A User Centric Visual Analytics Framework for News Discussions

Jakob Smedegaard Andersen
Department of Computer Science
HAW Hamburg
Hamburg, Germany
jakob.andersen@haw-hamburg.de

ABSTRACT

Visual Analytics has achieved a lot of attention for its abilities to support exploratory knowledge discovery in large data sets. In this work-in-progress paper, we develop a Visual Analytics framework for user comments in the domain of online journalism. First, we examine how journalists' needs can be mapped to a visual interactive interface to make sense of user comments. We further investigate how different classes of Machine Learning algorithms like supervised and unsupervised learning can be integrated into Visual Analytics to enable a more user centric analysis. Due to the variety of Machine Learning approaches, we expect that different forms of integration will be needed. Our goal is to place the domain experts (e.g. journalists) in the loop to improve analytical reasoning.

1 INTRODUCTION

As data accessibility increases analytical methods and techniques to handle the data become more important. Visual Analytics (VA) is a novel approach to gain insights from heterogeneous and unstructured data [3, 9]. The basic concept of VA is to combine the processing capabilities of machines with human abilities of pattern detection to overcome the flaws of pure analytical or visual approaches. Interactive Visualisation (IV) is used to bridge these parts together and to enable a more user centric discourse with data. The goal of VA is to provide tools to effectively gain knowledge out of data for better decision-making.

In order to create such tools for big data scenarios, a precise understanding of the coupling of Machine Learning (ML) and IV is required. As not much work is done on how to modify and steer ML methods through interactive interfaces, we focus on user centric possibilities to adapt ML algorithms in the progress of the analysis.

In this paper, we investigate how different types of ML like supervised and unsupervised learning as well as specific methods like clustering, classification, regression and dimension reduction can be integrated in VA. Differences are expected, because of the diversity of those approaches. We use our findings to create a VA framework for user comments in the domain of online journalism.

First, we shortly introduce the VA approach, the concept of "Human in the Loop" and the domain of making sense of news discussions. Secondly, we present our VA Framework for user comments. Hereafter, we argue in section 3.2 for a "Human in the Loop" approach in VA in order to improve our framework. Then, we outline questions for our upcoming research. Finally, we describe related work and end up with our conclusion.

2 PRELIMINARIES

2.1 Visual Analytics

This section shortly introduces the central concepts of VA.

VA is defined as "the science of analytical reasoning facilitated by interactive visual interfaces" [2]. By combining methods from IV with ML and other automated techniques, VA seeks to improve the process of knowledge discovery out of complex structured and unstructured data [9]. This approach is gaining more and more importance due its abilities of integrating human knowledge into computational data processing. VA enables explorative data analytics in large scale data scenarios. Subject is the effective acquisition, expansion and generation of knowledge to finally make better decisions. Within VA the user takes an active role, as he or she steers and supervises the analysis. The interaction becomes the crucial part in which the user communicates his or her knowledge. Several approaches have emerged effectively combining the strengths of human cognition and machine processing [5, 8, 10]. However, further research regarding the user centric coupling of ML and interactive interfaces is needed.

2.2 The Human in the Loop Paradigm

In the early stages of ML the overall question was, "how to construct computer programs that automatically improve with experience" [12]. However, fully automated ML (aML) is not applicable for all real world scenarios. Pure automatic approaches presuppose a good understanding of the problem to achieve beneficial results. They are unsuitable for ill-defined or a priori undefined questions or if needed training data is not available. VA addresses tasks which are explorative in nature [3]. A more seamless approach is needed which fits into the existing interactive process.

A ML approach which can utilise domain knowledge is described by the phrase "Human in the Loop" (HitL). HitL is a special case of interactive ML (iML) and can be defined as algorithms that can optimize their learning behaviour through the interaction with humans [7]. The human is directly integrated in the train, tune and test phase of the algorithm to obtain a higher quality of results. Although ML is a central component of VA, the integration of HitL is not much investigated.

The primary advantage of HitL is the ability to reach inside the models black box. The approach enables the advanced use of the human knowledge and expertise inside a continuous feedback loop. User interaction empowers model steering and is not limited to model selection and parameterization. However, the HitL approach rises new challenges in getting the most out of the participation with humans. Novel approaches are needed to make this interplay more intuitive and beneficial.

A typical use case or HitL is supervised learning. In this setting, the goal is to learn a mapping between X and Y given a set of training pairs (x_i, y_i), whereas $x_i \in X$ are called examples and $y_i \in Y$ are labels for $i \in \{1, ..., n\}$. A typical supervised HitL approach is to increase n as the number of example pairs to gain more accuracy through enhancing the grounded truth. However, HitL can also be applied to unsupervised learning settings. Here only a set of n examples $X = \{x_1, ..., x_n\}$ is given. The goal is to find interesting structures in the examples X. By applying HitL the problem can be transformed into a semi-supervised learning problem if a user provides l labels for some examples in X. The examples get divided in $X_l := \{x_1, ..., x_l\}$ for which the labels $Y_l := \{y_1, ..., y_l\}$ are given and $X_u := \{x_{l+1}, ..., x_n\}$ as the set with no labels. In this case, the training pairs can be used to dynamically guide the computation by constraints or additional information. See Capelle et al. [1] for details.

2.3 User Feedback in News Discussions

This section introduces and motivates the domain of analysing user feedback in news discussions. User feedbacks refers to various forms of user participation regarding journalistic contents. The focus lies on textual comments from multiple channels like news webpages, emails and social media.

There is a high demand of gaining insights from user comments, as they are a valuable source of information. They can contain useful aspects like feedback, critics, new perspectives and expertise. Furthermore, comments mirror personal opinions which are normally hard to capture [13]. On the backside, comments can include insults, hustles and advertisements which can negatively affect the overall quality of the discussion.

Studies show that user comments support the daily work of journalists and editors [16]. This includes the obtaining of new ideas for further articles, additional facts and direct feedback for improving the work of journalists. Thus, observing user comments has clearly a positive value. However, the heterogeneity and large volume raises a number of challenges, such as moderation overhead costs and overview of the current state of the discussion [11].

Consequently, newsrooms are faced with an increasing demand for computer supported ways to analyse, aggregate and visualize user comments. Unfortunately, there is a lack of analytical tools to provide high quality comments that can be leveraged for journalistic purposes [11].

From a data science point of view, user comments are documents with heterogeneous information contexts and several connections between each other. They consist of multiple attributes like a commenter identification, the related article, a timestamp, a title, a ranking from other readers and several annotations from manually or automated classifications like sentiment analysis and swearword detection.

3 MAKING SENSE OF USER COMMENTS

As the manual analysis of user comments is resource consuming and unpractical with a rising volume, velocity and variety of user comments, various researchers focus on approaches to detect patterns automatically [6, 19]. In order to enable better analytical reasoning, we constructed a VA framework for annotated user comments.

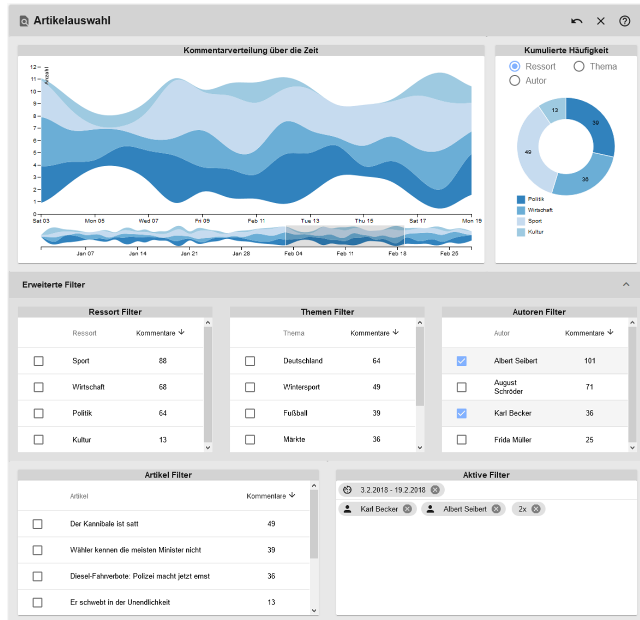

Figure 1: The *Article Selection* view.

3.1 Visual Analytics Framework

We have developed a fully functional VA framework for user comments. The framework covers the needs of journalists. Our research builds upon the findings of Loosen et al. [11]. They propose requirements for an analytics tool in the field of user comments which covers the needs of journalists. The aim of our work is develop and further examine these findings in a VA tool. The primary question is how to map the analytical requirements into a suitable combination of visualisations and interactions to fulfil journalists' needs. As our data collection, we use a set of pre annotated user comments. Our framework consists of the following analytical dimensions: See Loosen et al. [11] for further details.

- **Article Selection**
  The occurrence of comments in time/progress of a discussion. The user can select samples of articles from which the comments are analysed.
- **Topics and Addressees**
  What is discussed and who is mentioned and directly addressed in a sample of user comments.
- **Discussion and Argumentation**
  The direction of a discussion and risen arguments over the time. The user can analyse the development of pro- and contra-arguments towards a certain question or topic over time.
- **Quality**
  Metrics to quantify the quality of the user comments to offer a condensed overview.
- **Selected User Comments**
  A close read function for selected user comments.

Each dimension is implemented as a separate view within a web application. Views supply a set of visualisations and interactions to

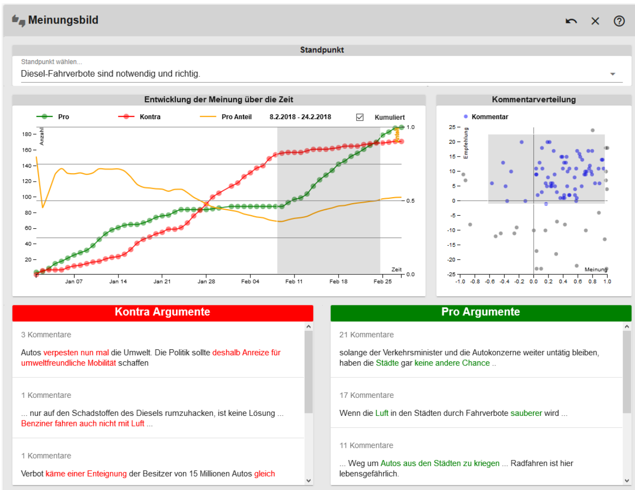

Figure 2: The *Discussion and Argumentation* view.

Figure 3: The *Quality* view.

enable and support an analytical discourse with the user. The views are coordinated and implement different interaction strategies like *selections*, *filtering*, *focus+context* and *linking+brushing*. Every view is an optional part of the analysis. The user decides dynamically, which views he or she wants to use. The layout of the views are based on a "*filter flow*" metaphor. The views possess a specified ordering in which they are placed on the screen. Changes inside a view are only forwarded to subsequent views. The user is able to filter in arbitrary order and there are no top-down restrictions.

Figure 1 shows the *Article Selection view*. The upper part of the figure depicts the distribution of comments over time, whereas the lower part offers several selection options. Comments can be selected individually or grouped by sections, topics and authors. The *Discussion and Argumentation view* is depicted in figure 2. After the user selected a stance or topic the distribution of the for and against comments are shown in a line-graph. Related arguments are listed in the bottom part. Furthermore, selected comments are plotted regarding to their sentiment and user ratings. The grey boxes relate to filtering operation and only comments within are considered. Figure 3 illustrates the *Quality view*. User comments are represented as a set of different indicators. In the upper part user comments are categorised along the dimensions *article reference, compliance* and *originality*. For further analysis, each of the above selected comments are than depicted as polylines inside parallel coordinates. As indicators e.g. the *length*, *sentiment*, or *number of references* are used.

Another central part in our investigations is the evaluation of our prototype. We conduct a quantitative usability study after the end of the first implementation cycle to assess the overall system and to counteract any weak points. Seven participants are observed while solving real world problems with our prototype. In addition, they answer a questionnaire about the usability. The findings are used for improvements and further requirements. The evaluation also reveals that the participants unconditionally trust the results. This is not surprising as the tool does not show the accuracy of the analytical results.

3.2 From Interactive Filtering to Human in the Loop

In our first prototype the processing capabilities of VA is not exhausted. One drawback of our prototype is the lack of interactive model adaptation. The prototype lacks in integrating users' expertise and experience in the analysis. The current interaction takes place as a sequence of *selections*, *overviews* and *filter* operations. The integration of ML is limited to pre-processing.

To enable HitL and interactive model steering and thus exploit the possibilities of VA, we extend our prototype with interactive ML components. We have to develop specific use cases that include HitL approach, which can also be used in a more general form across domains.

The question arises why it could be beneficial to place the journalist in the loop. At first, our framework is based on several classification models that provide annotations for interactive visualisation. Since these algorithms are based on simplified models of reality, misclassifications are to be expected. The misclassification likelihood can be quantified by relative error frequencies. By using HitL, the user can interactively correct spotted misclassifications and initiate new training cycles to improve the model's accuracy, which affect the visual correctness.

Secondly, HitL opens up new possibilities of individualisation. It is difficult to satisfy the desire for customizable queries through a predefined set of filter operations. HitL makes it possible to dynamically select data that are of special interest. The user can be offered opportunities to create and apply generic classification models at runtime.

Furthermore, many ML algorithms work on restricted information basis. There is only a limited amount of data available, only a sample is used for computation or the algorithms are incorrectly configured. This leads to different understandings of similarities or weights between human expectations and computational results. HitL makes it possible through experience of the users to correct these deviations and thus to improve the results.

4 NEXT STEPS

For our next steps to integrate different ML methods in VA, the following questions arises:

RQ1: To what extent can different ML methods be adapted within an iterative process?

A taxonomy for user centric adaptations of ML methods will be developed to guide the development of VA applications. The integration of HitL aspects is carried out with regard to the difference of supervised and unsupervised approaches. In addition, specific methods like classification, regression, clustering and dimension reductions will be considered.

RQ2: How can model adaptations and responses be translated and mapped into a visual metaphor?

It is a difficult task to steer computational models to match expectations. There is a gap between computational processing on the machine side and cognition on the human side that can lead to hard usability problems [17]. The user has to translate his or her mental model into numeric variables to steer the computation. The challenge is to provide a visual layer which abstracts the computational perspective, as pretty much every journalist (or end user) do not want to deal with ML details.

RQ3: How can the uncertainties of ML method be visually communicated and utilised in a constructive manner?

Uncertainty is created and communicated over the complete VA process [15]. We will focus on uncertainties created by ML algorithms. There exist several quality measures which quantify the accuracy of ML algorithms. We want to discover how these uncertainties can be communicated in an appropriate manner to create awareness. Furthermore, we want to utilize the uncertainty in combination with the HitL approach to reduce misclassifications.

In order to answer these questions, we examine the integration of ML and IV from a user centric point of view. To support the suitability of our findings, we develop and evaluate several prototypes in the domain of news discussions. A goal is to cover different ML approaches. We will carry out quantitative usability tests with representatives to spot and document strengths and weaknesses. The observation will take place inside our usability lab which consists of eye-tracers, cameras, screen captures and key loggers.

5 RELATED WORK

In [18, 20] the authors present VA tools to make sense of text collections. In contrast to our approach they do not place the human in the loop and rely heavily on pre-processing. Our modelling of the human interactions with ML is similar to [14], but they do not distinguish between various ML strategies like supervised and unsupervised learning. Additionally, [4] is related to our work as they come up with a novel principle for analytical interactions called *semantic interactions*. We will build upon these findings to provide interactions that derive from the user's analytic process.

6 CONCLUSION

We have developed a VA tool for user comments in online journalism and have outlined next steps for a user centric integration of ML in our prototype. The next steps cover how different ML methods like supervised and unsupervised learning can be adapted within an iterative process, how these adaptations can be visual translated and mapped to a visual metaphor as well as how occurred uncertainties can be communicated and constructively used. A broader understanding of the coupling of ML and IV is necessary to fully exploit the strengths of VA.

ACKNOWLEDGMENTS

The paper was supported by BWFG Hamburg within the "Forum 4.0" project as part of the *ahoi.digital* funding line.

REFERENCES

[1] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. 2009. *Semi-supervised learning.* MIT Press.

[2] Kristin A. Cook and James J. Thomas. 2005. *Illuminating the path: The research and development agenda for visual analytics.* IEEE Computer Society.

[3] Geoffrey Ellis and Florian Mansmann. 2010. *Mastering the information age solving problems with visual analytics.* Taylor & Francis Group.

[4] Alex Endert, Patrick Fiaux, and Chris North. 2012. Semantic interaction for visual text analytics. In *Proceedings of the SIGCHI conference on Human factors in computing systems.* ACM, 473–482.

[5] Alex Endert, William Ribarsky, Cagatay Turkay, BL William Wong, Ian Nabney, I Díaz Blanco, and Fabrice Rossi. 2017. The state of the art in integrating machine learning into visual analytics. In *Computer Graphics Forum*, Vol. 36. Wiley Online Library, 458–486.

[6] Marlo Häring, Wiebke Loosen, and Walid Maalej. 2018. Who is addressed in this comment?: Automatically classifying meta-comments in news comments. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 67.

[7] Andreas Holzinger. 2016. Interactive machine learning for health informatics: When do we need the human-in-the-loop? *Brain Informatics* 3, 2 (2016), 119–131.

[8] Liu Jiang, Shixia Liu, and Changjian Chen. 2018. Recent research advances on interactive machine learning. *Journal of Visualization* (2018), 1–17.

[9] Daniel A. Keim, Florian Mansmann, Jörn Schneidewind, Jim Thomas, and Hartmut Ziegler. 2008. Visual Data Mining. In *Theory, techniques and tools for visual analytics.* Springer, Chapter Visual analytics: Scope and challenges, 76–90.

[10] Shixia Liu, Xiting Wang, Mengchen Liu, and Jun Zhu. 2017. Towards better analysis of machine learning models: A visual analytics perspective. *Visual Informatics* 1, 1 (2017), 48–56.

[11] Wiebke Loosen, Marlo Häring, Zijad Kurtanović, Lisa Merten, Julius Reimer, Lies van Roessel, and Walid Maalej. 2018. Making sense of user comments: Identifying journalists' requirements for a comment analysis framework. *SCM Studies in Communication and Media* 6, 4 (2018), 333–364.

[12] Tom M. Mitchell. 1997. *Machine learning.* McGraw-Hill.

[13] Deokgun Park, Simranjit Sachar, Nicholas Diakopoulos, and Niklas Elmqvist. 2016. Supporting comment moderators in identifying high quality online news comments. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16).* ACM, 1114–1125.

[14] Dominik Sacha, Michael Sedlmair, Leishi Zhang, John A. Lee, Jaakko Peltonen, Daniel Weiskopf, Stephen C. North, and Daniel A. Keim. 2017. What you see is what you can change: Human-centered machine learning by interactive visualization. *Neurocomputing* 268 (2017), 164–175.

[15] Dominik Sacha, Hansi Senaratne, Bum C. Kwon, Geoffrey Ellis, and Daniel A. Keim. 2016. The role of uncertainty, awareness, and trust in visual analytics. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016), 240–249.

[16] Arthur D. Santana. 2011. Online readers' comments represent new opinion pipeline. *Newspaper Research Journal* 32, 3 (2011), 66–81.

[17] Jessica Z. Self, Radha K. Vinayagam, J. T. Fry, and Chris North. 2016. Bridging the gap between user intention and model parameters for human-in-the-loop data analytics. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics (HILDA '16).* ACM, New York, NY, USA, Article 3, 6 pages.

[18] John Stasko, Carsten Görg, and Zhicheng Liu. 2008. Jigsaw: Supporting investigative analysis through interactive visualization. *Information Visualization* 7, 2 (2008), 118–132.

[19] Gregor Wiedemann, Eugen Ruppert, Raghav Jindal, and Chris Biemann. 2018. Transfer learning from LDA to BiLSTM-CNN for offensive language detection in twitter. *Austrian Academy of Sciences, Vienna September 21, 2018* (2018), 85–94.

[20] Yi Yang, Quanming Yao, and Huamin Qu. 2017. VISTopic: A visual analytics system for making sense of large document collections using hierarchical topic modeling. *Visual Informatics* 1, 1 (2017), 40 – 47.