Ewerth. "Investigating Correlations of Automatically Extracted Multimodal Features and Lecture Video Quality". In: *Proceedings of the 1st International Workshop on Search as Learning with Multimedia Information*. SALMM '19. Nice, France: Association for Computing Machinery, 2019, pp. 11–19. ISBN: 9781450369190. DOI: 10.1145/3347451.3356731.

[304] Hang Zhou, Christian Otto, and Ralph Ewerth. "Visual Summarization of Scholarly Videos Using Word Embeddings and Keyphrase Extraction". In: *Digital Libraries for Open Knowledge - 23rd International Conference on Theory and Practice of Digital Libraries, TPDL 2019, Oslo, Norway, September 9-12, 2019, Proceedings*. Vol. 11799. Lecture Notes in Computer Science. Springer, 2019, pp. 327–335. DOI: 10.1007/978-3-030-30760-8_28.

Other Publications

[1] Waleed Abdulla. *Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow*. https://github.com/matterport/Mask_RCNN. 2017.

[2] *Age-of-acquisition (AoA) norms for over 50 thousand English words*. http://crr.ugent.be/archives/806. Accessed: 16 November 2021.

[3] Malihe Alikhani, Piyush Sharma, Shengjie Li, Radu Soricut, and Matthew Stone. "Clue: Cross-modal Coherence Modeling for Caption Generation". In: *CoRR* abs/2005.00908 (2020). arXiv: 2005.00908. URL: https://arxiv.org/abs/2005.00908.

[4] Hakan Altinpulluk, Hakan Kilinc, Mehmet Firat, and Onur Yumurtaci. "The influence of segmented and complete educational videos on the cognitive load, satisfaction, engagement, and academic achievement levels of learners". In: *Journal of Computers in Education* 7.2 (2020), pp. 155–182. ISSN: 2197-9995.

[5] F. Amadieu, J. Lemarié, and A. Tricot. "How may multimedia and hypertext documents support deep processing for learning?" In: *Psychologie Française* 62.3 (2017). Cognition et multimédia : les atouts du numérique en situation d'apprentissage, pp. 209–221. ISSN: 0033-2984.

[6] Lorin W Anderson, David R Krathwohl, P Airasian, K Cruikshank, R Mayer, P Pintrich, J Raths, and M Wittrock. "A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy". In: *New York. Longman Publishing. Artz, AF, & Armour-Thomas, E.(1992). Development of a cognitive-metacognitive framework for protocol analysis of mathematical problem solving in small groups. Cognition and Instruction* 9.2 (2001), pp. 137–175.

[7] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. "Bottom-up and top-down attention for image captioning and visual question answering". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 6077–6086.

[8] Evlampios E. Apostolidis, Eleni Adamantidou, Alexandros I. Metsai, Vasileios, and Ioannis Patras. "Video Summarization Using Deep Neural Networks: A Survey". In: *Proc. IEEE* 109.11 (2021), pp. 1838–1863. DOI: 10.1109/JPROC.2021.3117472.

[9] Jana Arndt, Anne Schüler, and Katharina Scheiter. "Investigating the Influence of Simultaneous– Versus Sequential–Text-Picture Presentation on Text-Picture Integration". In: *The Journal of Experimental Education* 87.1 (2019), pp. 116–127.

[10] James Badger. "Learning in non-formal settings: Investigating cemetery guides' talk during school visits". In: *International Journal of Educational Research* 109 (2021), p. 101852. ISSN: 0883-0355. DOI: https://doi.org/10.1016/j.ijer.2021.101852. URL: https://www.sciencedirect.com/science/article/pii/S088303552100121X.

[11] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. 2015.

[12] Saeid Balaneshin-kordan and Alexander Kotov. "Deep Neural Architecture for Multi-Modal Retrieval based on Joint Embedding Space for Text and Images". In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM. 2018, pp. 28–36.

[13] David A Balota, Melvin J Yap, Keith A Hutchison, Michael J Cortese, Brett Kessler, Bjorn Loftis, James H Neely, Douglas L Nelson, Greg B Simpson, and Rebecca Treiman. "The English lexicon project". In: *Behavior research methods* 39.3 (2007), pp. 445–459.

[14] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. "Multimodal Machine Learning: A Survey and Taxonomy". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 41.2 (2019), pp. 423–443.

[15] Roland Barthes. "Image-Music-Text, ed. and trans". In: *S. Heath, London: Fontana* 332 (1977).

[16] John Bateman. *Text and image: A critical introduction to the visual/verbal divide*. Routledge, 2014.

[17] Kevin S. Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. "When Is "Nearest Neighbor" Meaningful?" In: *Database Theory - ICDT '99, 7th Int. Conference, Jerusalem, Israel, January 10-12, 1999, Proceedings*. Ed. by Catriel Beeri and Peter Buneman. Vol. 1540. Lecture Notes in Computer Science. Springer, 1999, pp. 217–235. DOI: 10.1007/3-540-49257-7_15.

[18] Nilavra Bhattacharya and Jacek Gwizdka. "Measuring Learning During Search: Differences in Interactions, Eye-Gaze, and Semantic Similarity to Expert Knowledge". In: *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. ACM. 2019, pp. 63–71.

[19] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python. Analyzing Text with the Natural Language Toolkit*. Sebastopol, USA: O'Reilly Media Inc., 2009. ISBN: 978-0-596-51649-9.

[20] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation". In: *J. Mach. Learn. Res.* 3 (2003), pp. 993–1022. URL: http://jmlr.org/papers/v3/blei03a.html.

[21] Paul Boersma and David Weenink. *Praat: doing phonetics by computer [Computer program]*. http://www.fon.hum.uva.nl/praat. [Version 6.0.37; retrieved 15-January-2019]. 2018.

[22] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomás Mikolov. "Enriching Word Vectors with Subword Information". In: *Trans. Assoc. Comput. Linguistics* 5 (2017), pp. 135–146.

[23] Florian Boudin. "pke: an open source python-based keyphrase extraction toolkit". In: *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference System Demonstrations, December 11-16, 2016, Osaka, Japan*. Ed. by Hideo Watanabe. ACL, 2016, pp. 69–73.

[24] Florian Boudin. "Unsupervised Keyphrase Extraction with Multipartite Graphs". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*. Ed. by Marilyn A. Walker, Heng Ji, and Amanda Stent. Association for Computational Linguistics, 2018, pp. 667–672. DOI: 10.18653/v1/n18-2105.

[25] Adrien Bougouin, Florian Boudin, and Béatrice Daille. "TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction". In: *Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan, October 14-18, 2013*. Asian Federation of Natural Language Processing / ACL, 2013, pp. 543–551. URL: https://aclanthology.org/I13-1062/.

[26] *BPG image format (2014)*. http://bellard.org/bpg/. Accessed: 14 May 2022.

[27] Cynthia Brame. "Effective Educational Videos: Principles and Guidelines for Maximizing Student Learning from Video Content". In: vol. 15. Oct. 2016, es6–es6. DOI: 10.1187/cbe.16-03-0125.

[28] Amy Brand, Liz Allen, Micah Altman, Marjorie M. K. Hlava, and Jo Scott. "Beyond authorship: attribution, contribution, collaboration, and credit". In: *Learn. Publ.* 28.2 (2015), pp. 151–155. DOI: 10.1087/20150211.

[29] Ivar Bråten, Christian Brandmo, and Yvonne Kammerer. "A Validation Study of the Internet-Specific Epistemic Justification Inventory With Norwegian Preservice Teachers". In: *Journal of Educational Computing Research* 57.4 (2019), pp. 877–900. DOI: 10.1177/0735633118769438. eprint: https://doi.org/10.1177/0735633118769438. URL: https://doi.org/10.1177/0735633118769438.

[30] Leo Breiman. *Classification and regression trees*. Routledge, 2017.

[31] Leo Breiman. "Statistical modeling: The two cultures (with comments and a rejoinder by the author)". In: *Statistical science* 16.3 (2001), pp. 199–231.

[32] Francesco Brigo, Willem M. Otte, Stanley C. Igwe, Frediano Tezzon, and Raffaele Nardone. "Clearly written, easily comprehended? The readability of websites providing information on epilepsy". In: *Epilepsy & Behavior* 44 (2015), pp. 35–39. ISSN: 1525-5050. DOI: https://doi.org/10.1016/j.yebeh.2014.12.029. URL: https://www.sciencedirect.com/science/article/pii/S1525505014006994.

[33] Julie L Brockman and John M Dirkx. "Learning to become a machine operator: The dialogical relationship between context, self, and content". In: *Human Resource Development Quarterly* 17.2 (2006), pp. 199–221.

[34] Andrei Z. Broder. "A taxonomy of web search". In: *SIGIR Forum* 36.2 (2002), pp. 3–10. DOI: 10.1145/792550.792552.

[35] Carl Brucker. "Arkansas tech writing". In: *English* 2053.June (2009), p. 109.

[36] Serhat S Bucak, Rong Jin, and Anil K Jain. "Multiple kernel learning for visual object recognition: A review". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.7 (2014), pp. 1354–1369.

[37] Sahan Bulathwela, María Pérez-Ortiz, Aldo Lipani, Emine Yilmaz, and John Shawe-Taylor. "Predicting Engagement in Video Lectures". In: *Proceedings of the 13th International Conference on Educational Data Mining, EDM 2020, Fully virtual conference, July 10-13, 2020*. Ed. by Anna N. Rafferty, Jacob Whitehill, Cristóbal Romero, and Violetta Cavalli-Sforza. International Educational Data Mining Society, 2020. URL: https://educationaldatamining.org/files/conferences/EDM2020/papers/paper_62.pdf.

[38] Wen-Hsuan Chang, Jie-Chi Yang, and Yu-Chieh Wu. "A Keyword-based Video Summarization Learning Platform with Multimodal Surrogates". In: *ICALT 2011, 11th IEEE International Conference on Advanced Learning Technologies, Athens, Georgia, USA, 6-8 July 2011*. IEEE Computer Society, 2011, pp. 37–41. DOI: 10.1109/ICALT.2011.19.

[39] Wen-Hsuan Chang, Jie-Chi Yang, and Yu-Chieh Wu. "A Keyword-based Video Summarization Learning Platform with Multimodal Surrogates". In: *ICALT 2011, 11th IEEE International Conference on Advanced Learning Technologies, Athens, Georgia, USA, 6-8 July 2011*. IEEE Computer Society, 2011, pp. 37–41. DOI: 10.1109/ICALT.2011.19.

[40] Lei Chen, Chee Wee Leong, Gary Feng, and Chong Min Lee. "Using Multimodal Cues to Analyze MLA'14 Oral Presentation Quality Corpus: Presentation Delivery and Slides Quality". In: *Proceedings of the 2014 ACM Workshop on Multimodal Learning Analytics Workshop and Grand Challenge*. MLA. Istanbul, Turkey: ACM, 2014, pp. 45–52. ISBN: 978-1-4503-0488-7. DOI: 10.1145/2666633.2666640. URL: http://doi.acm.org/10.1145/2666633.2666640.

[41] Kyunghyun Cho. "Simple Sparsification Improves Sparse Denoising Autoencoders in Denoising Highly Corrupted Images". In: *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*. Vol. 28. JMLR Workshop and Conference Proceedings. JMLR.org, 2013, pp. 432–440. URL: http://proceedings.mlr.press/v28/cho13.html.

[42] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. "On the Properties of Neural Machine Translation: Encoder-Decoder Approaches". In: *Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar, 25 October 2014*. Ed. by Dekai Wu, Marine Carpuat, Xavier Carreras, and Eva Maria Vecchi. Association for Computational Linguistics, 2014, pp. 103–111. DOI: 10.3115/v1/W14-4012. URL: https://aclanthology.org/W14-4012/.

[43] Yoojin Choi, Mostafa El-Khamy, and Jungwon Lee. "Variable Rate Deep Image Compression With a Conditional Autoencoder". In: *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 3146–3154. DOI: 10.1109/ICCV.2019.00324.

[44] J Clement. "Hours of video uploaded to YouTube every minute". In: *Statista. com* (2019). URL: https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/.

[45] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. "Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)". In: *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2016. URL: http://arxiv.org/abs/1511.07289.

[46] Romain Cohendet, Claire-Héléne Demarty, Ngoc Q. K. Duong, and Martin Engilberge. "VideoMem: Constructing, Analyzing, Predicting Short-Term and Long-Term Video Memorability". In: *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 2531–2540. DOI: 10.1109/ICCV.2019.00262.

[47] Michael J Cole, Jacek Gwizdka, Chang Liu, Nicholas J Belkin, and Xiangmin Zhang. "Inferring user knowledge level from eye movement patterns". In: *Information Processing & Management* 49.5 (2013), pp. 1075–1091.

[48] Meri Coleman and Ta Lin Liau. "A computer readability formula designed for machine scoring." In: *Journal of Applied Psychology* 60.2 (1975), p. 283.

[49] Kevyn Collins-Thompson, Soo Young Rieh, Carl C. Haynes, and Rohail Syed. "Assessing Learning Outcomes in Web Search: A Comparison of Tasks and Query Strategies". In: *Proceedings of the 2016 ACM Conference on Human Information Interaction and Retrieval, CHIIR 2016, Carrboro, North Carolina, USA, March 13-17, 2016*. Ed. by Diane Kelly, Robert Capra, Nicholas J. Belkin, Jaime Teevan, and Pertti Vakkari. ACM, 2016, pp. 163–172. DOI: 10.1145/2854946.2854972.

[50] Xin Cong, Jiawei Sheng, Shiyao Cui, Bowen Yu, Tingwen Liu, and Bin Wang. "Relation-Guided Few-Shot Relational Triple Extraction". In: *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*. Ed. by Enrique Amigó, Pablo Castells,

Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai. ACM, 2022, pp. 2206–2213. DOI: 10.1145/3477495.3531831.

[51] Andrew R. A. Conway, Michael J. Kane, Michael F. Bunting, D. Zach Hambrick, Oliver Wilhelm, and Randall W. Engle. "Working memory span tasks: A methodological review and user's guide". In: *Psychonomic Bulletin & Review* 12.5 (2005), pp. 769–786. ISSN: 1531-5320. URL: https://doi.org/10.3758/BF03196772.

[52] Maria Cotofan. *Work and Well-being during COVID-19: Impact, Inequalities, Resilience, and the Future of Work*. https://worldhappiness.report/ed/2021/work-and-well-being-during-covid-19-impact-inequalities-resilience-and-the-future-of-work/#accommodation-food-service-and-temporary-workers-have-been-hit-hardest. Accessed: 1 November 2021.

[53] *Coursera*. https://www.coursera.org/. Accessed: 2 November 2021.

[54] Paul Covington, Jay Adams, and Emre Sargin. "Deep Neural Networks for YouTube Recommendations". In: *Proceedings of the 10th ACM Conference on Recommender Systems, Boston, MA, USA, September 15-19, 2016*. Ed. by Shilad Sen, Werner Geyer, Jill Freyne, and Pablo Castells. ACM, 2016, pp. 191–198. DOI: 10.1145/2959100.2959190.

[55] Creative Commons. *Attribution-NonCommercial-ShareAlike 4.0 International*. https://creativecommons.org/licenses/by-nc-sa/4.0. [Online; accessed 18-April-2019]. 2019.

[56] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, and Dasarathi Sampath. "The YouTube video recommendation system". In: *Proceedings of the 2010 ACM Conference on Recommender Systems, RecSys 2010, Barcelona, Spain, September 26-30, 2010*. Ed. by Xavier Amatriain, Marc Torrens, Paul Resnick, and Markus Zanker. ACM, 2010, pp. 293–296. DOI: 10.1145/1864708.1864770. URL: https://doi.org/10.1145/1864708.1864770.

[57] *DBpedia*. http://wiki.dbpedia.org. Accessed: 9 November 2021.

[58] Erhan Delen, Jeffrey Liew, and Victor Willson. "Effects of interactivity and instructional scaffolding on learning: Self-regulation in online video-based environments". In: *Computers & Education* 78 (2014), pp. 312–320. ISSN: 0360-1315.

[59] Diana DeStefano and Jo-Anne LeFevre. "Cognitive load in hypertext reading: A review". In: *Computers in Human Behavior* 23.3 (2007), pp. 1616–1641.

[60] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*. 2019, pp. 4171–4186.

[61] *Education Gap Grows Between Rich and Poor, Studies Say*. https://cepa.stanford.edu/news/education-gap-grows-between-rich-and-poor-studies-say. Accessed: 6 March 2023.

[62] *edX*. https://www.edx.org/. Accessed: 16 November 2021.

[63] Carsten Eickhoff, Jaime Teevan, Ryen White, and Susan T. Dumais. "Lessons from the journey: a query log analysis of within-session learning". In: *Seventh ACM International Conference on Web Search and Data Mining, WSDM 2014, New York, NY, USA, February 24-28, 2014*. Ed. by Ben Carterette, Fernando Diaz, Carlos Castillo, and Donald Metzler. ACM, 2014, pp. 223–232. DOI: 10.1145/2556195.2556217.

[64] Ehsan Elhamifar and M. Clara De Paolis Kaluza. "Online Summarization via Submodular and Convex Optimization". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 1818–1826. DOI: 10.1109/CVPR.2017.197.

[65] Stefan Engeser. "Messung des expliziten Leistungsmotivs: Kurzform der Achievement Motives Scale". 2005. URL: https://www.uni-trier.de/fileadmin/fb1/prof/PSY/PGA/bilder/Engeser__2005__Kurzform_der_AMS.pdf.

[66] *Englisch Hilfen*. https://github.com/GokulVSD/FOGIndex. Accessed: 15 November 2021.

[67] *English Lexicon Project Web Site*. https://elexicon.wustl.edu/. Accessed: 16 November 2021.

[68] Gonenc Ercan and Ilyas Cicekli. "Using lexical chains for keyword extraction". In: *Inf. Process. Manag.* 43.6 (2007), pp. 1705–1714. DOI: 10.1016/j.ipm.2007.01.015.

[69] Ralph Ewerth, Stefan Dietze, Anett Hoppe, and Ran Yu. "SALMM'19: First International Workshop on Search as Learning with Multimedia Information". In: *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*. Ed. by Laurent Amsaleg, Benoit Huet, Martha A. Larson, Guillaume Gravier, Hayley Hung, Chong-Wah Ngo, and Wei Tsang Ooi. ACM, 2019, pp. 2724–2725.

[71] Ralph Ewerth, Matthias Springstein, Lo An Phan-Vogtmann, and Juliane Schütze. ""Are Machines Better Than Humans in Image Tagging?" - A User Study Adds to the Puzzle". In: *Advances in Information Retrieval - 39th European Conference on IR Research, ECIR 2017, Aberdeen, UK, April 8-13, 2017, Proceedings*. Ed. by Joemon M. Jose, Claudia Hauff, Ismail Sengör Altingövde, Dawei Song, Dyaa Albakour, Stuart N. K. Watt, and John Tait. Vol. 10193. Lecture Notes in Computer Science. 2017, pp. 186–198. DOI: 10.1007/978-3-319-56608-5_15.

[72] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. "Recent Developments in openSMILE, the Munich Open-source Multimedia Feature Extractor". In: *Proceedings of the 21st ACM International Conference on Multimedia*. MM '13. New York, NY, USA: ACM, 2013, pp. 835–838. ISBN: 978-1-4503-2404-5. DOI: 10.1145/2502081.2502224. URL: http://doi.acm.org/10.1145/2502081.2502224.

[73] Mengdi Fan, Wenmin Wang, Peilei Dong, Liang Han, Ronggang Wang, and Ge Li. "Cross-media Retrieval by Learning Rich Semantic Embeddings of Multimedia". In: *Proceedings of the 2017 ACM Conference on Multimedia, MM 2017, Mountain View, CA, USA, October 23-27, 2017*. 2017, pp. 1698–1706.

[74] *fastText*. https://fasttext.cc/docs/en/crawl-vectors.html. Accessed: 2 November 2021.

[75] Andreas Fischer, Volkmar Frinken, Alicia Fornés, and Horst Bunke. "Transcription alignment of Latin manuscripts using hidden Markov models". In: *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing, HIP@ICDAR 2011, Beijing, China, September 16-17, 2011*. Ed. by Bill Barrett, Michael S. Brown, R. Manmatha, and Jake Gehring. ACM, 2011, pp. 29–36. DOI: 10.1145/2037342.2037348.

[76] Corina Florescu and Cornelia Caragea. "PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*. Ed. by Regina Barzilay and Min-Yen Kan. Association for Computational Linguistics, 2017, pp. 1105–1115. DOI: 10.18653/v1/P17-1102.

[77] Shane Frederick. "Cognitive Reflection and Decision Making". In: *Journal of Economic Perspectives* 19.4 (Dec. 2005), pp. 25–42. DOI: 10.1257/089533005775196732.

[78] Ujwal Gadiraju, Ran Yu, Stefan Dietze, and Peter Holtz. "Analyzing Knowledge Gain of Users in Informational Search Sessions on the Web". In: *2018 ACM on Conference on Human Information Interaction and Retrieval (CHIIR)*. ACM. 2018.

[79] Ujwal Gadiraju, Ran Yu, Stefan Dietze, and Peter Holtz. "Analyzing Knowledge Gain of Users in Informational Search Sessions on the Web". In: *Proceedings of the 2018 Conference on Human Information Interaction and Retrieval, CHIIR 2018, New Brunswick, NJ, USA, March 11-15, 2018*. Ed. by Chirag Shah, Nicholas J. Belkin, Katriina Byström, Jeff Huang, and Falk Scholer. ACM, 2018, pp. 2–11. DOI: 10.1145/3176349.3176381.

[80] Erlijn van Genuchten, Katharina Scheiter, and Anne Schüler. "Examining learning from text and pictures for different task types: Does the multimedia effect differ for conceptual, causal, and procedural tasks?" In: *Computers in Human Behavior* 28.6 (2012), pp. 2209–2218. DOI: 10.1016/j.chb.2012.06.028.

[81] Mehmet Gönen and Ethem Alpaydin. "Multiple Kernel Learning Algorithms". In: *Journal of Machine Learning Research* 12 (2011), pp. 2211–2268.

[82] Jane Greenberg. "Metadata and the world wide web". In: *Encyclopedia of library and information science* 3 (2003), pp. 1876–1888.

[83] Raphaël Grolimund. *Individual work of Bachelor students monitored using a dynamic assessment dashboard*. Zenodo, Feb. 2017. DOI: 10.5281/zenodo.290129. URL: https://doi.org/10.5281/zenodo.290129.

[84] Philip J. Guo, Juho Kim, and Rob Rubin. "How video production affects student engagement: an empirical study of MOOC videos". In: *First (2014) ACM Conference on Learning @ Scale, L@S 2014, Atlanta, GA, USA, March 4-5, 2014*. 2014, pp. 41–50. DOI: 10.1145/2556325.2566239.

[85] Jacek Gwizdka and Xueshu Chen. "Towards Observable Indicators of Learning on Search." In: *SAL@ SIGIR*. 2016.

[86] Michael Gygli, Helmut Grabner, and Luc Van Gool. "Video summarization by learning submodular mixtures of objectives". In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 3090–3098. DOI: 10.1109/CVPR.2015.7298928.

[87] AmirHossein Habibian, Ties van Rozendaal, Jakub M. Tomczak, and Taco Cohen. "Video Compression With Rate-Distortion Autoencoders". In: *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 7032–7041. DOI: 10.1109/ICCV.2019.00713.

[88] Mounia Haddoud and Saïd Abdeddaïm. "Accurate keyphrase extraction by discriminating overlapping phrases". In: *J. Inf. Sci.* 40.4 (2014), pp. 488–500. DOI: 10.1177/0165551514530210.

[89] Matthias Hagen, Martin Potthast, Benno Stein, Christof Bräutigam, and Anna Beyer. *Webis Query Segmentation Corpus 2010 (Webis- QSeC-10)*. Zenodo, July 2010. DOI: 10.5281/zenodo.3256198. URL: https://doi.org/10.5281/zenodo.3256198.

[90] Matthias Hagen, Martin Potthast, Benno Stein, Marcel Gohsen, and Anja Rathgeber. *Webis Query Spelling Corpus 2017 (Webis-QSpell-17)*. Version 2 incl. error annotations. Zenodo, Aug. 2017. DOI: 10.5281/zenodo.3570912. URL: https://doi.org/10.5281/zenodo.3570912.

[91] Matthias Hagen, Martin Potthast, Michael Völske, Jakob Gomoll, and Benno Stein. "How Writers Search: Analyzing the Search and Writing Logs of Non-fictional Essays". In: *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*. ACM. 2016, pp. 193–202.

[92] Matthias Hagen, Benno Stein, Jakob Gomoll, and Anna Beyer. *Webis Search Mission Corpus 2012 (Webis-SMC-12)*. Zenodo, May 2013. DOI: 10.5281/zenodo.3265962. URL: https://doi.org/10.5281/zenodo.3265962.

[93] Fasih Haider, Loredana Cerrato, Nick Campbell, and Saturnino Luz. "Presentation Quality Assessment Using Acoustic Information and Hand Movements". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing*. Shanghai, China: IEEE, Mar. 2016, pp. 2812–2816. DOI: 10.1109/ICASSP.2016.7472190.

[94] Michael Alexander Kirkwood Halliday and CMIM Matthiessen. "An introduction to functional grammar third edition". In: *London: Edward Arnold* (2004).

[95] Zellig S Harris. "Distributional structure". In: *Word* 10.2-3 (1954), pp. 146–162.

[96] Kazi Saidul Hasan and Vincent Ng. "Automatic Keyphrase Extraction: A Survey of the State of the Art". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*. The Association for Computer Linguistics, 2014, pp. 1262–1273. DOI: 10.3115/v1/p14-1119.

[97] Beatrice Susanne Hasler, Bernd Kersten, and John Sweller. "Learner control, cognitive load and instructional animation". In: *Applied Cognitive Psychology* 21.6 (2007), pp. 713–729.

[98] Susan Havre, Elizabeth G. Hetzler, Paul Whitney, and Lucy T. Nowell. "ThemeRiver: Visualizing Thematic Changes in Large Document Collections". In: *IEEE Trans. Vis. Comput. Graph.* 8.1 (2002), pp. 9–20. DOI: 10.1109/2945.981848.

[99] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. 2016, pp. 770–778.

[100] Xiangnan He, Zhankui He, Jingkuan Song, Zhenguang Liu, Yu-Gang Jiang, and Tat-Seng Chua. "NAIS: Neural Attentive Item Similarity Model for Recommendation". In: *IEEE Trans. Knowl. Data Eng.* 30.12 (2018), pp. 2354–2366. DOI: 10.1109/TKDE.2018.2831682. URL: https://doi.org/10.1109/TKDE.2018.2831682.

[101] Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. "Combining lexical and grammatical features to improve readability measures for first and second language texts". In: *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*. 2007, pp. 460–467.

[102] Andreas M Hein. "Identification and bridging of semantic gaps in the context of multi-domain engineering". In: *Forum on Philosophy, Engineering & Technology*. 2010, pp. 58–57.

[103] Christian Andreas Henning and Ralph Ewerth. "Estimating the Information Gap between Textual and Visual Representations". In: *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, ICMR 2017, Bucharest, Romania, June 6-9, 2017*. 2017, pp. 14–22.

[104] Christian Andreas Henning and Ralph Ewerth. "Estimating the information gap between textual and visual representations". In: *IJMIR* 7.1 (2018), pp. 43–56.

[105] Hynek Hermansky, Nathaniel Morgan, A Bayya, and P Kohn. "RASTA-PLP speech analysis technique". In: *ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 1. San Francisco, CA, USA: IEEE, Apr. 1992, pp. 121–124. ISBN: 0-7803-0532-9. DOI: 10.1109/ICASSP.1992.225957.

[106] Daniel Hienert, Matthew Mitsui, Philipp Mayr, Chirag Shah, and Nicholas J. Belkin. "The Role of the Task Topic in Web Search of Different Task Types". In: *Proceedings of the 2018 Conference on Human Information Interaction and Retrieval, CHIIR 2018, New Brunswick, NJ, USA, March 11-15, 2018*. Ed. by Chirag Shah, Nicholas J. Belkin, Katriina Byström, Jeff Huang, and Falk Scholer. ACM, 2018, pp. 72–81. DOI: 10.1145/3176349.3176382.

[107] Rebecca Hincks. "Measures and perceptions of liveliness in student oral presentation speech: A proposal for an automatic feedback mechanism". In: *System* 33.4 (2005), pp. 575–591. ISSN: 0346-251X. DOI: https://doi.org/10.1016/j.system.2005.04.002. URL: http://www.sciencedirect.com/science/article/pii/S0346251X05000679.

[108] Avery Hlousek and Bentley Krause. *The impact of learning modality on team-based learning (TBL) outcomes in anatomical sciences education*. Zenodo, July 2020. DOI: 10.5061/dryad.cfxpnvx35. URL: https://doi.org/10.5061/dryad.cfxpnvx35.

[109] Jerry R Hobbs. *Why is discourse coherent*. Tech. rep. SRI INTERNATIONAL MENLO PARK CA, 1978.

[110] Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, Jürgen Schmidhuber, et al. *Gradient flow in recurrent nets: the difficulty of learning long-term dependencies*. 2001.

[111] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-Term Memory". In: *Neural Comput.* 9.8 (1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735. URL: https://doi.org/10.1162/neco.1997.9.8.1735.

[112] Richang Hong, Jinhui Tang, Hung-Khoon Tan, Chong-Wah Ngo, Shuicheng Yan, and Tat-Seng Chua. "Beyond search: Event-driven summarization for web videos". In: *ACM Trans. Multim. Comput. Commun. Appl.* 7.4 (2011), 35:1–35:18. DOI: 10.1145/2043612.2043613.

[113] Anett Hoppe, Peter Holtz, Yvonne Kammerer, Ran Yu, Stefan Dietze, and Ralph Ewerth. "Current Challenges for Studying Search as Learning Processes". In: (Linked Learning Workshop – Learning and Education with Web Data (LILE), in conjunction with ACM Conference on Web Science). 2018.

[114] Anett Hoppe, Peter Holtz, Yvonne Kammerer, Ran Yu, Stefan Dietze, and Ralph Ewerth. "Current Challenges for Studying Search as Learning Processes". In: *Linked Learning Workshop - Learning and Education with Web Data (LILE), in conjunction with ACM Conference on Web Science* (Linked Learning Workshop – Learning and Education with Web Data (LILE), in conjunction with ACM Conference on Web Science). 2018.

[115] Benjamin D Horne and Sibel Adali. "This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news". In: *arXiv preprint arXiv:1703.09398* (2017).

[116] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications". In: *CoRR* abs/1704.04861 (2017).

[118] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. "Densely Connected Convolutional Networks". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. 2017, pp. 2261–2269.

[119] Ting-Hao (Kenneth) Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross B. Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. "Visual Storytelling". In: *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*. 2016, pp. 1233–1239.

[120] Anette Hulth. "Reducing false positives by expert combination in automatic keyword indexing". In: *Recent Advances in Natural Language Processing III, Selected Papers from RANLP 2003, Borovets, Bulgaria*. Ed. by Nicolas Nicolov, Kalina Bontcheva, Galia Angelova, and Ruslan Mitkov. Vol. 260. Current Issues in Linguistic Theory (CILT). John Benjamins, Amsterdam/Philadelphia, 2003, pp. 367–376.

[121] Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. "Automatic Understanding of Image and Video Advertisements". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. 2017, pp. 1100–1110.

[122] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. "What makes an image memorable?" In: *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*. IEEE Computer Society, 2011, pp. 145–152. DOI: 10.1109/CVPR.2011.5995721.

[123] Bernard J Jansen, Danielle Booth, and Brian Smith. "Using the taxonomy of cognitive learning to model online searching". In: *Information Processing & Management* 45.6 (2009), pp. 643–663.

[124] Natasha Jaques, Sara Taylor, Akane Sano, and Rosalind Picard. "Multi-task, multi-kernel learning for estimating individual wellbeing". In: *Proc. NIPS Workshop on Multimodal Machine Learning, Montreal, Quebec*. Vol. 898. 2015.

[125] Taeho Jo. "Neural Based Approach to Keyword Extraction from Documents". In: *Computational Science and Its Applications - ICCSA 2003, International Conference, Montreal, Canada, May 18-21, 2003, Proceedings, Part I*. Ed. by Vipin Kumar, Marina L. Gavrilova, Chih Jeng Kenneth Tan, and Pierre L'Ecuyer. Vol. 2667. Lecture Notes in Computer Science. Springer, 2003, pp. 456–461. DOI: 10.1007/3-540-44839-X_49.

[126] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. "DenseCap: Fully Convolutional Localization Networks for Dense Captioning". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. 2016, pp. 4565–4574.

[127] Martin Johnson and Dominika Majewska. "Formal, Non-Formal, and Informal Learning: What Are They, and How Can We Research Them? Research Report." In: *Cambridge University Press & Assessment* (2022).

[128] Nivja H. de Jong and Ton Wempe. "Praat script to detect syllable nuclei and measure speech rate automatically". In: vol. 41. 2. May 2009, pp. 385–390. DOI: 10.3758/BRM.41.2.385. URL: https://doi.org/10.3758/BRM.41.2.385.

[129] Martin Joos. "Description of language design". In: *Journal of the Acoustical Society of America* 22.6 (1950), pp. 701–707.

[130] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. "Bag of Tricks for Efficient Text Classification". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*. Ed. by Mirella Lapata, Phil Blunsom, and Alexander Koller. Association for Computational Linguistics, 2017, pp. 427–431. DOI: 10.18653/v1/e17-2068.

[131] Rishita Kalyani and Ujwal Gadiraju. "Understanding User Search Behavior Across Varying Cognitive Levels". In: *Proceedings of the 30th ACM Conference on Hypertext and Social Media, HT 2019, Hof, Germany, September 17-20, 2019*. Ed. by Claus Atzenbeck, Jessica Rubart, and David E. Millard. ACM, 2019, pp. 123–132. DOI: 10.1145/3342220.3343643.

[132] Nick Kanopoulos, Nagesh Vasanthavada, and Robert L Baker. "Design of an image edge detection filter using the Sobel operator". In: *IEEE Journal of solid-state circuits* 23.2 (1988), pp. 358–367.

[133] Saraschandra Karanam and Herre van Oostendorp. "Age-related Differences in the Content of Search Queries when Reformulating". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7-12, 2016*. Ed. by Jofish Kaye, Allison Druin, Cliff Lampe, Dan Morris, and Juan Pablo Hourcade. ACM, 2016, pp. 5720–5730. DOI: 10.1145/2858036.2858444. URL: https://doi.org/10.1145/2858036.2858444.

[134] Saraschandra Karanam, Herre van Oostendorp, and Bipin Indurkhya. "Evaluating CoLiDeS + Pic: the role of relevance of pictures in user navigation behaviour". In: *Behav. Inf. Technol.* 31.1 (2012), pp. 31–40.

[135] Andrej Karpathy, Armand Joulin, and Fei-Fei Li. "Deep Fragment Embeddings for Bidirectional Image Sentence Mapping". In: *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. 2014, pp. 1889–1897.

[136] Khalid Al-Khatib, Michael Völske, Shahbaz Syed, Nikolay Kolyada, and Benno Stein. *Webis ChangeMyView Corpus 2020 (Webis-CMV-20)*. Zenodo, Apr. 2020. DOI: 10.5281/zenodo.3778298. URL: https://doi.org/10.5281/zenodo.3778298.

[137] J. Kiefer and J. Wolfowitz. "Stochastic Estimation of the Maximum of a Regression Function". In: *The Annals of Mathematical Statistics* 23.3 (1952), pp. 462–466. ISSN: 00034851. URL: http://www.jstor.org/stable/2236690 (visited on 09/02/2022).

[138] Su Nam Kim and Min-Yen Kan. "Re-examining Automatic Keyphrase Extraction Approaches in Scientific Articles". In: *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications, MWE@IJCNLP 2009, Singapore, August 6, 2009*. Ed. by Dimitra Anastasiou, Chikara Hashimoto, Preslav Nakov, and Su Nam Kim. Association for Computational Linguistics, 2009, pp. 9–16. URL: https://aclanthology.org/W09-2902/.

[139] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. "Derivation of new readability formulas (automated readability index, fog count and reading ease formula) for navy enlisted personnel". In: (1975).

[140] Rolf Kloepfer. *Komplementarität von Sprache und Bild: am Beispiel von Comic, Karikatur und Reklame*. Akad. Verlag-Gesell. Athenaion, 1977.

[141] Klaus Krippendorff. "Estimating the reliability, systematic error and random error of interval data". In: *Educational and Psychological Measurement* 30.1 (1970).

[142] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*. 2012, pp. 1106–1114.

[143] Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. "Integrating Text and Image: Determining Multimodal Document Intent in Instagram Posts". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*. 2019, pp. 4621–4631. DOI: 10.18653/v1/D19-1469.

[144] Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. "Age-of-acquisition ratings for 30,000 English words". In: *Behavior Research Methods* 44.4 (2012), pp. 978–990.

[145] Mohamed Zakaria Kurdi. "Lexical and Syntactic Features Selection for an Adaptive Reading Recommendation System Based on Text Complexity". In: *Proceedings of the 2017 International Conference on Information System and Data Mining*. ICISDM '17. Charleston, SC, USA: Association for Computing Machinery, 2017, pp. 66–69. ISBN: 9781450348331. DOI: 10.1145/3077584.3077595. URL: https://doi.org/10.1145/3077584.3077595.

[146] Xuan Nhat Lam, Thuc Vu, Trong Duc Le, and Anh Duc Duong. "Addressing cold-start problem in recommendation systems". In: *Proceedings of the 2nd International Conference on Ubiquitous Information Management and Communication, ICUIMC 2008, Suwon, Korea, January 31 - February 01, 2008*. Ed. by Won Kim and Hyung-Jin Choi. ACM, 2008, pp. 208–211. DOI: 10.1145/1352793.1352837.

[147] Weiyu Lan, Xirong Li, and Jianfeng Dong. "Fluency-Guided Cross-Lingual Image Captioning". In: *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*. 2017, pp. 1549–1557.

[148] J. Lemarié, L. Castillan, and H. Eyrolle. "Effects of expertise and multimedia presentation on the enactment and recall of procedural instructions". In: *Psychologie Française* 62.4 (2017), pp. 351–359. ISSN: 0033-2984.

[149] Zechao Li, Jinhui Tang, Xueming Wang, Jing Liu, and Hanqing Lu. "Multimedia News Summarization in Search". In: *ACM Trans. Intell. Syst. Technol.* 7.3 (2016), 33:1–33:20. DOI: 10.1145/2822907.

[150] Jian Liang, Zhihang Li, Dong Cao, Ran He, and Jingdong Wang. "Self-Paced Cross-Modal Subspace Matching". In: *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*. 2016, pp. 569–578.

[151] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. "Microsoft COCO: Common Objects in Context". In: *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*. 2014, pp. 740–755.

[152] *List of antonyms and opposites*. http://www.myenglishpages.com/site_php_files/vocabulary-lesson-opposites.php. Accessed: 23 November 2017.

[153] Fayao Liu, Luping Zhou, Chunhua Shen, and Jianping Yin. "Multiple kernel learning in the primal for multimodal Alzheimer's disease classification." In: *IEEE J. Biomedical and Health Informatics* 18.3 (2014), pp. 984–990.

[154] Fei Liu, Feifan Liu, and Yang Liu. "A Supervised Framework for Keyword Extraction From Meeting Transcripts". In: *IEEE Trans. Speech Audio Process.* 19.3 (2011), pp. 538–548. DOI: 10.1109/TASL.2010.2052119.

[155] Hanrui Liu, Chang Liu, and Nicholas J Belkin. "Investigation of users' knowledge change process in learning-related search tasks". In: *Proceedings of the Association for Information Science and Technology* 56.1 (2019), pp. 166–175.

[156] Jingzhou Liu, Dominic J. D. Hughes, and Yiming Yang. "Unsupervised Extractive Text Summarization with Distance-Augmented Sentence Graphs". In: *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*. Ed. by Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai. ACM, 2021, pp. 2313–2317. DOI: 10.1145/3404835.3463111.

[157] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. "A ConvNet for the 2020s". In: *CoRR* abs/2201.03545 (2022). arXiv: 2201.03545. URL: https://arxiv.org/abs/2201.03545.

[158] Margaret C Lohman. "Work situations triggering participation in informal learning in the workplace: A case study of public school teachers". In: *Performance Improvement Quarterly* 16.1 (2003), pp. 40–54.

[159] Lu Lu, Yeonjong Shin, Yanhui Su, and George Em Karniadakis. "Dying ReLU and Initialization: Theory and Numerical Examples". In: *ArXiv* abs/1903.06733 (2019).

[160] Zheng Lu and Kristen Grauman. "Story-Driven Summarization for Egocentric Video". In: *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*. IEEE Computer Society, 2013, pp. 2714–2721. DOI: 10.1109/CVPR.2013.350.

[161] Yu-Fei Ma, Xian-Sheng Hua, Lie Lu, and HongJiang Zhang. "A generic framework of user attention model and its application in video summarization". In: *IEEE Trans. Multim.* 7.5 (2005), pp. 907–919. DOI: 10.1109/TMM.2005.854410.

[162] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. "Rectifier nonlinearities improve neural network acoustic models". In: *Proc. icml*. Vol. 30. 1. Citeseer. 2013, p. 3.

[163] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. "Unsupervised Video Summarization with Adversarial LSTM Networks". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 2982–2991. DOI: 10.1109/CVPR.2017.318.

[164] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. "The Stanford CoreNLP Natural Language Processing Toolkit". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, System Demonstrations*. The Association for Computer Linguistics, 2014, pp. 55–60. DOI: 10.3115/v1/p14-5010.

[165] Gary Marchionini. "Exploratory search: from finding to understanding". In: *Commun. ACM* 49.4 (2006), pp. 41–46. DOI: 10.1145/1121949.1121979.

[166] Emily E Marsh and Marilyn Domas White. "A taxonomy of relationships between images and text". In: *Journal of Documentation* 59.6 (2003), pp. 647–672.

[167] Radan Martinec and Andrew Salway. "A system for image–text relations in new (and old) media". In: *Visual Communication* 4.3 (2005), pp. 337–371.

[168] Roberto Martínez Maldonado, Vanessa Echeverría, Gloria Fernandez Nieto, and Simon Buckingham Shum. "From Data to Insights: A Layered Storytelling Approach for Multimodal Learning Analytics". In: *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*. Ed. by Regina Bernhaupt, Florian 'Floyd' Mueller, David Verweij, Josh Andres, Joanna McGrenere, Andy Cockburn, Ignacio Avellino, Alix Goguey, Pernille Bjøn, Shengdong Zhao, Briane Paul Samson, and Rafal Kocielnik. ACM, 2020, pp. 1–15. DOI: 10.1145/3313831.3376148. URL: https://doi.org/10.1145/3313831.3376148.

[169] *MASSIVE OPEN ONLINE COURSE (MOOC) MARKET*. https://www.mordorintelligence.com/industry-reports/massive-open-online-course-mooc-market. Accessed: 12 July 2021.

[170] Gerald Matthews, Sian E Campbell, Shona Falconer, Lucy A Joyner, Jane Huggins, Kirby Gilliland, Rebecca Grier, and Joel S Warm. "Fundamental dimensions of subjective state in performance settings: task engagement, distress, and worry". In: *Emotion (Washington, D.C.)* 2.4 (Dec. 2002), pp. 315–340. ISSN: 1528-3542. DOI: 10.1037/1528-3542.2.4.315. URL: https://doi.org/10.1037/1528-3542.2.4.315.

[171] Richard E Mayer and Roxana Moreno. "A split-attention effect in multimedia learning: Evidence for dual processing systems in working memory." In: *Journal of educational psychology* 90.2 (1998), p. 312.

[172] Richard E. Mayer and Roxana Moreno. "Nine Ways to Reduce Cognitive Load in Multimedia Learning". In: *Educational Psychologist* 38.1 (2003), pp. 43–52.

[173] Masoud Mazloom, Robert Rietveld, Stevan Rudinac, Marcel Worring, and Willemijn van Dolen. "Multimodal Popularity Prediction of Brand-related Social Media Posts". In: *Proceedings of the 2016 ACM Conference on Multimedia MM 2016, Amsterdam, The Netherlands, October 15-19, 2016*. 2016, pp. 197–201.

[174] G Harry Mc Laughlin. "SMOG grading-a new readability formula". In: *Journal of reading* 12.8 (1969), pp. 639–646.

[175] Brian McClanahan and Swapna S. Gokhale. "Interplay between video recommendations, categories, and popularity on YouTube". In: *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation, SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI 2017, San Francisco, CA, USA, August 4-8, 2017*. IEEE, 2017, pp. 1–7. DOI: 10.1109/UIC-ATC.2017.8397661.

[176] Scott McCloud. "Understanding comics: The invisible art". In: *Northampton, Mass* (1993).

[178] Andreia-Simona Melnic and Nicoleta Botez. "Formal, non-formal and informal interdependence in education". In: *Economy Transdisciplinarity Cognition* 17.1 (2014), pp. 113–118.

[179] Martin Merkt and Stephan Schwan. "How does interactivity in videos affect task performance?" In: *Computers in Human Behavior* 31 (2014), pp. 172–181. ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2013.10.018. URL: http://www.sciencedirect.com/science/article/pii/S0747563213003683.

[180] Martin Merkt, Sonja Weigand, Anke Heier, and Stephan Schwan. "Learning with videos vs. learning with print: The role of interactive features". In: *Learning and Instruction* 21.6 (2011), pp. 687–704. ISSN: 0959-4752. DOI: https://doi.org/10.1016/j.learninstruc.2011.03.004. URL: http://www.sciencedirect.com/science/article/pii/S0959475211000247.

[181] Rada Mihalcea and Paul Tarau. "TextRank: Bringing Order into Text". In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain*. ACL, 2004, pp. 404–411. URL: https://aclanthology.org/W04-3252/.

[182] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).

[183] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. "Distributed Representations of Words and Phrases and their Compositionality". In: *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.* 2013, pp. 3111–3119.

[184] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. "Learning Joint Embedding with Multimodal Cues for Cross-Modal Video-Text Retrieval". In: *Proceedings of the 2018 ACM International Conference on Multimedia Retrieval*. ACM. 2018, pp. 19–27.

[185] Niluthpol Chowdhury Mithun, Rameswar Panda, Evangelos E. Papalexakis, and Amit K. Roy-Chowdhury. "Webly Supervised Joint Embedding for Cross-Modal Image-Text Retrieval". In: *Proceedings of the 26th ACM International Conference on Multimedia*. MM '18. Seoul, Republic of Korea: ACM, 2018, pp. 1856–1864. ISBN: 978-1-4503-5665-7.

[186] Antonija Mitrovic, Matthew Gordon, Alicja Piotrkowicz, and Vania Dimitrova. "Investigating the Effect of Adding Nudges to Increase Engagement in Active Video Watching". In: *Artificial Intelligence in Education*. Ed. by Seiji Isotani, Eva Millán, Amy Ogan, Peter Hastings, Bruce McLaren, and Rose Luckin. Cham: Springer International Publishing, 2019, pp. 320–332.

[187] Felipe Moraes, Sindunuraga Rikarno Putra, and Claudia Hauff. "Contrasting Search as a Learning Activity with Instructor-designed Learning". In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018*. Ed. by Alfredo Cuzzocrea, James Allan, Norman W. Paton, Divesh Srivastava, Rakesh Agrawal, Andrei Z. Broder, Mohammed J. Zaki, K. Selçuk Candan, Alexandros Labrinidis, Assaf Schuster, and Haixun Wang. ACM, 2018, pp. 167–176. DOI: 10.1145/3269206.3271676. URL: https://doi.org/10.1145/3269206.3271676.

[190] Eric Müller-Budack, Jonas Theiner, Sebastian Diering, Maximilian Idahl, and Ralph Ewerth. "Multimodal Analytics for Real-world News using Measures of Cross-modal Entity Consistency". In: *Proceedings of the 2020 on International Conference on Multimedia Retrieval, ICMR 2020, Dublin, Ireland, June 8-11, 2020*. Ed. by Cathal Gurrin, Björn Þór Jónsson, Noriko Kando, Klaus Schöffmann, Yi-Ping Phoebe Chen, and Noel E. O'Connor. ACM, 2020, pp. 16–25. ISBN: 9781450370875. DOI: 10.1145/3372278.3390670.

[191] Vinod Nair and Geoffrey E. Hinton. "Rectified Linear Units Improve Restricted Boltzmann Machines". In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*. Ed. by Johannes Fürnkranz and Thorsten Joachims. Omnipress, 2010, pp. 807–814. URL: https://icml.cc/Conferences/2010/papers/432.pdf.

[192] Winfried Nöth. *Handbook of semiotics*. Indiana University Press, 1995.

[193] Heather L. O'Brien, Andrea Kampen, Amelia W. Cole, and Kathleen Brennan. "The Role of Domain Knowledge in Search as Learning". In: *CHIIR '20: Conference on Human Information Interaction and Retrieval, Vancouver, BC, Canada, March 14-18, 2020*. Ed. by Heather L. O'Brien, Luanne Freund, Ioannis Arapakis, Orland Hoeber, and Irene Lopatovska. ACM, 2020, pp. 313–317. DOI: 10.1145/3343413.3377989.

[194] Xavier Ochoa, Marcelo Worsley, Katherine Chiluiza, and Saturnino Luz. "MLA'14: Third Multimodal Learning Analytics Workshop and Grand Challenges". In: *Proceedings of the 16th International Conference on Multimodal Interaction, ICMI 2014, Istanbul, Turkey, November 12-16, 2014*. Istanbul, Turkey: ACM, 2014, pp. 531–532. DOI: 10.1145/2663204.2668318.

[195] *Online learning cannot just be for those who can afford its technology*. https://www.nature.com/articles/d41586-020-02709-3. Accessed: 2 November 2021.

[196] Herre van Oostendorp, Saraschandra Karanam, and Bipin Indurkhya. "CoLiDeS+ Pic: a cognitive model of web-navigation based on semantic information from pictures". In: *Behav. Inf. Technol.* 31.1 (2012), pp. 17–30.

[203] Paul Over, Alan F. Smeaton, and George Awad. "The trecvid 2008 BBC rushes summarization evaluation". In: *Proceedings of the 2nd ACM Workshop on Video Summarization, TVS 2008, Vancouver, British Columbia, Canada, October 31, 2008*. Ed.

by Paul Over and Alan F. Smeaton. ACM, 2008, pp. 1–20. DOI: 10.1145/1463563.1463564.

[204] W Bradford Paley. "TextArc: Showing word frequency and distribution in text". In: *Poster presented at IEEE Symposium on Information Visualization*. Vol. 2002. 2002.

[205] Shailendra Palvia, Prageet Aeron, Parul Gupta, Diptiranjan Mahapatra, Ratri Parida, Rebecca Rosner, and Sumita Sindhi. "Online Education: Worldwide Status, Challenges, Trends, and Implications". In: *Journal of Global Information Technology Management* 21.4 (2018), pp. 233–241. DOI: 10.1080/1097198X.2018.1542262.

[206] Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. "VALSE: A Task-Independent Benchmark for Vision and Language Models Centered on Linguistic Phenomena". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Association for Computational Linguistics, 2022, pp. 8253–8280. URL: https://aclanthology.org/2022.acl-long.567.

[207] Georg Pardi, Johannes von Hoyer, Peter Holtz, and Yvonne Kammerer. "The Role of Cognitive Abilities and Time Spent on Texts and Videos in a Multimodal Searching as Learning Task". In: *CHIIR '20: Conference on Human Information Interaction and Retrieval, Vancouver, BC, Canada, March 14-18, 2020*. Ed. by Heather L. O'Brien, Luanne Freund, Ioannis Arapakis, Orland Hoeber, and Irene Lopatovska. ACM, 2020, pp. 378–382. DOI: 10.1145/3343413.3378001.

[208] Georg Pardi, Johannes von Hoyer, Peter Holtz, and Yvonne Kammerer. "The Role of Cognitive Abilities and Time Spent on Texts and Videos in a Multimodal Searching as Learning Task". In: *CHIIR '20: Conference on Human Information Interaction and Retrieval, Vancouver, BC, Canada, March 14-18, 2020*. Ed. by Heather L. O'Brien, Luanne Freund, Ioannis Arapakis, Orland Hoeber, and Irene Lopatovska. ACM, 2020, pp. 378–382. DOI: 10.1145/3343413.3378001.

[209] Georg Pardi, Yvonne Kammerer, and Peter Gerjets. "Search and Justification Behavior During Multimedia Web Search for Procedural Knowledge". In: *Companion Publication of the 10th ACM Conference on Web Science*. WebSci '19. Boston, Massachusetts, USA: ACM, 2019, pp. 17–20. ISBN: 978-1-4503-6174-3. DOI: 10.1145/3328413.3329405. URL: http://doi.acm.org/10.1145/3328413.3329405.

[210] Wuxu Peng, Linda Huang, Julia Jia, and Emma Ingram. "Enhancing the Naive Bayes Spam Filter Through Intelligent Text Modification Detection". In: *17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications / 12th IEEE International Conference On Big Data Science And Engineering, TrustCom/BigDataSE 2018, New York, NY, USA, August 1-3, 2018*. IEEE, 2018, pp. 849–854. DOI: 10.1109/TrustCom/BigDataSE.2018.00122.

[211] Gustavo Penha, Alexandru Balan, and Claudia Hauff. "Introducing MANtIS: a novel Multi-Domain Information Seeking Dialogues Dataset". In: *CoRR* abs/1912.04639 (2019). arXiv: 1912.04639. URL: http://arxiv.org/abs/1912.04639.

[212] James W Pennebaker, Martha E Francis, and Roger J Booth. "Linguistic inquiry and word count: LIWC 2001". In: *Mahway: Lawrence Erlbaum Associates* 71.2001 (2001), p. 2001.

[213] B Phillips. "A calculus of cohesion". In: *Fourth LACUS Forum, Montreal, Canada*. 1977.

[214] John Platt. "Sequential minimal optimization: A fast algorithm for training support vector machines". In: (1998).

[215] Soujanya Poria, Erik Cambria, and Alexander F. Gelbukh. "Deep Convolutional Neural Network Textual Features and Multiple Kernel Learning for Utterance-level Multimodal Sentiment Analysis". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*. 2015, pp. 2539–2544.

[216] Verónica Proaño-Ríos and Roberto I. González-Ibáñez. "Dataset of Search Results Organized as Learning Paths Recommended by Experts to Support Search as Learning". In: *Data* 5.4 (2020), p. 92. DOI: 10.3390/data5040092. URL: https://doi.org/10.3390/data5040092.

[217] Jinwei Qi, Yuxin Peng, and Yunkan Zhuo. "Life-long Cross-media Correlation Learning". In: *2018 ACM Conference on Multimedia, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*. 2018, pp. 528–536.

[218] Vasili Ramanishka, Abir Das, Dong Huk Park, Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, and Kate Saenko. "Multimodal Video Description". In: *Proceedings of the 2016 ACM Conference on Multimedia, MM 2016, Amsterdam, The Netherlands, October 15-19, 2016*. 2016, pp. 1092–1096.

[219] Nils Reimers and Iryna Gurevych. "Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2020. URL: https://arxiv.org/abs/2004.09813.

[220] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*. Ed. by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan. Association for Computational Linguistics, 2019, pp. 3980–3990. DOI: 10.18653/v1/D19-1410.

[221] Ulrike Reiner. "Automatic Analysis of Dewey Decimal Classification Notations". In: *Data Analysis, Machine Learning and Applications - Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation e.V., Albert-Ludwigs-Universität Freiburg, March 7-9, 2007*. Ed. by Christine Preisach, Hans Burkhardt, Lars Schmidt-Thieme, and Reinhold Decker. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, 2007, pp. 697–704. DOI: 10.1007/978-3-540-78246-9_82.

[222] Dongxiao Ren, Weihua Xu, Zhonghua Wang, and Qinxiu Sun. "Deep Label Feature Fusion Hashing for Cross-Modal Retrieval". In: *IEEE Access* 10 (2022), pp. 100276–100285. DOI: 10.1109/ACCESS.2022.3208147. URL: https://doi.org/10.1109/ACCESS.2022.3208147.

[223] Juliane Richter, Katharina Scheiter, and Alexander Eitel. *Signaling text-picture relations in multimedia learning: The influence of prior knowledge.* Richter, Juliane: Leibniz-Institut für Wissensmedien, Schleichstrasse 6, Tübingen, Switzerland, 72076, j.richter@iwm-tuebingen.de, 2018. DOI: 10.1037/edu0000220.

[224] Nirmal Roy, Felipe Moraes, and Claudia Hauff. "Exploring Users' Learning Gains within Search Sessions". In: *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. 2020, pp. 432–436.

[225] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. "Learning representations by back-propagating errors". In: *nature* 323.6088 (1986), pp. 533–536.

[226] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision* 115.3 (2015), pp. 211–252.

[227] Dario D. Salvucci and Joseph H. Goldberg. "Identifying fixations and saccades in eye-tracking protocols". In: *Proceedings of the Eye Tracking Research & Application Symposium, ETRA 2000, Palm Beach Gardens, Florida, USA, November 6-8, 2000*. Ed. by Andrew T. Duchowski. ACM, 2000, pp. 71–78. DOI: 10.1145/355017.355028.

[228] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. "MobileNetV2: Inverted Residuals and Linear Bottlenecks". In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 4510–4520. DOI: 10.1109/CVPR.2018.00474.

[229] Rossano Schifanella, Paloma de Juan, Joel R. Tetreault, and Liangliang Cao. "Detecting Sarcasm in Multimodal Social Platforms". In: *Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016, Amsterdam, The Netherlands, October 15-19, 2016*. 2016, pp. 1136–1145.

[230] Florian Schmidt-Weigand and Katharina Scheiter. "The role of spatial descriptions in learning from multimedia". In: *Computers in Human Behavior* 27.1 (2011), pp. 22–28.

[231] Wolfgang Schneider, Matthias Schlagmüller, and Marco Ennemoser. *LGVT 6-12: Lesegeschwindigkeits-und-verständnistest für die Klassen 6-12*. Hogrefe Göttingen, 2007.

[232] Anne Schüler, Jana Arndt, and Katharina Scheiter. "Does text–picture integration also occur with longer text segments?" In: *Applied Cognitive Psychology* 33.6 (2019), pp. 1137–1146.

[233] Anne Schüler, Francesca Pazzaglia, and Katharina Scheiter. "Specifying the boundary conditions of the multimedia effect: The influence of content and its distribution between text and pictures". In: *British Journal of Psychology* 110.1 (2019), pp. 126–150.

[234] Björn W. Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus R. Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, Marcello Mortillaro, Hugues Salamin, Anna Polychroniou, Fabio Valente, and Samuel Kim. "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism". In: *INTERSPEECH 2013: 14th Annual Conference of the International Speech Communication Association*. Lyon, France: INTERSPEECH 2013, 2013.

[235] *Sentence-Transformer Model*. https://huggingface.co/sentence-transformers/roberta-large-nli-stsb-mean-tokens. Accessed: 16 November 2021.

[236] Anand Senthi. *Advantages & Disadvantage of Web Based Learning*. http://anandsenthi.emergucate.com/advantages-disadvantage-of-web-based-learning/. Accessed: 1 November 2021.

[237] Mathias Seuret, Michele Alberti, Marcus Liwicki, and Rolf Ingold. "PCA-Initialized Deep Neural Networks Applied to Document Image Analysis". In: *14th IAPR International Conference on Document Analysis and Recognition, ICDAR 2017, Kyoto, Japan, November 9-15, 2017*. IEEE, 2017, pp. 877–882. DOI: 10.1109/ICDAR.2017.148.

[238] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. "Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*. Ed. by Iryna Gurevych and Yusuke Miyao. Association for Computational Linguistics, 2018, pp. 2556–2565. DOI: 10.18653/v1/P18-1238.

[239] Baoguang Shi, Xiang Bai, and Cong Yao. "An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 39.11 (2017), pp. 2298–2304. DOI: 10.1109/TPAMI.2016.2646371.

[241] Ekaterina Shutova, Douwe Kiela, and Jean Maillard. "Black Holes and White Rabbits: Metaphor Identification with Visual Features". In: *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*. 2016, pp. 160–170.

[242] Shivangi Singhal, Anubha Kabra, Mohit Sharma, Rajiv Ratn Shah, Tanmoy Chakraborty, and Ponnurangam Kumaraguru. "SpotFake+: A Multimodal Framework for Fake News Detection via Transfer Learning (Student Abstract)". In: *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 2020, pp. 13915–13916. URL: https://ojs.aaai.org/index.php/AAAI/article/view/7230.

[243] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh C. Jain. "Content-Based Image Retrieval at the End of the Early Years". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 22.12 (2000), pp. 1349–1380.

[244] Aaron Smith, Skye Toor, and Patrick Van Kessel. "Many turn to YouTube for children's content, news, how-to lessons". In: *Pew Research Centre* 7 (2018). URL: https://www.pewresearch.org/internet/2018/11/07/many-turn-to-youtube-for-childrens-content-news-how-to-lessons/.

[245] Edgar A Smith and J Peter Kincaid. "Derivation and validation of the automated readability index for use with technical materials". In: *Human factors* 12.5 (1970), pp. 457–564.

[246] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. "MPNet: Masked and Permuted Pre-training for Language Understanding". In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Ed. by Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin. 2020. URL: https://proceedings.neurips.cc/paper/2020/hash/c3a690be93aa602ee2dc0ccab5b7b67e-Abstract.html.

[247] Sanja Stajner, Richard Evans, Constantin Orasan, and Ruslan Mitkov. "What can readability measures really tell us about text complexity?" In: *Proceedings of Workshop on natural language processing for improving textual accessibility* (Jan. 2012), pp. 14–22.

[248] *Stanford Log-linear Part-Of-Speech Tagger*. https://nlp.stanford.edu/software/tagger.shtml. Accessed: 2 November 2021.

[249] Jan Stewart. "Recalibrating the Flesch Readability Index for the twenty-first century". In: *Bulletin of the International Cultural Research Institute of Chikushi Jogakuen University/Junior College.* (2003). URL: http://www.stewartenglish.com/Recalibrating_Abstract.pdf.

[250] Rohail Syed and Kevyn Collins-Thompson. "Exploring Document Retrieval Features Associated with Improved Short- and Long-term Vocabulary Learning Outcomes". In: *Proceedings of the 2018 Conference on Human Information Interaction and Retrieval, CHIIR 2018, New Brunswick, NJ, USA, March 11-15, 2018*. Ed. by Chirag

Shah, Nicholas J. Belkin, Katriina Byström, Jeff Huang, and Falk Scholer. ACM, 2018, pp. 191–200. DOI: 10.1145/3176349.3176397.

[251] Rohail Syed and Kevyn Collins-Thompson. "Retrieval algorithms optimized for human learning". In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM. 2017, pp. 555–564.

[252] *SyllaPy*. https://github.com/mholtzscher/syllapy. Accessed: 15 November 2021.

[253] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning". In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*. 2017, pp. 4278–4284.

[254] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going deeper with convolutions". In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. 2015, pp. 1–9.

[255] Davide Taibi, Francesca Bianchi, Philipp Kemkes, and Ivana Marenzi. "Learning Analytics for Interpreting". In: *Proceedings of the 10th International Conference on Computer Supported Education, CSEDU 2018, Funchal, Madeira, Portugal, March 15-17, 2018, Volume 1*. Ed. by Bruce M. McLaren, Rob Reilly, Susan Zvacek, and James Onohuome Uhomoibhi. SciTePress, 2018, pp. 145–154. DOI: 10.5220/0006774801450154.

[256] Mingxing Tan and Quoc V. Le. "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks". In: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 6105–6114. URL: http://proceedings.mlr.press/v97/tan19a.html.

[257] David S. Taubman and Michael W. Marcellin. *JPEG2000 - image compression fundamentals, standards and practice*. Vol. 642. The Kluwer international series in engineering and computer science. Kluwer, 2002. ISBN: 978-0-7923-7519-7. DOI: 10.1007/978-1-4615-0799-4.

[258] Ann Taylor, Mitchell Marcus, and Beatrice Santorini. "The Penn Treebank: An Overview". In: *Treebanks: Building and Using Parsed Corpora*. Ed. by Anne Abeillé. Dordrecht: Springer Netherlands, 2003, pp. 5–22. ISBN: 978-94-010-0201-1. DOI: 10.1007/978-94-010-0201-1_1. URL: https://doi.org/10.1007/978-94-010-0201-1_1.

[259] *TIB AV-Portal*. https://av.tib.eu. Accessed: 2 November 2021.

[260] *TIB AV-Portal Open Data*. https://av.tib.eu/opendata. Accessed: 9 November 2021.

[261] Christos Troussas, Maria Virvou, Kurt Junshean Espinosa, Kevin Llaguno, and Jaime D. L. Caro. "Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning". In: *4th International Conference on Information, Intelligence, Systems and Applications, IISA 2013, Piraeus, Greece, July 10-12, 2013*. Ed. by Nikolaos G. Bourbakis, George A. Tsihrintzis, and Maria Virvou. IEEE, 2013, pp. 1–6. DOI: 10.1109/IISA.2013.6623713.

[262] Peter D. Turney. "Learning Algorithms for Keyphrase Extraction". In: *Inf. Retr.* 2.4 (2000), pp. 303–336. DOI: 10.1023/A:1009976227802.

[263] *Udacity*. https://udacity.com. Accessed: 2 November 2021.

[264] Len Unsworth. "Image/text relations and intersemiosis: Towards multimodal text description for multiliteracies education". In: *Proceedings of the 33rd International Systemic Functional Congress*. 2007, pp. 1165–1205.

[265] Arash Vahdat, Guang-Tong Zhou, and Greg Mori. "Discovering Video Clusters from Visual Features and Noisy Tags". In: *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI*. Ed. by David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars. Vol. 8694. Lecture Notes in Computer Science. Springer, 2014, pp. 526–539. DOI: 10.1007/978-3-319-10599-4_34.

[266] Pertti Vakkari. "Searching as learning: A systematization based on literature". In: *J. Inf. Sci.* 42.1 (2016), pp. 7–18. DOI: 10.1177/0165551515615833.

[267] Erlijn Van Genuchten, Katharina Scheiter, and Anne Schüler. "Examining learning from text and pictures for different task types: Does the multimedia effect differ for conceptual, causal, and procedural tasks?" In: *Computers in human behavior* 28.6 (2012), pp. 2209–2218.

[268] Theo Van Leeuwen. *Introducing Social Semiotics*. Psychology Press, 2005.

[269] Theo Van Leeuwen and Gunther Kress. "Multimodal discourse: the modes and media of contemporary communication". In: *London: Arnold* (2001).

[270] Deepika Varshney and Dinesh Kumar Vishwakarma. "A unified approach for detection of Clickbait videos on YouTube using cognitive evidences". In: *Appl. Intell.* 51.7 (2021), pp. 4214–4235. DOI: 10.1007/s10489-020-02057-9.

[271] Alakananda Vempala and Daniel Preotiuc-Pietro. "Categorizing and Inferring the Relationship between the Text and Image of Twitter Posts". In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. Ed. by Anna Korhonen, David R. Traum, and Lluís Màrquez. Association for Computational Linguistics, 2019, pp. 2830–2840. DOI: 10.18653/v1/p19-1272.

[272] Katrien Verbert, Hendrik Drachsler, Nikos Manouselis, Martin Wolpers, Riina Vuorikari, and Erik Duval. "Dataset-driven research for improving recommender systems for learning". In: *Proceedings of the 1st International Conference on Learning Analytics and Knowledge, LAK 2011, Banff, AB, Canada, February 27 - March 01, 2011*. Ed. by Phillip Long, George Siemens, Gráinne Conole, and Dragan Gasevic. ACM, 2011, pp. 44–53. DOI: 10.1145/2090116.2090122. URL: https://doi.org/10.1145/2090116.2090122.

[273] Michael Völske, Matthias Hagen, and Benno Stein. *Webis Query-Task-Mapping Corpus 2019 (Webis- QTM-19)*. Zenodo, May 2019. DOI: 10.5281/zenodo.3257431. URL: https://doi.org/10.5281/zenodo.3257431.

[274] Jörg Waitelonis and Harald Sack. "Augmenting Video Search with Linked Open Data". In: *5th International Conference on Semantic Systems, Graz, Austria, September 2-4, 2009. Proceedings*. Ed. by Adrian Paschke, Hans Weigand, Wernher Behrendt, Klaus Tochtermann, and Tassilo Pellegrini. Verlag der Technischen Universität Graz, 2009, pp. 550–558. URL: https://www.fiz-karlsruhe.de/sites/default/files/FIZ/Dokumente/Forschung/ISE/Publications/Conferences-Workshops/2009a_Waitelonis.pdf.

[275] Jörg Waitelonis and Harald Sack. "Towards exploratory video search using linked data". In: *Multim. Tools Appl.* 59.2 (2012), pp. 645–672. DOI: 10.1007/s11042-011-0733-1.

[276] Jiabing Wang and Hong Peng. "Keyphrases Extraction from Web Document by the Least Squares Support Vector Machine". In: *2005 IEEE / WIC / ACM International Conference on Web Intelligence (WI 2005), 19-22 September 2005, Compiegne, France*. Ed. by Andrzej Skowron, Rakesh Agrawal, Michael Luck, Takahira Yamaguchi, Pierre Morizet-Mahoudeaux, Jiming Liu, and Ning Zhong. IEEE Computer Society, 2005, pp. 293–296. DOI: 10.1109/WI.2005.87.

[277] Jiabing Wang, Hong Peng, and Jing-Song Hu. "Automatic Keyphrases Extraction from Document Using Neural Network". In: *Advances in Machine Learning and Cybernetics, 4th International Conference, ICMLC 2005, Guangzhou, China, August 18-21, 2005, Revised Selected Papers*. Ed. by Daniel S. Yeung, Zhi-Qiang Liu, Xizhao Wang, and Hong Yan. Vol. 3930. Lecture Notes in Computer Science. Springer, 2005, pp. 633–641. DOI: 10.1007/11739685_66.

[278] Jingya Wang, Xiatian Zhu, and Shaogang Gong. "Video Semantic Clustering with Sparse and Incomplete Tags". In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*. Ed. by Dale Schuurmans and Michael P. Wellman. AAAI Press, 2016, pp. 3618–3624. URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12034.

[279] Meng Wang, Richang Hong, Guangda Li, Zheng-Jun Zha, Shuicheng Yan, and Tat-Seng Chua. "Event Driven Web Video Summarization by Tag Localization

and Key-Shot Identification". In: *IEEE Trans. Multim.* 14.4 (2012), pp. 975–985. DOI: 10.1109/TMM.2012.2185041.

[280] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. "Multiscale structural similarity for image quality assessment". In: *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. Vol. 2. Ieee. 2003, pp. 1398–1402.

[281] Mark Warschauer, Lynn Reimer, Kameryn Denaro, Gabe Orona, Katerina Schenke, Tutrang Nguyen, Amanda Niili, Di Xu, Sabrina Solanki, and Tate Tamara. *Evaluating Promising Practices in Undergraduate STEM Lecture Courses*. Zenodo, Feb. 2021. DOI: 10.7280/D11M5Q. URL: https://doi.org/10.7280/D11M5Q.

[282] *Which digital learning materials do you use in your classroom in a typical week?* https://www.statista.com/statistics/658475/us-classroom-digital-learning-materials-weekly-usage/. Accessed: 12 July 2021.

[283] Christoph Wick and Frank Puppe. "Fully Convolutional Neural Networks for Page Segmentation of Historical Document Images". In: *13th IAPR International Workshop on Document Analysis Systems, DAS 2018, Vienna, Austria, April 24-27, 2018*. IEEE Computer Society, 2018, pp. 287–292. DOI: 10.1109/DAS.2018.39.

[284] Jennifer Wiley, Susan R. Goldman, Arthur C. Graesser, Christopher A. Sanchez, Ivan K. Ash, and Joshua A. Hemmerich. "Source Evaluation, Comprehension, and Learning in Internet Science Inquiry Tasks". In: *American Educational Research Journal* 46.4 (2009), pp. 1060–1106.

[285] Teena Willoughby, S. Alexandria Anderson, Eileen Wood, Julie Mueller, and Craig Ross. "Fast searching for information on the Internet to use in a learning context: The impact of domain knowledge". In: *Comput. Educ.* 52.3 (2009), pp. 640–648. DOI: 10.1016/j.compedu.2008.11.009.

[286] Max L. Wilson, Chaoyu Ye, Michael B. Twidale, and Hannah Grasse. "Search Literacy: Learning to Search to Learn". In: *Proceedings of the Second International Workshop on Search as Learning, SAL 2016, co-located with the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 21st, 2016*. Ed. by Jacek Gwizdka, Preben Hansen, Claudia Hauff, Jiyin He, and Noriko Kando. Vol. 1647. CEUR Workshop Proceedings. CEUR-WS.org, 2016.

[287] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. "Aggregated Residual Transformations for Deep Neural Networks". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 5987–5995. DOI: 10.1109/CVPR.2017.634.

[288] Nan Xu and Wenji Mao. "MultiSentiNet: A Deep Semantic Network for Multimodal Sentiment Analysis". In: *Proceedings of the 2017 ACM Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*. 2017, pp. 2399–2402.

[289] Xing Xu, Jingkuan Song, Huimin Lu, Yang Yang, Fumin Shen, and Zi Huang. "Modal-adversarial Semantic Learning Network for Extendable Cross-modal Retrieval". In: *Proceedings of the 2018 ACM International Conference on Multimedia Retrieval, ICMR 2018, Yokohama, Japan, June 11-14, 2018*. 2018, pp. 46–54.

[290] Keiji Yanai and Kobus Barnard. "Image region entropy: a measure of "visualness" of web images associated with one concept". In: *ACM International Conference on Multimedia, Singapore*. 2005, pp. 419–422.

[291] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. "Hierarchical Attention Networks for Document Classification". In: *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*. 2016, pp. 1480–1489.

[292] Yi-Ren Yeh, Ting-Chu Lin, Yung-Yu Chung, and Yu-Chiang Frank Wang. "A Novel Multiple Kernel Learning Framework for Heterogeneous Feature Fusion and Variable Selection". In: *IEEE Transactions on Multimedia* 14.3-1 (2012), pp. 563–574.

[293] Ran Yu, Ujwal Gadiraju, Peter Holtz, Markus Rokicki, Philipp Kemkes, and Stefan Dietze. "Predicting User Knowledge Gain in Informational Search Sessions". In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*. Ed. by Kevyn Collins-Thompson, Qiaozhu Mei, Brian D. Davison, Yiqun Liu, and Emine Yilmaz. ACM, 2018, pp. 75–84. DOI: 10.1145/3209978.3210064.

[294] Arianna Yuan and Yang Li. "Modeling Human Visual Search Performance on Realistic Webpages Using Analytical and Deep Learning Methods". In: *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*. Ed. by Regina Bernhaupt, Florian 'Floyd' Mueller, David Verweij, Josh Andres, Joanna McGrenere, Andy Cockburn, Ignacio Avellino, Alix Goguey, Pernille Bjøn, Shengdong Zhao, Briane Paul Samson, and Rafal Kocielnik. ACM, 2020, pp. 1–12. DOI: 10.1145/3313831.3376870. URL: https://doi.org/10.1145/3313831.3376870.

[295] Chengzhi Zhang. "Automatic keyword extraction from documents using conditional random fields". In: *Journal of Computational Information Systems* 4.3 (2008), pp. 1169–1180.

[296] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. "Summary Transfer: Exemplar-Based Subset Selection for Video Summarization". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June*

27-30, 2016. IEEE Computer Society, 2016, pp. 1059–1067. DOI: 10.1109/CVPR.2016.120.

[297] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. "Video Summarization with Long Short-Term Memory". In: *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VII*. Ed. by Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling. Vol. 9911. Lecture Notes in Computer Science. Springer, 2016, pp. 766–782. DOI: 10.1007/978-3-319-46478-7_47.

[298] Mingda Zhang, Rebecca Hwa, and Adriana Kovashka. "Equal But Not The Same: Understanding the Implicit Relationship Between Persuasive Images and Text". In: *British Machine Vision Conference 2018, BMVC 2018, Northumbria University, Newcastle, UK, September 3-6, 2018*. 2018, p. 8.

[299] Xiangmin Zhang, Michael Cole, and Nicholas Belkin. "Predicting users' domain knowledge from search behaviors". In: *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM. 2011, pp. 1225–1226.

[300] Xinghua Zhang, Bowen Yu, Yubin Wang, Tingwen Liu, Taoyu Su, and Hongbo Xu. "Exploring Modular Task Decomposition in Cross-domain Named Entity Recognition". In: *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*. Ed. by Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai. ACM, 2022, pp. 301–311. DOI: 10.1145/3477495.3531976.

[301] Bin Zhao, Xuelong Li, and Xiaoqiang Lu. "HSA-RNN: Hierarchical Structure-Adaptive RNN for Video Summarization". In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 7405–7414. DOI: 10.1109/CVPR.2018.00773.

[302] Yu-Meng Zhao, Yun Jing, Shuo Gao, and Liu Limin. "News Image-Text Matching With News Knowledge Graph". In: *IEEE Access* 9 (2021), pp. 108017–108027. DOI: 10.1109/ACCESS.2021.3093650. URL: https://doi.org/10.1109/ACCESS.2021.3093650.

[303] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. "PubLayNet: largest dataset ever for document layout analysis". In: *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE. Sept. 2019, pp. 1015–1022. DOI: 10.1109/ICDAR.2019.00166.

[305] Martina Ziefle. "Effects of display resolution on visual performance". In: *Human factors* 40.4 (1998), pp. 554–568.

Christian Otto

Curriculum Vitae

chro@outlook.de
0000-0003-0226-3608

Work Experience

12/2021 – today **Leibniz University Hannover, L3S Research Center**
- Project coordinator: Leibniz AI Academy – Hybrid Micro-Degrees for Academic Training
- Research assistant in the research group "Visual Analytics"

10/2019 – 06/2020 **Continuing education "Promotion Plus"**
- Acquisition of management skills for non-university careers.

09/2016 – 11/2021 **Technische Informationsbibliothek (TIB) Hannover**
- Research assistant in the research group "Visual Analytics"
- Projects
  - TIB AV-Analytics – Development of a software platform for systematic film and video analysis
  - SALIENT – Investigating, Enhancing, and Predicting Learning during Multimodal Web Search
  - Consulting project for content garden technologies GmbH on analysis of image content in advertisements

03/2017– 10/2018 **Leibniz University Hannover, L3S Research Center**
- Research assistant in the research group "Visual Analytics"
- Project: FaAM – Fully-automated Alpha Matting for portrait photography

10/2014– 08/2016 **University of Applied Sciences Jena**
- Research assistant, department Electrical Engineering and Information Technology
- Project: GoVideo: Automatic methods for cost-efficient annotation of documentary film and video content

03/2011– 09/2014 **Fraunhofer IOF Jena**
- Student assistant with focus on software development in the department Imaging and Sensing

Education

04/2012 – 03/2014 **Master of Science in Computer Science**
Friedrich Schiller University Jena
Thesis Title: *Verification of a method for 3D measurement with direct phase deconvolution.*

10/2008 – 03/2012 **Bachelor of Science in Computer Science**
Friedrich Schiller University Jena
Thesis Title: *Software concept of a goniometer measuring station*

09/2000 – 06/2007 **Abitur**
Staatliches Gymnasium "Leuchtenburg" Kahla

Technical Skills

Operating Systems	Linux, Microsoft Windows
Office	Microsoft Office, LibreOffice, GIMP 2.0
Programming	Java, LaTeX, Python, C/C++, MySQL, JavaScript
Python Libraries	Tensorflow, OpenCV, Numpy, PyQt, PyTorch, Pandas, DjangoDB, Scikit-Learn, Flask, ElasticSearch
Other	Vue 2, Docker, Git, Anaconda

Languages

German	Mother tongue
English	Level C1

Research Activities

06/2022	**Invited reviewer for *Artificial Intelligence Review* Journal**
03/2022	**Invited Talk on the 1st International Workshop on *Multimodal Understanding for the Web and Social Media* (MUWS) @ CVPR**, *"Characterization and Classification of Semantic Image-Text Relations"*
03/2022	**Programme Committee of the 1st International Workshop on *Multimodal Understanding for the Web and Social Media* (MUWS)**
02/2022	**Invited reviewer for International Conference on *Artificial Intelligence in Education* (AIED) 2022**
2017 – 2022	**Supervision of 4 master's theses and 6 bachelor's theses**

Honors and Awards

06/2019	**Best Paper Award**, *ACM Internatioal Conference on Multimedia Retrieval (ICMR)*