Visual Interactive Labeling of Large Multimedia News Corpora

Qi Han, Markus John, Kuno Kurzhals, Johannes Messner and Thomas Ertl

Institute for Visualization and Interactive Systems, University of Stuttgart, Stuttgart, Germany
E-mail: firstname.lastname@vis.uni-stuttgart.de

***Abstract*—The semantic annotation of large multimedia corpora is essential for numerous tasks. Be it for the training of classification algorithms, efficient content retrieval, or for analytical reasoning, appropriate labels are often the first necessity before automatic processing becomes efficient. However, manual labeling of large datasets is time-consuming and tedious. Hence, we present a new visual approach for labeling and retrieval of reports in multimedia news corpora. It combines automatic classifier training based on caption text from news reports with human interpretation to ease the annotation process. In our approach, users can initialize labels with keyword queries and iteratively annotate examples to train a classifier. The proposed visualization displays representative results in an overview that allows to follow different annotation strategies (e.g., active learning) and assess the quality of the classifier. Based on a usage scenario, we demonstrate the successful application of our approach. Therein, users label several topics which interest them and retrieve related documents with high confidence from three years of news reports.**

***Index Terms*—Visual Analytics, Classification, Visual Interactive Labeling, Multimedia Analysis, Multimodal Learning, Visual Active Learning**

I. Introduction

Nowadays, news agencies and broadcasting services publish their news reports digitally, providing large databases of multimedia content consisting of video supported by closed captions. Private users upload countless videos to online portals, which in many cases are enriched by manually edited or automatically generated subtitles. The vast amount of data in such collections contains valuable information and knowledge for a wide range of new applications. For example, students can retrieve multimedia material related to the topics which they are interested in to learn [1]. That way, educational institutions can analyze the content and the quality of their provided courses. Researchers and journalists can study the dynamics of specific news events and compare the coverage of the events by different news agencies through analyzing multiple news corpora [2]. Over the years, many approaches for retrieving and analyzing multimedia data have been introduced to fulfill those needs. In general, textual data from subtitles provides richer semantical information for retrieval tasks than an analysis solely based on computer vision. Hence, in this work, we focus on large corpora of video data with additional textual data, such as automatically generated subtitles.

LEVIA'18: Leipzig Symposium on Visualization in Applications 2018

The semantic annotation of these archives plays an essential role in many multimedia analysis tasks: A human annotator assigns one or multiple class labels to data instances in a corpus. Typically, those classes represent high-level semantic categories which human users can interpret easily. The labeled data instances can be leveraged to train a classifier using machine learning methods [3], which can analyze and classify a large amount of similar unseen instances in a short period of time. Once all data instances in the corpus are annotated, they can be used as meta-data in various ways for analysis and retrieval tasks. The annotations can also help analysts to group and aggregate the data into fewer classes, for analysis on a higher level of abstraction. For example, all reports about a specific conflict could be summarized to provide an overview of the temporal development of this topic and compare it to other conflicts. Last but not least, annotations can be leveraged to enable retrieval queries based on high-level semantics [4]. However, the manual labeling of large multimedia datasets is time-consuming and tedious because videos have to be investigated individually, often accompanied by annotators having to skim through the video. Hence, several approaches have been proposed to reduce the burden of data labeling and to accelerate the process.

One research direction considers visualization-based approaches to reduce the effort of manual labeling by making it more intuitive, transparent, and integrated into the analysis process. For example, ChronoVis [5] and ANVIL [6] aid annotation and analysis of multimodal data by providing synchronized visualizations views of annotation and the data. Hagedorn et al. [7] introduce approaches to support the optimization of an annotation workflow by providing adequate representation and interaction possibilities. Kurzhals et al. [8] show that their visual approach is more efficient for the annotation of eye tracking videos than a traditional one. Van der Corput and Van Wijk [9] propose to integrate categorization into the analysis of large image collections by treating all meta-data as multivariate data. Kurzhals et al. [10] also suggest an approach for analyzing movies with video and textual data. Our work extends on these ideas by further integrating an automatic classifier to suggest labels and reduce the number of manual labels needed. Uncertainty visualization is introduced to help users by assessing the quality of the suggested labels.

Active learning [11] has been proposed to reduce the number of labeled instances needed to build a precise classifier which can label a dataset automatically. Although active learning has

shown success in many applications, some limitations have hindered a more comprehensive adaptation of it. Specifically, the optimal strategy of choosing an instance to label manually depends on the structure of the data, the type of classifier, and the already labeled data [12]. Thus, several visualization approaches have been proposed to improve the effectiveness of active learning by integrating the users' knowledge. Visualizations allow users to obtain an overview of the dataset quickly, to see the progress of the annotation process, and to interpret the classification results using their domain and world knowledge [13], [14]. Altogether, visualization can make the annotation process more efficient and the analysis more effective. For example, the *Visual Classifier* [15] allows users to interactively label documents for recall-oriented retrieval. Han et al. [16] propose a visual approach which allows users to visually inspect, correct, and accept documents labels proposed by classifiers trained using very few initially labeled data instances. Those newly labeled instances can be used to retrain the classifier so that the classifier accuracy is iteratively improved and more labeled data is available. Liu et al. [17] propose an interactive visual approach for discovery and retrieval of interesting data from microblogs. The visualization enables users to find and refine the most uncertain results effectively.

Bernard et al. [18] present a unified framework for combining visual analytics and active learning that is called *visual-interactive labeling (VIAL)*. We propose an approach based on VIAL for large multimedia news corpora with visual active learning. We allow users to initiate labels via keywords queries. The approach depicts the intermediate results of the classifier with uncertainty information in a sorted table visualization and an overview visualization inspired by treemaps [19]. Users can assess the proposed annotation and refine the labeling with different strategies. On demand, they can obtain detailed information for a single video instance via a detail view. Users can decide if they are satisfied with the current results of annotation and want to stop the labeling process using the visualizations. Overall, the approach allows effective retrieval and analysis of large news corpora by combining automatic classification and human interpretation via a flexible visual interface.

The primary contributions of this work are thus threefold:

1) We introduce a mixed-initiative visual interactive approach that combines efforts from users and a classifier for efficient semantic annotation of a large multimedia corpus.
2) The instances are shown in a grid layout. Users can use multiple strategies to rank, select and label the most informative instances for the classifier to speedup the labeling. (a) They can choose and annotate instances of different classifier confidence levels. (b) They can identify and correct errors of the classifier. (c) Or, they can use keyword queries to retrieve and label the most representative instances for a class. Those strategies can potentially provide more information than labeling most uncertain instances alone, which is the most commonly used active learning strategy.
3) We present a usage scenario to demonstrate the applicability and effectiveness of the proposed approach.

II. Approach

In the following, we first describe the dataset that our approach expected and the applied text preprocessing steps (see Figure 1). Last, we introduce our visual approach and explain the different views and possibilities in detail.

A. Dataset and Data Preprocessing

As the starting point, we assume that the news broadcasts have been segmented into coherent video segments and for each segment, we can obtain its subtitle manually or automatically. If additional meta-data, such as publication date or title of the video is available, our approach can also leverage those data, for example, to provide an additional method for sorting the videos. In this work, we call each video segment together with its subtitle and meta-data as one *report*. After the dataset is loaded into our system, the subtitles of the reports are processed in a linguistic analysis pipeline using the Stanford CoreNLP toolkit [20]. The pipeline consists of tokenization, sentence splitting, and lemmatization [21]. Based on the lemmas, we can define a vector for each report. Afterwards, we use the "term frequency - inverse document frequency" (*tf-idf*) weighting scheme to assign a value to each lemma that reflects how representative the lemma is to the report.

B. Visual Approach

The main workspace of our approach is depicted in Figure 2. It consists of different panels with various visualizations and interactive possibilities that we will introduce in the following. The control panel (see Figure 2 (a)) at the top includes several control options and a search box. At the left most part of the panel is the search box that enables users to search for reports using one or multiple keywords. It allows users to initial a labeling process by finding reports related to a certain keyword or concept. The search method is based on the Lucene library [23] and returns the reports ordered by the *tf-idf* measure by default. Using the buttons on the right side of the panel, users can create new labels or retrain classifiers. For each new label, users assign a name and a certain color for it. The created categories are listed at the bottom of the panel with the assigned color. The "Train & Predict" method allows users to train a classifier once the users have labeled some data. Once the classifier is trained, it will also classify the unlabeled reports accordingly. We will explain how the classifier works later in this section. On the left side is the filter panel (see Figure 2 (a)) that enables users to customize the shown search results. For example, users can sort the result set by date or similarity or they can select only reports about their confidence (high, middle, or low) to a specific label. That way, we enable users to inspect the classification results of different levels of uncertainty. For example, they can inspect only low confidence classifications, which are good candidates for manual inspection, since these classifications have probably

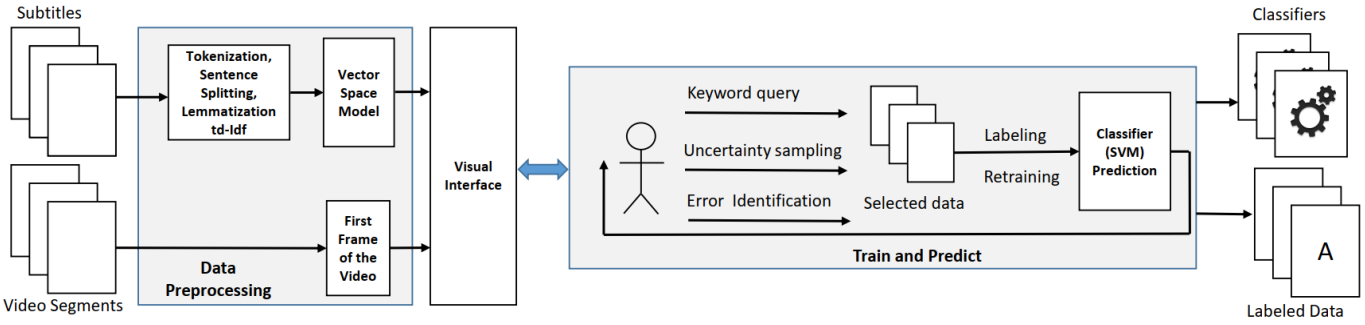

Fig. 1: The workflow of our approach comprises the data preprocessing steps and the "Train & Predict" process.

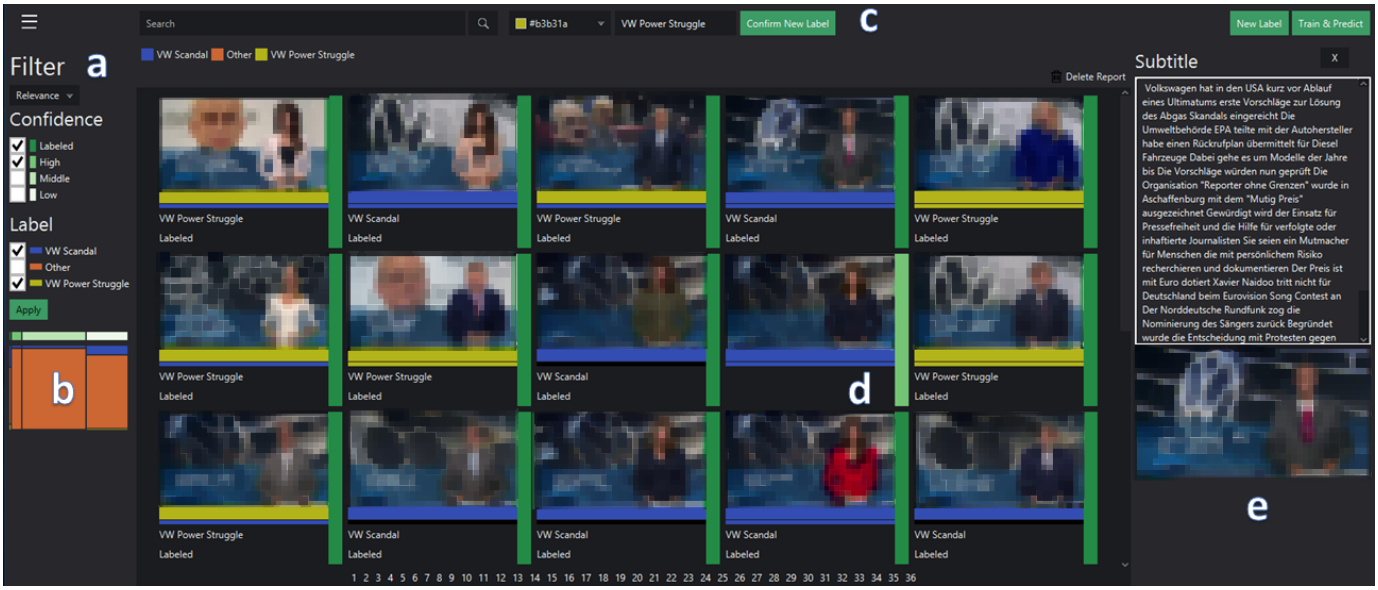

Fig. 2: Our approach consists of different views and interactive possibilities that support the iterative annotation of reports in order to train a classifier. (video source: ARD Mediathek [22])

a high impact on a new prediction. Hence, the quality of the resulting classifier can be improved with fewer labeling actions.

Furthermore, we provide an overview visualization that represents the overall classification progress for the defined labels (see Figure 2 (b)). The visualization consists of two parts: The stacked bars on the top depict the number of reports that are labeled by the users (dark green) or by the classifier with a high (green), middle (light-green), and low (white) confidence. Below those bars is a matrix representation of the classified reports. It represents the created labels with the different assigned confidence values. The columns represent whether the reports are labeled by the users or the classifier with different confidence levels. The rows of the matrix represent different labels filled with the color of the labels. The width and length of the cells represent the number of reports in the respective groupings. This way, users can quickly get an overview of the relative portions of the different labels with the confidence level in the whole dataset. For example, users could find a label that is over represented in the high confidence area. This would suggest some bias in the classifier that could be correct by labeling. By correcting the wrongly assigned reports, the classifier can be greatly improved. Additionally, users can click on the different cells to focus on the group of reports in the content panel (see Figure 2 (c)).

The detail view provides information about each of the reports as depicted in Figure 2 (d). We display the start frame of the reports, which is usually very meaningful. Next to the image, a bar chart represents the confidence value of the classifier. If the reports have not yet been assigned a label by the users or the classifier, the bar chart has a black color. Two bars under the thumbnails show the current label of the report and the label from the previous round. With those two bars, users can quickly identify reports which have changed their labels recently which could be errors. By right-clicking on a report, users can assign an existing label or create a new one in a pop-up menu. Additionally, users can click on a report to open the subtitles right next to the detail view. Double-click on a report will replay the corresponding video file in a separate window.

Based on the labels and the user annotated reports, users can activate the "Train & Predict" method (see Figure 1). This method uses a Support-Vector Machine (SVM) [24] method with a linear kernel [25]. We choose the linear kernel because the training and the prediction are fast and it works well for textual data [25]. The training of the SVM model takes the *tf-idf* vectors of the reports and the assigned labels as input. Once a model is trained, the SVM predicts a value to the different labels for each unlabeled report. The higher the value to the label, the more certain the classification. Whereas, if the classifier returns a negative value to a label, it is a very uncertain assignment.

III. Usage Scenario

In this section, we demonstrate the analytical ability of our proposed approach with a concrete usage scenario. We firstly introduce the used dataset and afterwards, we describe the usage scenario in detail.

A. Dataset

In this paper, we used the German news broadcast station "ARD" episodes from the last three years [22]. Each day, an episode appears that lasts 15 minutes and covers different topics such as sport, weather, or politic news. Since there is no information when a topic changes in an episode, we developed an approach to segment the episodes in a previous student thesis. The approach is based on image recognition and text processing, which, because of the available space, is not focus in this paper. The segmented data are stored in a csv file and include date, title, frame number, textual subtitle, and the start and end frame.

B. The Volkswagen Emissions Scandal

In this usage scenario, we describe how our approach facilitates a journalist to investigate the Volkswagen (VW) emissions scandal happened in 2015. A young journalist Alice wants to write a report on the influence of the automobile industry on the environment. She thinks that the VW Scandal, where VW car reported fake emission values, will be a perfect opportunity to investigate the dynamics of such environmental crisis events.

She launches our system and loads the "ARD" corpus, which we described above. As there are too many reports in the dataset, she can not read each one of them. Therefore, she uses the filter panel to filter all related reports to the VW Scandal. To begin with, she creates two labels: "VW Scandal" and "Other". She searches for reports that contain either the words "VW" and "Skandal" (Scandal). She inspects the results in the detail view and sorts them by their similarity. Each report is represented by a thumbnail including the title heading and the first frame of the corresponding video segments. She immediately notices that many thumbnails contain the VW log and headlines shown containing related words like "Dieselmotoren" (Diesel engines), "Abgas-Skandal" (Exhaust gas scandal) or "Umweltbundesamt" (Federal environment office). Encouraged by the results, she clicks on one of the obvious reports and chooses to view the detailed information of that report. She quickly skims the subtitle and the video frame confirms that the report is indeed about the VW Scandal. Subsequently, she labels the reports as "VW" topic. With the help of visualization, she easily identifies and labels more related reports about the VW Scandal. Similarly, she also picks several reports that are not related to the "VW Scandal" and labels them as the "Other" topic.

After she has annotated about ten reports for each label, she decides to train a classifier in order to find more related reports automatically. Therefore, she starts the "Train & Predict" method to train a multi-class classifier for those two labels. Once the classifier is trained, the predicted labels for each report and the overview visualization are updated. The color of the bar chart for each report is also updated that represents the classifier's confidence of the predicted label. With the help of this information, she notices that a similar amount of reports are predicted for the created two labels with high confidence. She has not expected that result, because the corpus consists of all news reports over the last three years and reports about the "VW Scandal" should only be a small portion of it.

As a next step, she filters the reports that have high confidence in the overview visualization. By analyzing them, she finds out that some reports have been wrongly assigned. The images for the reports give her some hints about the reason for the possible errors. For example, one of the reports shows the logo of a bank. She is surprised and wants to find out more about it. Therefore, she selects the thumbnail to investigate the segment in more detail. As soon as she sees the detailed content of the report, she understands the reason. The report is about a scandal of a commercial bank and uses quite similar words as the "VW Scandal". Since such cases confuse the classifier, she relabels the report as "Other". Furthermore, she uses the option to find similar reports and inspect them. She relabels all the wrongly assigned reports with the multi-selection functionality.

During the analysis, she finds two other clusters that confuse the classifier. One of them is about the "Machtkampf" (power struggle) in the VW Company and the other one discusses the "Pkw-Maut" (car charge). As these two events are only loosely related to her investigation goal, she creates two other labels to hold those two clusters of reports and retrains the classifier. This time, the number of reports labeled as "VM Scandal" is reduced a lot. Next, she decides to focus on the reports with low classifier confidence. She corrects the wrongly assigned reports and thus, she resolves the classifier's confusion and improves its ability to differentiate the reports. Figure 2 shows a screenshot of our desktop during the annotation.

She determines that most of the instances for the "VW Scandal" are classified with high confidence now. She is satisfied with the proposed labels for the rest of the reports and finishes her retrieval effort. Compared with manually watching through all videos in the corpus and annotation them accordingly, she just inspected and annotated a much smaller number of videos suggested by the classifier and the visualization overview. Compared with finding the reports using just keywords filters, she achieved a better recall rate,

because the classifier can discover and consider many keywords, which are harder for human users to come up with, from the annotated data.

IV. Discussion and Future Work

In the usage scenario, we show that users can use different strategies for selecting reports which fit better to the different phases of the labeling process. Not only can they use the information from the classifier but also their domain knowledge about the labels to effectively select reports. They can, for example, not only select and annotate the most uncertain reports but also detect and correct high confident errors of the classifier. We hypothesize that our approach will enable users to label more effectively than traditional active learning strategies, which normally rely on the distribution of the data and the prediction confidence of the classifier. We are planning to conduct a comparative study to test this hypothesis. Further, we could allow users to specify different initial term weights for different labels to integrate their domain knowledge more efficiently. In addition, we want to implement an overview, which visualizes and tracks the reports that change their labels during the annotation process. One potential limitation of our approach is the classifier that we used. Even though it is fast enough for the current dataset, however, it could be difficult to keep the retraining in an interactive rate for much bigger datasets (e.g., all tweets in one year). Online machine learning methods [26] could be used to overcome this limitation.

V. Conclusion

In this paper, we presented an approach that supports users in initializing labels with keyword queries and iteratively annotating examples to train a classifier. The proposed approach provides an overview that displays representative results and enables users to follow different annotation strategies (e.g., active learning) and assess the quality of the classifier. These richer set of annotation strategies can potentially allow users to train the classifier more efficiently by providing the classifier more informative examples than labeling the most uncertain data point. A usage scenario demonstrates the applicability of our approach.

References

[1] T. Liu and J. R. Kender, "Lecture videos for e-learning: current research and challenges," in *Proceedings of the IEEE International Symposium on Multimedia Software Engineering*, Dec 2004, pp. 574–578.

[2] M. Brehmer, S. Ingram, J. Stray, and T. Munzner, "Overview: The design, adoption, and analysis of a visual document mining tool for investigative journalists," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 2271–2280, 2014.

[3] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," *Emerging artificial intelligence applications in computer engineering*, vol. 160, pp. 3–24, 2007.

[4] H. Yi, D. Rajan, and L.-T. Chia, "Semantic video indexing and summarization using subtitles," in *Proceedings of the Advances in Multimedia Information Processing (PCM 2004)*, K. Aizawa, Y. Nakamura, and S. Satoh, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 634–641.

[5] A. Fouse, N. Weibel, E. Hutchins, and J. D. Hollan, "Chronoviz: A system for supporting navigation of time-coded data," in *Proceedings of the Extended Abstracts on Human Factors in Computing Systems (CHI '11)*, ser. CHI EA '11. New York, NY, USA: ACM, 2011, pp. 299–304. [Online]. Available: http://doi.acm.org/10.1145/1979742.1979706

[6] M. Kipp, "Anvil: A universal video research tool," Handbook of Corpus Phonology, 2012.

[7] J. Hagedorn, J. Hailpern, and K. G. Karahalios, "Vcode and vdata: illustrating a new framework for supporting the video annotation workflow," in *Proceedings of the Working Conference on Advanced Visual Interfaces*. ACM, 2008, pp. 317–321.

[8] K. Kurzhals, M. Hlawatsch, C. Seeger, and D. Weiskopf, "Visual analytics for mobile eye tracking," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, 2017. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2016.2598695

[9] P. Van Der Corput and J. J. van Wijk, "Iclic: Interactive categorization of large image collections," in *Proceedings of IEEE Pacific Visualization Symposium (PacificVis)*. IEEE, 2016, pp. 152–159.

[10] K. Kurzhals, M. John, F. Heimerl, P. Kuznecov, and D. Weiskopf, "Visual movie analytics," *ACM Transactions on Multimedia*, vol. 18, no. 11, pp. 2149–2160, Nov. 2016. [Online]. Available: https://doi.org/10.1109/TMM.2016.2614184

[11] B. Settles, "Active learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 6, no. 1, pp. 1–114, 2012.

[12] J. Bernard, M. Hutter, M. Lehmann, M. Mller, M. Zeppelzauer, and M. Sedlmair, "Learning from the Best - Visual Analysis of a Quasi-Optimal Data Labeling Strategy," in *Proceedings of EuroVis 2018 - Short Papers*, J. Johansson, F. Sadlo, and T. Schreck, Eds. The Eurographics Association, 2018. [Online]. Available: https://diglib.eg.org/handle/10.2312/eurovisshort20181085

[13] S. Liu, X. Wang, M. Liu, and J. Zhu, "Towards better analysis of machine learning models: A visual analytics perspective," *Visual Informatics*, vol. 1, no. 1, pp. 48–56, 2017.

[14] R. Krüger, H. Bosch, D. Thom, E. Pttmann, Q. Han, S. Koch, F. Heimerl, and T. Ertl, "Prolix - visual prediction analysis for box office success," in *Proceedings of the IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2013.

[15] F. Heimerl, S. Koch, H. Bosch, and T. Ertl, "Visual classifier training for text document retrieval," *IEEE Transactions on Visualization and Computer Graphics*, no. 12, pp. 2839–2848, 2012.

[16] Q. Han, W. Zhu, F. Heimerl, S. Koch, and T. Ertl, "A visual approach for interactive co-training," *KDD 2016 Workshop on Interactive Data Exploration and Analytics*, pp. 46–52, 2016.

[17] M. Liu, S. Liu, X. Zhu, Q. Liao, F. Wei, and S. Pan, "An uncertainty-aware approach for exploratory microblog retrieval," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 250–259, 2016.

[18] J. Bernard, M. Zeppelzauer, M. Sedlmair, and W. Aigner, "Vial: a unified process for visual interactive labeling," *The Visual Computer*, vol. 34, no. 9, pp. 1189–1207, Sep 2018. [Online]. Available: https://doi.org/10.1007/s00371-018-1500-3

[19] B. Shneiderman, "Tree visualization with tree-maps: 2-d space-filling approach," *ACM Transactions on Graphics*, vol. 11, no. 1, pp. 92–99, Jan. 1992. [Online]. Available: http://doi.acm.org/10.1145/102377.115768

[20] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Proceedings of the Association for Computational Linguistics (ACL) Conference, System Demonstrations*, 2014, pp. 55–60. [Online]. Available: http://www.aclweb.org/anthology/P/P14/P14-5010

[21] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008.

[22] "ARD Mediathek," https://www.ardmediathek.de/, accessed: 2019-09-01.

[23] Apache Software Foundation, "Lucene Core," https://lucene.apache.org/core/, 1999.

[24] C. Cortes and V. Vapnik, "Support vector machine," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[25] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *Journal of machine learning research*, vol. 9, no. Aug, pp. 1871–1874, 2008.

[26] P. Laskov, C. Gehl, S. Krüger, and K.-R. Müller, "Incremental support vector learning: Analysis, implementation and applications," *Journal of machine learning research*, vol. 7, no. Sep, pp. 1909–1936, 2006.