3rd Conference on Production Systems and Logistics

Modularized Active Learning Solution For Labelling Text Data For Business Environment Analysis

Furkan Agacayaklar[1], Annika Lange[1], Julia-Anne Scholz[1], Thomas Knothe[1], Dirk Busse[2]

[1]*Fraunhofer Institute for Production Systems and Design Technology IPK, Berlin, Germany*
[2]*budatec GmbH, Berlin, Germany*

Abstract

In today's interconnected world, the pace of change is increasing gradually and the effects of an event can propagate and disrupt industries, organizations or companies more dramatically and quickly. Therefore, having a comprehensive overview of the environment is a precious asset for resilience and sustainable growth. One enabler of the above-mentioned interconnectedness is the rapid flow and vast availability of information in text form, which can be also used as the fundamental resource to understand the shifting environment. Hence, actors can be able to become aware of changes at an early stage. The underlying patterns to filter relevant information can be detected by learning from data, or more specifically machine learning. Natural language processing (NLP) techniques can be applied because text data is analyzed. However, to embed the expertise and perspective of the user into the initial model, data should be labeled. This requires valuable expert time from the organization for the labeling, thus it should be minimized. This study aims to present an efficient and user-friendly solution for data labeling. To achieve this, a modularized Active Learning-based backend is combined with an intuitive interface. The output of this labeling process will be used further to train a model for environment analysis. Nevertheless, the main focus of this paper is the development of a solution to maximize efficiency during data labeling for environment analysis. After an introduction to the problem, the overview of the suggested solution accompanied by a prototype will be demonstrated.

Keywords

Business Environment Analysis; Active Learning; Natural Language Processing; Machine Learning; Data Labeling

1. Introduction

A critical challenge for companies is planning future-oriented strategies, especially since today's environment is characterized by transformations [1]. Unpredictable, disruptive events (e.g. the Corona crisis) and dynamic developments are becoming increasingly difficult for companies to assess. Moreover, disruptive influences often put companies in the situation of having to make decisions with long-term effects within a short period of time [2]. Trade barriers, for example, have a direct influence on the supply chain and can force a company to quickly decide on alternative ways to source resources under uncertainty.

Foresight and corporate environment analysis are key to identifying risks and opportunities and ensuring the long-term competitiveness of the company [3]. JAIN [4] identified the connection of environmental influences with the business strategy as an essential component for a future-oriented, agile company. To this end, making trend predictions and differentiating between relevant and irrelevant environmental influences

DOI: https://doi.org/10.15488/12190

are elementary components [4]. However, especially in small and medium-sized enterprises (SMEs), there are various barriers to carrying out environmental analysis, including high personnel and financial costs [5,6]. Moreover, the complexity of dynamic influences requires a solid overview of relevant information for short- and long-term decision-making. For this reason, the Fraunhofer Institute for Production Systems and Design Technology (IPK) has developed a model-based interactive situation awareness monitor and liquidity assessment for enterprise resilience management to provide SMEs in particular with a well-founded and company-specific basis for decision-making, even in times of crisis. In addition to information on the corporate environment, information on suppliers, their production and orders are visualized. The aim of this study is to provide content for a news ticker which is displayed on the situation awareness monitor. [5]

Consideration of context while displaying relevant information from the corporate environment requires analyzing data intelligently instead of providing unfiltered data from certain resources. This could be realized by Machine Learning (ML), more specifically Natural Language Processing [7], if data is in text format. However, this context filter model should be tailored to each organization. The perspective of the organization can be transferred into the ML model by labeling data. One or more individuals from the organization should label a various number of information based on relevancy. However, this task is manual and time-consuming, which is a problem especially when expert time is involved. To minimize the effect of this barrier, an Active Learning (AL) based labeling solution is developed which will be presented in this paper. This labeled data can be used to develop the initial model, which can be utilized for further implementation in the environment analysis pipeline. Nevertheless, the scope of this paper is concentrated on data labeling and Active Learning. Additional steps in the environment analysis pipeline such as training and performance of NLP models are part of further research and therefore not covered in this paper.

The logic behind the approach using text data, supervised ML and Active Learning for environment analysis is addressed in section 2. Next, existing approaches from Active Learning are presented (section 3), followed by the application (section 4) and discussion (section 5). The paper concludes with a summary and outlook (section 6).

2. Problem Description

2.1 Input Data for Environment Analysis

Data can take different forms. A popular depiction of data is in tabular form, which is known as structured data. These types of data can be sorted, searched, modified and analyzed via conventional methods. However, they constitute a small portion of data because most of the data created today are unstructured in the form of text, image, sound, etc. [8]. This also applies to environment analysis, since real-time information coming from the outside world (e.g. news articles) is generally in raw format and not always pre-processed or put into the structured format. Therefore, analysis of unstructured data is key for environment analysis.

Text is an informative and effective form of unstructured data, considering news, tweets and other written media. This is how people consume information. Other forms such as images or videos are also valuable sources, but even those are captioned by text. For example, a picture of a ship lodged on the ground may not have much meaning. However, if this picture is captioned as *"The giant ship is causing a traffic jam in one of the world's busiest waterways"* [9] in coverage of the famous Suez Canal blockage, it has much more context and meaning to a decision-maker. Moreover, text data is relatively lightweight compared to images and video. Therefore, text data from news providers will be considered in this study.

2.2 Requirement for Machine Learning

The relevance of a text can be intuitively evaluated by a human being, but it is not straightforward for a machine to mimic this behavior. The first option to consider is rule-based methods such as regex-based filter or ElasticSearch [10]. The required information can be found by using keywords or search queries. Similarly, certain rules can be established to filter relevant information. However, some problems may arise. First of all, it is a static solution and does not learn from experience. Secondly, some information or themes can be overlooked because in a basic search it is known what to look for, but developing a filter to find relevant information is a broader task. This is where ML comes into consideration. Using ML allows one to inherit complex patterns in the data and develop more capable solutions. Therefore, an ML model will be trained to filter relevant news. The model can be considered as the context filter which is the first step of the pipeline and chained with further models or even rule-based filters for further refinement.

The problem could be described as a text classification problem but it can be approached as both supervised or unsupervised learning. Supervised learning methods learn from labeled data (i.e. input-output pairs) so that new inputs can be mapped to respective outputs [11]. In the case of environment analysis, tailored solutions should be developed for each organization. Thus, to apply supervised learning, data should be labeled based on the preferences of the corresponding organization. This prerequisite for supervised learning is an incentive to consider unsupervised learning.

Unsupervised learning algorithms try to discover patterns in data without labels [11]. For this problem, clustering algorithms [12], which is a sub-field of unsupervised learning, could be used to divide data into groups based on similarities of data. Each cluster can be analyzed and labeled based on relevancy and new data can be evaluated based on the cluster it is mapped to. However, dividing data into ideal, desired and homogeneous clusters is a challenge and tuning of a clustering algorithm requires effort which can be illustrated by varying results of different clustering algorithms on the same dataset [12]. Therefore, supervised learning is the more attractive option despite the requirement of labeling.

2.3 Data Labeling

In the case of environment analysis, one or more people from an organization with sufficient and comprehensive domain knowledge should evaluate the selected text data based on relevancy and label the data. This means a considerable time and effort of experienced employees should be dedicated to such a manual task. Therefore, an effective and efficient choice of data to be labeled is vital. Considering the variety of news providers and the amount of news published daily, the raw pool can be large. Titles and short descriptions of 29,040 articles from various German news providers are acquired as text data for environment analysis. The next step is to select the data to be labeled.

The first and the simplest option is random sampling. The problem with this approach is that it can lead to an unbalanced data set, which means the data could include many articles which will be labeled as irrelevant. This is detrimental for ML model training [13]. However, this phenomenon is highly likely for environment analysis because the articles which are specifically relevant for an organization will be small among all themes in the article pool such as sport, entertainment and technology. Assuming only 1,000 out of 29,040 articles were relevant, on average 35 articles are expected to be relevant in a sample of 1,000. The second option is inspired by how people look up information on the Web in daily life. The information is filtered by search queries or rule-based filters. Then, the desired information is searched within this distilled list of options. Similarly, the articles can be filtered based on one or more words and filtered articles could be labeled. However, that might result in overlooking some articles or themes as mentioned before. Another option is Active Learning [14–16] which is a method used to address the labeling problem in ML and provides informed choices on data for labeling and is elaborated in the following section.

3. State of the Art of Active Learning

Active Learning aims to make informed queries from unlabeled data based on the output of learning algorithms so that labeling of these queries will allow the model to perform better [14,16]. Therefore, target performance can be reached with fewer queries and faster compared to random selection [15,16]. This aspect is valuable for environment analysis data labeling since valuable expert time is minimized by fewer data to label.

The general procedure of AL is provided in Figure 1. It starts with small set of already labeled data which is enough to train a learning algorithm. These initial data can be selected by diversity sampling strategies (will be covered) or randomly and be labeled. This is depicted by the dashed line in the figure. This labeled (initial) data is used to train the model/learning algorithm. Then, the trained algorithm selects (queries) data from the unlabeled (pool) dataset, labeling of which is considered valuable by the algorithm for the goal being sought. The term "*goal*" is mentioned here because it is generally stated as accuracy in literature [16]. However, it can differ based on the type of query strategy, which will be covered. Selected data for labeling is examined by the user or expert, which is named as "*oracle*" in AL literature [14,16]. The oracle labels the data which is transferred from the pool dataset to the initial dataset, which is generally done automatically. Afterwards, the algorithm can be refitted based on the updated version of the labeled dataset, which allows it to make more informed choices. The updated algorithm can query the new data and the cycle repeats itself. In the environment analysis example, the algorithm selects an article from the unlabeled news dataset which is shown to the user. The user only reads the article and labels it as relevant(yes) or irrelevant(no). The labeled article can be removed from the pool dataset and appended to the labeled dataset in the background automatically. The algorithm can be retrained with labeled dataset and can query the next article. Although data can have more than two dimensions, the plot on the right shows an example in 2D just for visualization. The queried data point and labeled data can be observed.

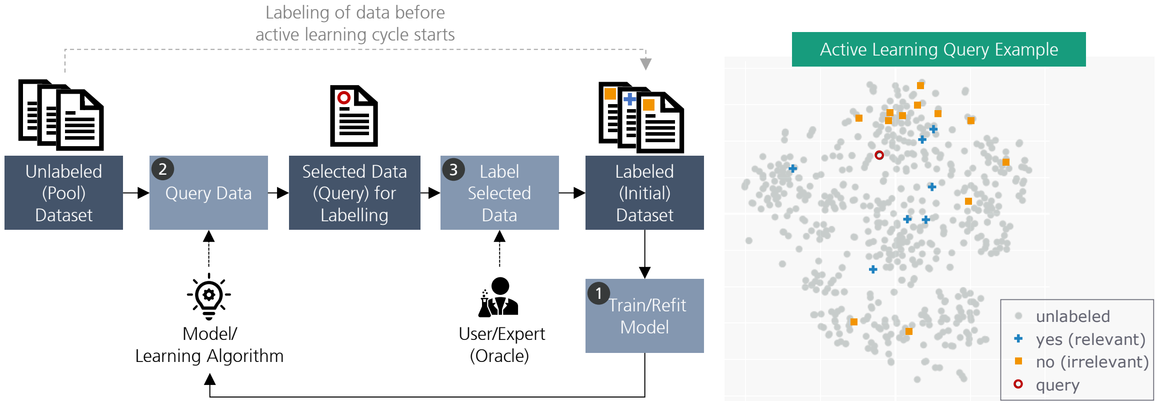


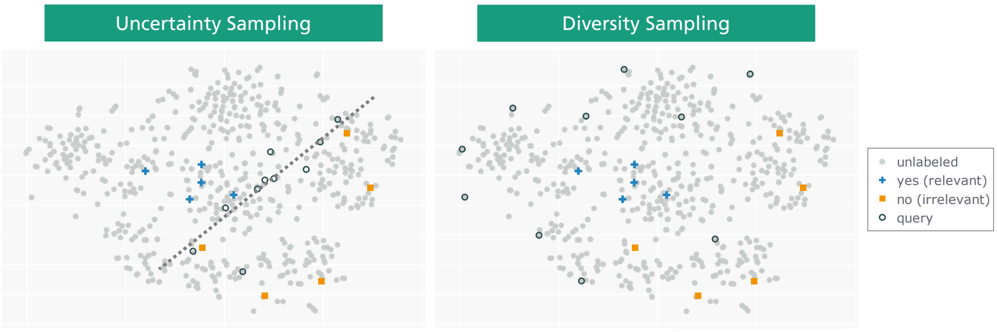
Figure 1: Generic flow of Active Learning and its representation on 2D scatter plot. Data were taken from the BBC News dataset [17]. Vectorized articles are projected on 2D via t-SNE [18] just for visualization purposes.

The goal, which is tried to be achieved by each query, can differ as mentioned before and AL query strategies can be classified based on this aspect. SETTLES [16] provides a good survey on query strategies, but they are not grouped. However, according to KUMAR, GUPTA [14], query strategies can be classified as informative-based and representative-based for classification problems. MUNRO [15] calls them uncertainty sampling and diversity sampling respectively, although uncertainty sampling is considered as a sub-category of informative-based strategies [14]. Nevertheless, informative-based strategies also focus on uncertainty, so underlying aims and logics are similar [14]. In this article, uncertainty sampling and diversity sampling are used for classification.

The knowledge quadrant by MUNRO [15] can be used to clarify the goals of strategies. Every data point belongs to one of the quadrants at a state which can be explained in the context of environment analysis. The first quadrant focuses on "*Known-knowns*", i.e. ML model knows that it can predict those data points (e.g. news articles) with high accuracy. This represents the "*current model state*". The points in the "*Unknown-knowns*" category are not correctly predicted by the ML model although the model is sure about it. This is bound to model performance and can be improved by "*transfer learning*" or a better model design. The third category is “*Known-Unknown*”. The ML model knows that these data points are confusing. For instance, the model cannot confidently tell whether these news articles are relevant or irrelevant. These data points are close to the decision boundaries of the model. Such data points are queried by “*uncertainty sampling*”. “*Unknown-unknown*” data is beyond the model’s knowledge since they are not labeled yet and could change the perspective of the model when labeled. This is caused by lacking diversity of data and can be tackled by “*diversity sampling*”. The model thinks it is sure about such news but has not got a labeled sample from this theme which could lead to false prediction.

3.1 Uncertainty Sampling

Classification models try to draw a decision boundary between labeled data points, which allows them to distinguish and assign these data points as shown in Figure 2.a. The data points are again projected to 2D via TSNE [18] just for better illustration. In the Figure, the model is trained based on labeled data (yes/no) and the decision boundary is illustrated by the hypothetical (manually drawn) line. In this example case, the model predicts points on the left and right of the boundary as relevant (yes) and irrelevant (no) respectively. The model tends to predict the points away from the decision boundary confidently. However, the points closer to the decision boundary are risky for the model to predict. Uncertainty sampling aims to query such points. In the below example, the points queried by the uncertainty sampling strategy are in the region of the boundary as expected. Uncertainty query strategies try to select data for labeling which could contribute to the model performance the most [16]. Thus, after labeling each query, the model can distinguish categories better which improves accuracy. For instance, the environment context model becomes better in distinguishing the relevancy of known themes after each uncertainty sampling query. However, what if the labeled data only consists of certain topics? Then the model assumes only these topics exist and tries to classify only them. Therefore, the variety of data is vital, which is achieved via diversity sampling.



(a) (b)

Figure 2: Queries taken via simulation based on true labels starting from the same state for ten iterations with different strategies (a) Uncertainty Sampling vs. (b) Diversity Sampling. A hypothetical decision boundary is manually drawn for uncertainty sampling to emphasize pattern. Data were taken from a BBC News dataset [17].

3.2 Diversity Sampling

Diversity sampling aims to query representative data points from the unlabeled data set [14] . The target is to solve the problem mentioned above. Uncertainty sampling without diversity sampling would focus more on the insights provided by the currently labeled data, and if it is not representative, the model will have gaps [15]. This phenomenon is essential for environment analysis since the expected result is to provide comprehensive information from the environment to the user as much as possible. If certain topics are overlooked and not labeled, the model would be trained without this knowledge and make inaccurate predictions. For example, if the labeled data contains only news about electronics, the uncertainty sampling tries to clarify the decision boundary for this topic (see Figure 2.a). In this case, topics such as Covid-19 or logistics which might be further away from electronics articles in vector space will be ignored. This will result in an incomplete recommendation from the context model. Diversity sampling helps to overcome this issue as shown in Figure 2.b. Nevertheless, diversity sampling will not focus on the data which might be predicted incorrectly by the model and this results in a less marginal positive impact on model performance compared to uncertainty sampling.

3.3 Hybrid and Advanced Strategies

The trade-off between diversity and uncertainty sampling strategies can be addressed by hybrid and combined models. Hybrid strategies merge both diversity and uncertainty into a single strategy [14]. Querying Informative and Representative Examples (QUIRE) [19], Learning Active Learning (LAL) [20], Self-Paced Active Learning (SPAL) [21] are some applications. However, these methods try to solve the problem end-to-end via complex metrics [14] which hinders interpretability for use cases and algorithms may become slower. This causes latency and more expert time for labeling. To address this issue, MUNRO [15] suggests advanced strategies which try to combine the advantages of two strategies more explicitly by applying them sequentially. For example, uncertainty sampling could be used to query some data points which could be further filtered by diversity sampling. However, this involves narrowing down samples via simple strategies one by one.

◔ Low

◑ Below Average

◕ Above Average

● High

		Query Strategies								
		Uncertainty Sampling		Diversity Sampling		Hybrid Strategies			Combined Strategies	
		Least Confidence / Margin of Confidence / Entropy [14–16]	Query-By-Committee [14,16]	Density-Based/ Representative Sampling [14,16]	Cluster-based Sampling [14,15]	QUIRE [19,22]	LAL [20,22]	SPAL [21,22]	Least Confidence with Cluster-based Sampling [15]	Representative Cluster-based Sampling [15]
Criteria	Variety of samples for complete topic coverage	◔	◔	◕	◕	◕	◑	◕	◑	●
	Contribution to model performance for better distinction of relevancy	●	●	◔	◔	◕	◕	◑	◕	◔
	Interpretability of metric used during query to check use case compatibility	●	◕	●	◕	◔	◔	◔	◕	◕
	Speed of queries to reduce latency and expert time spent	●	◕	●	◕	◔	◔	◑	◕	◕

Figure 3: Comparison of existing strategies based on designated criteria for environment analysis

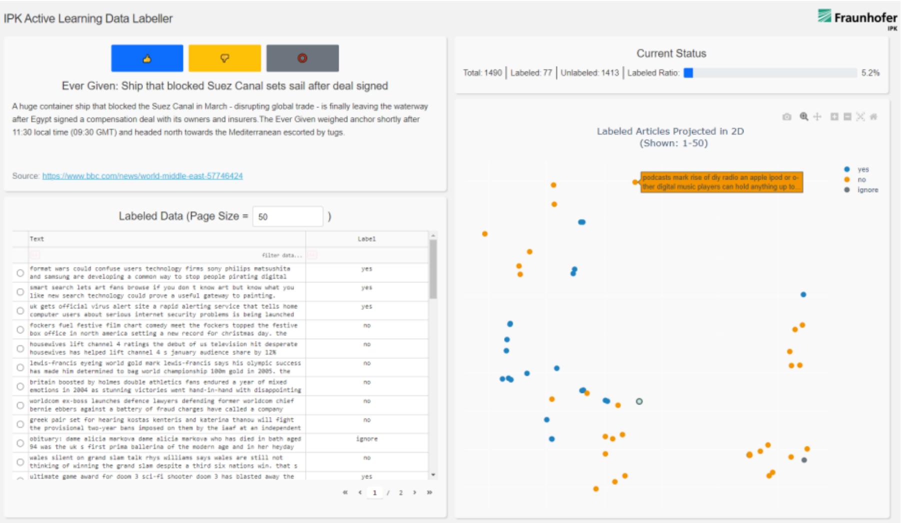
Figure 3 provides the evaluation of strategies from each type based on the criteria essential for the environment analysis labeling procedure. It can be seen that combining strategies can supplement diversity and uncertainty simultaneously at the cost of speed and simplicity.

4. Application

AL query strategy categories for classification problems and the trade-off between them are explained in the previous section. The context model for environment analysis solves a text classification problem and AL is a suitable method because oracle (i.e. expert, user) can label the data intuitively. However, expert time for labeling should be minimized and the labeled dataset should be adequately diverse and informative for the model. Moreover, the labeling experience for the user should be smooth. Considering all these aspects, a web-based prototype for AL with a modular backend for different strategies is developed and actively used for labeling by the industry partners. Python programming language is used for development.

4.1 User Interface

AL queries data based on the current state of the initial (labeled) and pool (unlabeled) dataset and learning algorithm for data selection should be regularly updated after each iteration of labeling, especially for uncertainty sampling. Therefore, it is not possible to give a static datasheet to a user to label data. There are AL strategies that enable batch sampling. These are especially valuable if the query strategy is time-consuming and computationally heavy, but they are likely to suffer from adaptivity and redundancy [14]. Therefore, they are not preferred especially for uncertainty sampling.



IPK Active Learning Data Labeller

Ever Given: Ship that blocked Suez Canal sets sail after deal signed

A huge container ship that blocked the Suez Canal in March - disrupting global trade - is finally leaving the waterway after Egypt signed a compensation deal with its owners and insurers.The Ever Given weighed anchor shortly after 11:30 local time (09:30 GMT) and headed north towards the Mediterranean escorted by tugs.

Source: https://www.bbc.com/news/world-middle-east-57746424

Labeled Data (Page Size = 50)

	Text	Label
	filter data...	
○	format wars could confuse users technology firms sony philips matsushita and samsung are developing a common way to stop people pirating digital	yes
○	smart search lets art fans browse if you don t know art but know what you like new search technology could prove a useful gateway to painting.	yes
○	uk gets official virus alert site a rapid alerting service that tells home computer users about serious internet security problems is being launched	yes
○	fockers fuel festive film chart comedy meet the fockers topped the festive box office in north america setting a new record for christmas day. the	no
○	housewives lift channel 4 ratings the debut of us television hit desperate housewives has helped lift channel 4 s january audience share by 12%	no
○	lewis-francis eyeing world gold mark lewis-francis says his olympic success has made him determined to bag world championship 100m gold in 2005. the	no
○	britain boosted by holmes double athletics fans endured a year of mixed emotions in 2004 as stunning victories went hand-in-hand with disappointing	no
○	worldcom ex-boss launches defence lawyers defending former worldcom chief bernie ebbers against a battery of fraud charges have called a company	no
○	greek pair set for hearing kostas kenteris and katerina thanou will fight the provisional two-year bans imposed on them by the iaaf at an independent	no
○	obituary: dame alicia markova dame alicia markova who has died in bath aged 94 was the uk s first prima ballerina of the modern age and in her heyday	ignore
○	wales silent on grand slam talk rhys williams says wales are still not thinking of winning the grand slam despite a third six nations win. that s	no
○	ultimate game award for doom 3 sci-fi shooter doom 3 has blasted away the	yes

« ‹ 1 / 2 › »

Current Status

Total: 1490 | Labeled: 77 | Unlabeled: 1413 | Labeled Ratio: 5.2%

Labeled Articles Projected in 2D
(Shown: 1-50)

Fraunhofer
IPK
podcasts mark rise of diy radio an apple ipod or other digital music players can hold anything up to...
yes
no
ignore

Figure 4: Active Learning Interface

A web-based user interface is developed to provide a smooth and adaptive labeling experience (Figure 4). The learning algorithms and strategies are updated in the backend when necessary while displaying selected queries in the front end. The current news headline and its description are shown on the top left with three buttons for labeling: relevant (blue), not relevant (yellow), ignore/not sure (gray). Ignore option is included to reduce noisy labels from users when they are not confident and they are simply ignored. Such data points are no longer queried for labeling and are not used to train the model since the user is not sure about them. On the bottom left, a paginated list of labeled data (history) can be observed, filtered, selected and modified. The articles on the list are displayed by an accompanying scatter plot where they are projected on the 2D via dimensionality reduction (TSNE[18]) and colored in line with labeling buttons. The models are trained on high dimensional vectors, but reduced 2D coordinates are shown just for interactivity. Moreover, these compressed/reduced 2D vectors still contain patterns in original data, so users can see forming clusters. On the top right, the current status can be observed. This simple interface allows users to label the data quickly while getting adaptive queries from the backend as described in Figure 1.

4.2 Modular Backend: Chaining Strategies

The aim of the modular backend is plain. The query strategies are chained one by one in order. Active strategy selects instances and the user labels them. After the designated number of queries are made with the strategy, the next strategy is initiated. The idea can be described as follows. The chain comprises diversity and uncertainty sampling methods interchangeably. Diversity sampling tries to fill the gaps in the information available to the model by selecting representative instances based on the current state. This is expected to confuse the model because new topics or themes will be introduced. In other words, the decision boundaries in some areas of feature space become vague. In the next step, uncertainty sampling strategies try to reduce disruption caused by increased diversity. This is done by selecting points to improve the model's performance. This exchange between strategies is repeated as much as necessary. Active Learning strategies try to maximize marginal contribution on accuracy as covered in the literature on individual strategies [23,19,24–26]. The exchange between diversity and uncertainty sampling improves coverage on topics in the dataset and distinction capability on these topics respectively.

The modular structure is developed based on the modAL (Modular Active Learning) framework [27]. It contains some strategies by default, but it also provides plug-and-play option for custom-made query strategies. Strategies such as basic uncertainty sampling, margin sampling, entropy sampling and query by committee are inherited from default functions [27]. Nevertheless, strategies such as cluster-based AL [15], density-based sampling [14] or advanced strategies like uncertainty-representative sampling [15,16] are developed and implemented. Moreover, these strategies are optimized for low latency sampling. Custom and tailored methods are also developed. One strategy tries to boost the number of positively labeled data (relevant) to reduce unbalance. Another strategy aims to increase diversity based on Radial Basis Function (RBF) Kernel [28] based similarity. They are also implemented as a strategy in the modular chain.

Order, repetition and parameters for strategies can be set by a config file. If there is no initial data, the chain starts with a cluster-based or kernel-based strategy since they do not require a pre-trained model on labeled data to query. Moreover, the query with this initial strategy continues until the minimum required amount of data from each category (relevant-not relevant) is obtained. Then, the following strategies in the chain are activated in order. The latency is reduced since algorithms are optimized for speed or selected accordingly. Preprocessing of pool (unlabeled) data including vectorization, clustering and dimensionality reduction is also done beforehand to prevent unnecessary computations during deployment. The combination of these properties, chained strategies and interactive interface provides a seamless experience for the user to label the informed selection of samples.

5. Discussion

The modular structure uses individual strategies, which are not novel by themselves and most of them are implemented based on existing methods. The main objective or contribution of this presented solution is to utilize the labeling effort of the user most efficiently. A context model will be the output of labeled data. This context model should mimic the perspective and mindset of the users as much as possible, to filter new information based on their preferences.

The current AL solution by nature provides intelligence to the labeling process. However, manual chaining of strategies via a config file can be improved. The goal sought by each strategy can be defined. Then, the current state during the labeling can be monitored via metrics or algorithms. These metrics could be used to trigger the transition between strategies. For example, if model performance is no longer improved by the current iteration of uncertainty sampling, diversity sampling can be initiated. The other aspect to consider is the static pool of data used during the labeling process, even if the data is up-to-date at that time. The context model will be trained based on this data. Therefore, methods should be implemented to recommend new topics to the user and learn their preferences toward them. This can be achieved by integrating trend detection methods [29] or tools [30] in the pipeline.

6. Summary and outlook

Awareness and vigilance are valuable in today's turbulent business environment. The utilization of big data could help to reap this benefit. Environment analysis tries to find the relevant information for an organization from the outside world and present it in a structured manner. This paper presents an approach to develop the first step of the pipeline for this analysis. The necessity of labeled data is explained, and the application of Active Learning is shown to satisfy this prerequisite efficiently and seamlessly. The goal is to minimize the expert time spent and amount of data labeled as much as possible to have a comprehensive understanding of which information the corresponding organization deems worthy. The user interface and modular backend consisting of chained query strategies, some of which are optimized and tailored for the problem, enable the goal. The overcoming of limitations originated from lack of intelligent transition between query strategies and integration of new trends and topics into the finalized system should be considered. The next step is to filter and analyze the data further based on the integrated enterprise model (IEM) [31] and the data of enterprise IT systems from an industry partner. Finally, relevant and tailored information for the organization will be displayed on the situation awareness monitor [5].

Acknowledgements

This research and development project is funded by the German Federal Ministry of Education and Research (BMBF) within the "Innovations for Tomorrow's Production, Services, and Work" Program (02P18X000) and implemented by the Project Management Agency Karlsruhe (PTKA). The author is responsible for the contents of this publication.

References

[1] Vecchiato, R., 2015. Creating value through foresight: First mover advantages and strategic agility. Technological Forecasting and Social Change 101, 25–36.

[2] Rust, H., 2021. Weise Voraussicht und Erfolgsplanung. Springer Fachmedien Wiesbaden, Wiesbaden, 210 pp.

[3] Dadkhah, S., Bayat, R., Fazli, S., Tork, E.K., Ebrahimi, A., 2018. Corporate foresight: developing a process model. Eur J Futures Res 6 (1), 1–10.

[4] Jain, S., 1984. Environmental Scanning in U.S. Corporations. Long Range Planning (Vol. 17, No. 2), 117–128.

[5] Kohl, H., Knothe, T., Oertwig, N., Gering, P., Scholz, J.-A., 2021. Interaktives Lagebild: Ein Werkzeug für das Krisenmanagement in prozessorientieren Unternehmen. Industrie 4.0 Management 37 (1), 37–40.

[6] Will, M., 2008. Talking about the future within an SME?: Corporate foresight and the potential contributions to sustainable development. Management of Env Quality 19 (2), 234–242.

[7] IBM Cloud Education, 2020. Natural Language Processing (NLP). https://www.ibm.com/cloud/learn/natural-language-processing. Accessed 15 March 2022.

[8] IBM Cloud Education, 2021. Structured vs. Unstructured Data: What's the Difference? https://www.ibm.com/cloud/blog/structured-vs-unstructured-data. Accessed 5 January 2022.

[9] BBC, 2021. Egypt's Suez Canal blocked by huge container ship. https://www.bbc.com/news/world-middle-east-56505413. Accessed 5 January 2022.

[10]Elasticsearch B.V., 2021. What is Elasticsearch? | Elasticsearch Guide [7.16] | Elastic. https://www.elastic.co/guide/en/elasticsearch/reference/current/elasticsearch-intro.html. Accessed 5 January 2022.

[11]Delua, J., 2021. Supervised vs. Unsupervised Learning: What's the Difference?, March 12.

[12]scikit-learn developers, 2022. 2.3. Clustering. https://scikit-learn.org/stable/modules/clustering.html. Accessed 5 January 2022.

[13]Salas-Eljatib, C., Fuentes-Ramirez, A., Gregoire, T.G., Altamirano, A., Yaitul, V., 2018. A study on the effects of unbalanced data when fitting logistic regression models in ecology. Ecological Indicators 85, 502–508.

[14]Kumar, P., Gupta, A., 2020. Active Learning Query Strategies for Classification, Regression, and Clustering: A Survey. Journal of Computer Science and Technology 35 (4), 913–945.

[15]Munro, R., 2021. Human-in-the-Loop Machine Learning, 1st edition ed. Manning Publications; Safari, Erscheinungsort nicht ermittelbar, Boston, MA, 424 pp.

[16]Settles, B. Active Learning Literature Survey. University of Wisconsin-Madison Department of Computer Sciences. https://minds.wisconsin.edu/handle/1793/60660.

[17]Greene, D., Cunningham, P., 2006. Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering, in: Proc. 23rd International Conference on Machine learning (ICML'06). ACM Press, pp. 377–384.

[18]scikit-learn developers, 2022. TSNE. https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html. Accessed 14 March 2022.

[19]Huang, S.-J., Jin, R., Zhou, Z.-H., 2014. Active learning by querying informative and representative examples. IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (10), 1936–1949.

[20]Konyushkova, K., Sznitman, R., Fua, P., 2017. Learning Active Learning from Data. Advances in Neural Information Processing Systems 30.

[21]Tang, Y.-P., Huang, S.-J., 2019. Self-Paced Active Learning: Query the Right Thing at the Right Time. AAAI 33, 5117–5124.

[22]Tang, Y.-P., Li, G.-X., Huang, S.-J., 2019. ALiPy: Active Learning in Python. https://arxiv.org/pdf/1901.03802.

[23]Du, B., Wang, Z., Zhang, L., Zhang, L., Liu, W., Shen, J., Tao, D., 2015. Exploring representativeness and informativeness for active learning. IEEE transactions on cybernetics 47 (1), 14–26.

[24]Krempl, G., Kottke, D., Spiliopoulou, M. Probabilistic active learning: Towards combining versatility, optimality and efficiency, in: International Conference on Discovery Science. Springer, pp. 168–179.

[25]Lewis, D.D., Gale, W.A., 1994. A Sequential Algorithm for Training Text Classifiers. https://arxiv.org/pdf/cmp-lg/9407020.

[26]Nguyen, V.-L., Shaker, M.H., Hüllermeier, E., 2022. How to measure uncertainty in uncertainty sampling for active learning. Mach Learn 111 (1), 89–122.

[27]Danka, T., Horvath, P., 2018. modAL: A modular active learning framework for Python, 5 pp. http://arxiv.org/pdf/1805.00979v2.

[28]scikit-learn developers, 2022. RBF. https://scikit-learn.org/stable/modules/generated/sklearn.gaussian_process.kernels.RBF.html. Accessed 15 March 2022.

[29]Kontostathis, A., Galitsky, L.M., Pottenger, W.M., Roy, S., Phelps, D.J., 2004. A Survey of Emerging Trend Detection in Textual Data Mining, in: Berry, M.W. (Ed.), Survey of Text Mining. Clustering, Classification, and Retrieval. Springer, New York, NY, pp. 185–224.

[30]Synthesio, 2013. Trend Analysis Tools | Synthesio. https://www.synthesio.com/glossary/trend-analysis-tools/. Accessed 5 January 2022.

[31]Spur, G., Mertins, K., Jochem, R., 1993. Integrierte Unternehmensmodellierung, 1. Aufl. ed. Beuth, Berlin.

Biography

Furkan Agacayaklar (*1993) is a research associate in the business process and factory management department at the Fraunhofer Institute for Production Systems and Design Technology (IPK) Berlin since 2021 with focus on data applications. He has previously studied Mechanical Engineering (B.Sc.) at Bogazici University and Management and Engineering in Production Systems (M.Sc.) at RWTH Aachen.

Annika Lange (*1996) is a research associate in the business process and factory management department at the Fraunhofer Institute for Production Systems and Design Technology (IPK) Berlin since 2021. She has previously studied Sports Engineering at Otto-von-Guericke University Magdeburg (B.Sc.) with an exchange at the University college of southeast Norway and Mechanical Engineering at Leibniz University Hannover (M.Sc.).

Julia-Anne Scholz (*1992) is a research associate at the Fraunhofer Institute for Production Systems and Design Technology (IPK) and has been developing solutions for agile process management and resilient business alignment. She has previously studied Industrial Engineering at the Technical University of Berlin with an exchange at the Beijing Tsinghua University.

Prof. Dr.-Ing. Thomas Knothe (*1971) has been head of the Business Process and Factory Management Department at the Fraunhofer Institute for Production Systems and Design Technology (IPK) Berlin since 2010. He has previously studied Information Technology in Mechanical Engineering at the Technical University of Berlin and completed his doctorate in engineering.

Dirk Busse (*1969) is founder, shareholder and managing director of budatec GmbH since 2009. He previously made an apprenticeship as Precision Mechanic and studied Mechanical Engineering at the Technical University Berlin (Dipl-Ing.).