



Visual Analysis of Humor Assessment Annotations for News Headlines in the Humicroedit Data Set

K. Kucher¹ , E. Akkurt¹, J. Folde¹, and A. Kerren^{1,2} 

¹Department of Science and Technology, Linköping University, Sweden

²Department of Computer Science and Media Technology, Linnaeus University, Sweden

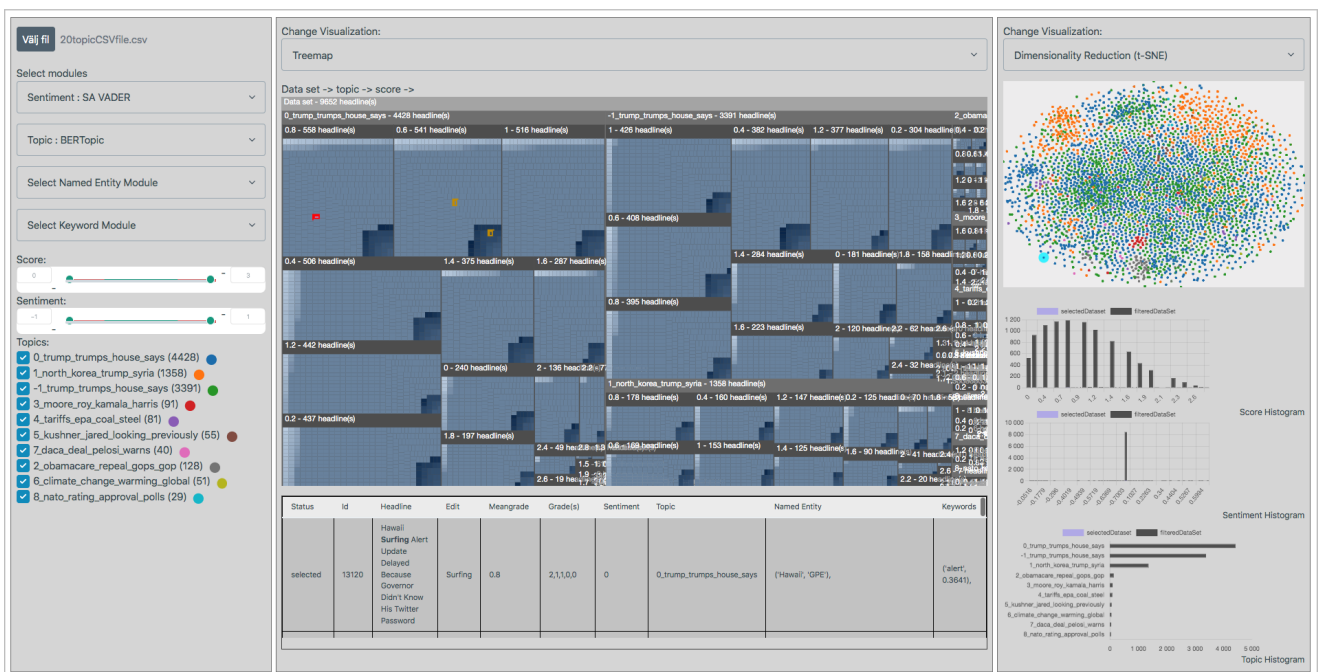


Figure 1: The user interface of our interactive visualization prototype. The left panel provides the selection of the input data and chosen computational analyses as well as filters; the central panel presents the main interactive view as well as the data table with details; and the right panel provides the secondary view as well as distribution plots.

Abstract

Effective utilization of training data is a critical component for the success of any artificial intelligence algorithm, including natural language processing (NLP) tasks. One particular task of interest is related to detecting or ranking humor in texts, as exemplified by the Humicroedit data set used for the SemEval 2020 task of assessing humor in micro-edited news headlines. Rather than focusing on text classification or prediction, in this study, we focus on gaining a deeper understanding and utilization of the data through the use of information visualization techniques facilitated by the established NLP methods such as sentiment analysis and topic modeling. We describe the design of an interactive visualization tool prototype that relies on multiple coordinated views to allow the user explore and analyze the relationships between the annotated humor scores, sentiments, and topics. Evaluation of the proposed approach involves a case study with the Humicroedit data set as well as domain expert reviews with four participants. The experts deemed the prototype useful for its purpose and saw potential in exploring similar data sets with it, as well as further potential applications in their line of work. Our study thus contributes to the body of work on visual text analytics for supporting computational humor analysis as well as annotated text data analysis in general.

CCS Concepts

• **Human-centered computing** → Visual analytics; Information visualization; • **Computing methodologies** → Natural language processing;

1. Introduction

Identification, prediction, and generation of text with highly subjective and context-dependent properties are important and difficult challenges in computational linguistics. Sarcasm and irony detection are examples of natural language processing (NLP) tasks that are considered challenging, including the issue of agreeing on the consistent annotations/labels for particular sentences or documents among human annotators with respect to such elusive categories. Computational approaches for identifying and analyzing humor in texts have also been in the focus of the NLP research community, with the shared task (contest) titled “Assessing Humor in Edited News Headlines” recognized as the best task at SemEval-2020 [HKGK20]. The respective task provides a data set titled *Humicroedit* that includes news headlines in English with minimal edits (e.g., single word replacements) made in order to make the respective headlines humorous [HKG19]. The actual level of humor/funniness was assessed by multiple annotators on an ordinal scale [HKG19, HSKS20]. The aim of the contest was to discover better-performing computational methods that would predict the funniness level or rank two edited versions of a headline. While a number of solutions focusing on such regression and classification tasks were proposed [HKGK20], there are further questions to be asked and insights to be discovered within the respective data, for instance, how consistent are the funniness level annotations across the topics, or to which extent can the funniness scores range across several related headlines.

The aim of this study (based on a thesis project [AF23]) was to create an interactive visualization tool (see Figure 1) to analyze and get a better understanding of how the contents of the Humicroedit data set affect its funniness level, e.g., what makes them more or less funny with regards to the topics and/or sentiment of the text, and to a lesser extent, detected named entities and keywords. The visualization tool is intended for natural language and linguistics researchers working on similar tasks, and, to the best of our knowledge, this is the first contribution from the visualization community so far to focus on the Humicroedit data set, and one of few contributions related to visualization of computational humor analysis, which have focused on other aspects and settings of this problem [WMW*22]. More specifically, we address the following research questions (in the context of the Humicroedit data set here and below, if not indicated otherwise):

- RQ1** How can we represent and interact with a humor annotation data set, so that insights could be drawn out of the relationship between the annotated humor/funniness scores and contents of such a data set?
- RQ2** How does the sentiment identified for original headlines, replacement terms, and resulting modified headlines relate to the annotated humor/funniness score in the complete data set?
- RQ3** What is the relationship between the annotated humor/funniness scores and the topics/terms as well as the particular named entities detected in the complete data?
- RQ4** How does the sentiment identified for original headlines, replacement terms, and resulting modified headlines relate to the topics/terms as well as the particular named entities detected in the complete data?

2. Related Work

As outlined in Section 1, the purpose of this project is to gain insight into humor-annotated data through the means of visual analytics (more concretely, to develop a visualization tool tailored to the data schema available in one such data set). Humor-annotated data has up to this point in time been a relatively unexplored territory in the world of information visualization, including the more specific areas of text visualization [KK15, AL19, KSD*22]. At the time of this writing, there are—to the best of our knowledge—no known visual analytical approaches that are completely compatible with the particular problem the data set provides; the very recent DeHumor approach by Wang et al. [WMW*22], for instance, focuses on in-depth multimodal analyses of comedy performances, which is inarguably relevant to the general challenges of both computational and visual/interactive humor analyses, but not specifically relevant to humor analysis for (micro-)edited news headlines.

Although there is a considerable knowledge gap about distinctly humor-annotated data, prior approaches exist in which visual analytics is utilized to delve into and visualize NLP-related data. The larger the data sets, the more valuable information might be contained and eventually discovered within them. However, as the data sets grow in volume, scaling issues follow and manual annotation becomes less doable for large quantities of data. Kuksenok et al. set to address this issue as this pertained to affecting labeled annotated data from digitally-mediated communication [KBR*12]. They presented a visual text analytics tool that enabled manual and automatic annotation of data, which was visualized as chat-log inputs in timelines and calendars. The respective tool [KBR*12], similar to our proposed approach, is intended to provide visual analytics to resolve tasks such as evaluating discrepancies and finding subsets in the data that enlist interest. However, the chat-log data is temporal, unlike the case of the Humicroedit data set. The edited headlines work as isolated headlines independent of time, which renders visualization in the form of timelines meaningless.

The study by Jentner et al. recently introduced an innovative method to visualize confusion in labeled data [JLH*23]. An image of a diamond is created with the use of color and icons to convey how the data might have been wrongly classified. The purpose of the visualization is to enhance the performance of machine learning by detecting errors in classifications and the data itself. Like the previous related work example, this method was tested on chat logs. Detecting errors of this kind could be useful in the setting of annotated humor. Irony and sarcasm are often utilized to generate humor. They are often illogical and reading between the lines is often required to grasp the concept, thus more trouble for some to detect through the means of machine learning and/or NLP processes. Therefore it could be argued to be reasonable to consider such potential errors in humor-annotated data and in turn how it could be visualized. However, this is outside the scope of this project, where more exploration-oriented tasks are more relevant to its core.

ALVA is a visualization analytic approach whose purpose is to aid the detection and classification of stances in textual data using machine learning and NLP processing as issues of classifying training and data collection often occur [KPSK17]. The solution ALVA presents is a novel representation of annotated data called CatCombs. As one document may be perceived to enclose multiple atti-

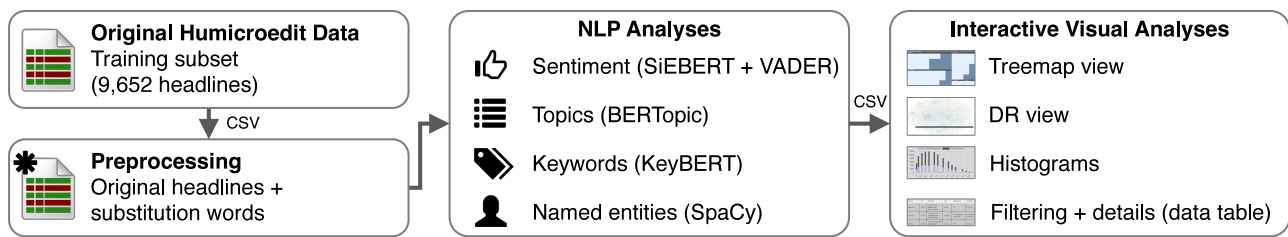


Figure 2: The pipeline comprising data preprocessing, computational, and interactive visual analyses for the Humicroedit data set adopted in this study. The current design separates the computational backend (with its output stored in separate CSV files) and interactive frontend.

tudes, this is illustrated as dots representing individual annotations that are rendered inside blocks. The blocks in turn are arranged according to the number of different combinations of stance categories the blocks contain. More information is obtained by hovering over and/or selecting a block. The selection of a block enables links to another block that represents the same document, but that has been annotated differently, and thus is represented by another combination of stance categories [KPSK17]. Still, ALVA is not directly applicable for our particular problem of annotated humor, and further solutions must be discovered.

3. Data and Computational Methods

As mentioned above, the aim of this study is to provide an insightful analysis of the Humicroedit data set [HKG19, HKGK20], e.g., understanding what makes headlines more or less funny. The actual data set consists of CSV files with the following columns: ID, original news headline (with the word to be edited indicated) in English, the edited word (i.e., the substitution), annotated funniness grades given to the edited headline (ranging from 0 “Not funny” to 3 “Funny”, and provided as an ordered list of individual scores such as “22100”, which makes it impossible to differentiate between the grades provided by an individual annotator over the complete data set, unfortunately), and the mean value of all given grades. The substitution words were either a noun, possibly a named entity, or a verb. The overall data set includes about 15,000 edited headlines that are split into training/validation/test subsets for the shared task purposes [HKGK20], and further similar edited headlines are provided by the authors as additional training resources [HKSK20]. Since our work does not aim to develop a computational model for predicting the funniness score or ranking two alternative edited headlines in terms of their funniness, we mainly focus on the data provided in the training subset in this paper (unless specified otherwise below), which includes 9,652 edited headlines.

The pipeline designed for this study is presented in Figure 2. The preprocessing step includes data cleaning, so that complete edited headlines are constructed from the original headlines + substitution words. Next, NLP techniques are applied to the preprocessed data with the goal of extracting sentiment, topics, keywords, and named entities from the respective texts. The modules currently used for each of these tasks are the following (while further implementations could be plugged in as part of future work): (1) SiEBERT [HHSS23], a fine-tuned model based on the RoBERTa model [LOG*19], and VADER [HG14] for sentiment analysis; (2)

BERTopic [Gro22] for topic modeling; (3) KeyBERT [Gro20] for keyword extraction; and lastly, (4) SpaCy [Spa23] for named entity recognition. The results of these computational analyses are fused and exported as a CSV file for further exploration using the interactive visualization prototype, as discussed in the next section.

4. Visualization Design

The initial design iteration for the interactive visual interface established the need to rely on multiple coordinated views and support for standard user tasks such as overview, filtering, brushing, and details on demand [Shn96, Rob07]. While discussing the design, we agreed that individual data distributions for the annotated funniness scores, sentiment polarity values, and topics could be represented with standard histograms—however, the choice of visual representations [GPQX07] that could provide an overview of several of these data facets was not a trivial task. As part of this step, several design alternatives were discussed, as presented in Figure 3. A parallel coordinate plot [HW13] is a classical multivariate data representation technique that could definitely support three quantitative (funniness and sentiment) as well as qualitative (topics) attributes (see Figure 3(a))—however, the downside for this technique with the intended thousands of data items would be severe cluttering that would require additional interactions and alleviation approaches [ZYQ*08]. Another alternative (see Figure 3(b)) would be to rely on a classical scatter plot, with two quantitative attributes of interest (funniness and sentiment) mapped to the axes, while the qualitative attributes (topics) could be represented with other visual variables/channels such as color. Here, a single scatter plot dot would represent the mean funniness and sentiment scores for a topic; another alternative would be to represent each annotated data item with a dot and to explicitly represent relationships between the items based on the same original headline (see Figure 3(c)). The main downside of both these alternatives was also related to the severe clutter, especially considering the potentially low variance of both funniness and sentiment polarity scores.

Based on these considerations, we chose the final design implemented in our visualization tool prototype, which is presented in Figure 1. The prototype is implemented with D3.js, Chart.js, and Pico CSS. The panel on the left provides the input file selection and computational analysis module selection (see Section 3) functionality, followed by filters for funniness and sentiment scores as well as topics. The central panel provides the main view, which is here chosen to use a treemap representation [STLD20] of the data

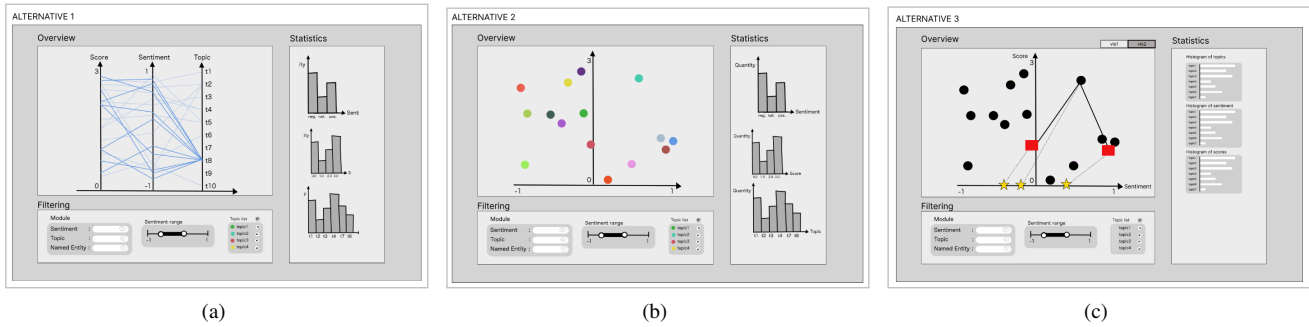


Figure 3: Initial sketches of the interactive visualization prototype, with the main visual representation chosen to be: (a) a parallel coordinates plot, (b) a scatter plot (average values of funniness score and sentiment across each topic), and (c) a linked scatter plot presenting relationships between edited headlines for the same original on demand.

set contents, with individual annotated items binned by funniness score and nested under topics. The adoption of the treemap is predicated on its efficacy in representing extensive data sets with relatively shallow hierarchies [MSJF20] conveniently, and its inherent capacity to depict multivariate data within categorized structures, thereby facilitating the portrayal of data distribution across the entirety of the data set. The treemap supports brushing (note a cell selection with red border in Figure 1, with two related headline cells highlighted with orange border), navigation between nested levels (e.g., the user can click on a parent node in order to expand its contents), and details on demand via hovering. Under the main view, the details are presented in a table: the edited headline selected in the main view is displayed in the first row of the table, with its original headline displayed in the second row, and afterwards, other related edited headlines. The panel on the right provides a secondary view (here, a dimensionality reduction (DR) plot for individual headline items based on t-SNE [VdMH08], with color indicating the respective topic) and several histograms representing distributions of the annotated funniness scores, sentiment polarity values, and topics for the selected data vs the complete data set. The DR technique serves as a useful starting point to identify similarities within a data set. In a cluster, it becomes easier to explore the relationships between headlines and their contents, including their sentiment, topic, named entities, and score. This can, for instance, be achieved through interactive features such as brushing and linking (the secondary view highlights the items selected in the main view) as well as hovering (details are presented with a tooltip and, furthermore, edges towards related headlines are rendered). The dimensionality reduction is based on the textual likeness of the edited headlines, which are transformed into embedding vectors with the SBERT model [RG19], concatenated with their funniness score and sentiment values. While our initial choices of DR techniques included PCA and t-SNE, only the latter was included in the prototype implementation; this choice in the context of providing a topical overview is corroborated by the prior work [NA19, EMK*21, ACT*24, ACS*24], although this is certainly not the sole viable option, and support for further techniques (similar to the computational analysis modules) can be considered part of future work. It should also be noted that the user is able to switch the representations used in the main and secondary

views independently, e.g., to use the DR plot as the main view or use two treemaps / two DR plots instead. Finally, the bottom part of the right panel includes several histogram views representing the distributions for the selected vs complete data with respect to the mean funniness score, sentiment polarity, and topic.

5. Evaluation

In this section, we discuss several forms of evaluation of the resulting prototype that took place as part of our study.

5.1. Case Study

In the case study, we engaged in an in-depth analysis of the available data set, applying the tool to discover interesting findings and obtain new insights. Focus has been dedicated to the interplay between the annotated funniness scores and the computational analysis results for the contents of headlines, e.g., sentiment, topics, keywords, and named entities, as mentioned among our research questions in Section 1. The first two cases that were investigated were low ([0,0.2]) and high ([2.8,3]) funniness score ranges vs different sentiment intervals such as negative ($[-1, -0.1]$), neutral ($[-0.1, 0.1]$), and positive ($[0.1, 1]$) polarity. Furthermore, similar exploration was conducted for funniness score vs topic and sentiment vs topic. Further insight into aspects such as the named entity and keywords have been considered, although to a lesser extent.

To briefly summarize the results, we have witnessed a noticeable difference between the results produced by VADER and SIEBERT/ROBERTa with respect to the produced polarity distributions: for instance, with ROBERTa, most headlines (1,064) lie in the negative interval, and with VADER, most headlines (1,201) lie in the neutral sentiment interval. This shows that ROBERTa has a tendency to classify text more negatively for this data set. Several individual cases of interest that were discovered for higher annotated funniness scores (and neutral sentiment with VADER, for instance) led us to the conjecture that it was the context of the headline text and the edited word together that made them funnier—the edited word was not in particular funny itself in these cases. Another note was that if the context of the headline was funny, the related edited headlines typically also had higher scores. For the cases involving

lower scores, we typically discovered either dark or serious subjects, or that the edited word generally did not fit very well into the context of the headline, e.g., there was no obvious connection or logic between the edited word and the context. Other findings include a note concerning potential biases with respect to higher numbers of occurrences of specific topics, etc. in the data (such as the former US President Trump) and the cases indicating relationships between specific keywords (often related to someone's physical appearance, intellectual abilities, or questionable morals), negative sentiment, and high funniness scores (which indicates more crude and insulting forms of humor, perhaps, which could still be appreciated by some annotators in certain contexts).

5.2. Expert Feedback

Semi-structured interviews with domain experts were conducted to evaluate the prototype [KSD*22]. Four participants were interviewed in total: two experts in the NLP research field, one expert in computational social science with expertise in the application of NLP techniques, as well as a linguist with experience in NLP. The roughly one-hour sessions were conducted over Zoom, and the protocol included a brief overview of the problem, data, and computational pipeline, followed by a live demo of the visualization prototype and a discussion. The questions ranged from how the interviewees experienced specific visualizations and features, to more general questions, for example, if they could see further use of a tool akin to the prototype in their own line of work.

5.2.1. Feedback for Treemap

The experts were asked if they gained a deeper understanding of the data set at hand after the presentation and live demo of the treemap feature. The consensus was that they did, as they knew nothing of the Humicroedit data set at the beginning of the session, but felt more informed at the end of it. Some commented specific take-aways they learned from the prototype were, for example, the distribution of the topics in the data set through the use of the treemap. On the other hand, they did not perceive the same level of comprehension regarding the funniness score and sentiment values. However, it was their impression that this lack of understanding was not because of some failure in the presentation of the data set nor the interface of the prototype, but rather due to the time constraints and that would need to interact with the prototype directly for some time to get used to the features at their own pace. Some brief analyses of the score and topic could be obtained with the help of the histograms. Still, more time was considered necessary for a more exhaustive understanding of the data set.

As for criticism of the treemap, although most experts found the hierarchy of the treemap adequate, some suggestions were voiced to reorder the treemap in order to enhance the view of a certain attribute at specific times/steps. From one perspective (from the linguist), it would also be useful to have a visualization that had a greater focus on the substitution word and detected keywords and how they correlate to sentiment and the annotated funniness level. Some confusion also arose regarding the sentiment color coding. The positive nodes had a darker color tone that created a greater contrast to the neutral sentiment color, which made the two groups easier to distinguish from each other. The negative nodes had a

lighter color and were similar in hue to the neutral sentiment color, leaving them less noticeable than the positive nodes. The treemap was laid out using the squarified method, which resulted in the negative nodes being placed alongside the top left edges of the parent node, while the positive nodes clumped together in the opposite bottom right corner. The color contrast in conjunction with the node placement led to some experts finding it difficult to study the negative headlines compared to the positive ones in the treemap.

Overall, the participants responded well to the treemap and its features. It provides the user with a straightforward exploration of the data set by relatively easy means. The distribution of different subsets and what those sets contained was displayed in a convenient way. One expert concluded, *"It is very good to get an overall idea about the data set and the predictions of the model"*. A trait of the treemap that the experts found favorable was how the user could gain an overview of the entire data set, while at the same time getting direct access to individual documents. One expert noted *"... I think it is good here that you actually get the possibility to get a representation of each document. That seems really useful."*, furthering this reasoning that in their experience in topic modeling often it is needed and desirable to view the original/raw data that have been processed. The usual procedure is to extract the most typical data items in the data set and view the corresponding documents separately. By using the proposed prototype, these extra steps can be reduced. Another aspect of the treemap the experts found interesting was how the related data items were highlighted when the user selects individual headline nodes. With this, they could see how small changes to the same headline could affect variables such as score and sentiment values. In addition to this, they could also view how headlines and their relatives were rearranged as different modules are enabled, with one expert stating *"Now this is really interesting to see where different scores (headlines) end up in the other blocks"* about the placement of a particular group of headlines as the sentiment module was switched from VADER to RoBERTA.

5.2.2. Feedback for DR Plot

When it came to the dimensionality reduction plot, the experts appreciated it as an alternative way to look at the data set. The zoom feature was valuable as it let the user conveniently study clusters of data points that were of interest, reflecting that knowing what makes the headlines similar in the data could be very useful. Such knowledge is intrinsic to understanding the data better.

One expert mentioned the possibility that some might be overwhelmed by the amount of information presented in the treemap if they are interested in just a particular set of data, or that they maybe do not have time at hand to skim the whole treemap to find what they need. In that case, the DR plot could make up for what the treemap lacks in the efficiency of detecting and selecting a subset of data. The back end of the DR plot was questioned by one expert, saying that they found little sense in including the sentiment and score value into the vectorization of the headlines as those values would not contribute much effect to the contents in the final graph as compared to hundreds of dimensions produced by the embedding approach itself. An additional comment to this choice of combining the score, sentiment, and headlines textual data could be hard for the user to grasp. They suggested keeping them separate,

keeping the text as the base for the DR computation and plotting, but representing the score and sentiment through other channels, such as the color. The same effect could be obtained with the prototype in its current state by filtering the data using the score and sentiment slider, but the proposed color coordination would convey information faster and with fewer steps in practice.

5.2.3. Comparing Treemap and DR Plot

When comparing the treemap and the DR plot, the experts contrasted the strengths and limitations of both visualization methods, seeing how they conceptually are quite different. Some preferred to view the data via the treemap as it presents the data as a whole and also conveys the score and sentiment value more candidly. Additionally, the treemap could demonstrate the result of the sentiment modules in a way that the dimensionality reduction plot could not. According to one of the experts, DR is already readily used in the NLP literature, and therefore they found the treemap to be more engaging for its novelty and ability to provide a lot of information instantly. A general case of confusion could be found in the table when the user changed the visualization method to interact. It was noted that the order of the same type of items did not stay consistent for the treemap vs the DR main view. Likewise, the selection of several data points could add to the disorientation. A recommendation from one of the experts was to clarify the switch of data in the table by changing the background color of rows, indicating different types of selections in the main view.

6. Discussion

Treemap Limitations Issues concerning scalability have become apparent during the implementation of the treemap. There is a large size difference between some of the topics, e.g., one may cover 30–50% of the treemap area, while another may just take up under 5%. The smaller topic formation congregates in the bottom right corner of the treemap, resulting in clutter where titles of parent nodes overlap and the headline nodes are too small to even distinguish at times. Various attempts have been made to produce a more readable treemap, including different binning, number of topics, and nesting structure, but further work is necessary to address the problem in a more general way, e.g., by considering and evaluating further hierarchical representations for this task [SHM14, MSJF20, KCW*21].

DR Plot Limitations While the overall DR plot functionality satisfies our design intentions and was received positively during the expert review, we should acknowledge the comment mentioned in Section 5.2.2 regarding the weak affect of the funniness score and sentiment values on spatialization, as the resulting positions seem to almost exclusively rely on the text embedding vectors.

Expert Review Limitations In general, the experts responded well to the prototype and the visualization provided information from both a wider perspective as well as on a detailed level. Some aspects were not too easy to grasp directly, e.g., the related headlines, but these confusions were sorted out in due time as the sessions proceeded. It should also be noted, though, that most participants would have liked to interact with the prototype longer and that direct interaction with the tool rather than a live demo via Zoom

(which also had implications with respect to the screen resolution, video feed quality, and performance) could be beneficial.

Annotation and Comparison of Multiple Data Sets Additional interactive workflows and tasks that were considered, but not implemented in our prototype due to time and scope limitations, included support for editing / creating annotations directly within our prototype as well as explicit support for comparing two or more data sets (e.g., training vs validation data, or funniness scores assigned independently by two annotators). With the prototype, the user could then examine if headlines of a certain sentiment or topic are more or less favorable across two data sets and thus detect biases in annotations. Support for these tasks is certainly interesting and can be considered part of future work; also, as stated in Section 3, annotations provided in the current Humicroedit data set cannot be traced to individual (even anonymous) annotators, which prevents the respective analysis of the existing data, unfortunately.

7. Conclusions and Future Work

This study has focused on creating a visualization prototype representing the Humicroedit data set with additional interactive functionalities and results of several additional NLP analyses of the annotated headline texts. The prototype has been evaluated in several ways, including the case study and domain expert review (RQ1). Using our proposed approach, we have discovered relationships between the annotated funniness scores and sentiment polarity, although the specific results also varied with respect to the sentiment analysis approach chosen, such as VADER vs SiBERT/RoBERTA (RQ2). No conclusive correlation between topics, keywords, and named entities vs the different annotated funniness score ranges was discovered, though (RQ3), and the same was the case for sentiment vs topics, keywords, and named entities in this data set (RQ4).

As discussed above, multiple further extensions and subsequent evaluations [KSD*22] (including task-based usability studies) are possible as part of future work for this study, including support for further computational analyses/models, workflows for editing/annotating and comparing several data sets, and visually representing and interacting with these annotations and computational results. The proposed pipeline is potentially generalizable to other tasks and data sets, especially if the number of elements used for the visual mapping (such as the max number of treemap cells) is adjusted by the user (however, for longer documents in particular, a more suitable details view would be beneficial rather than single table rows). With this study, we encourage the visualization community to contribute towards the challenges of (computational) humor analysis—but also the tasks of interactive / visual analytic support for data annotation in NLP.

Acknowledgments

This work was partially supported through (1) the ELLIIT environment for strategic research in Sweden and (2) the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. We also thank the domain experts for participating in the evaluation sessions.

References

- [ACS*24] ATZBERGER D., CECH T., SCHEIBEL W., DÖLLNER J., SCHRECK T.: Quantifying topic model influence on text layouts based on dimensionality reductions. In *Proc. of VISIGRAPP* (2024), SciTePress, pp. 593–602. doi:10.5220/0012391100003660. 4
- [ACT*24] ATZBERGER D., CECH T., TRAPP M., RICHTER R., SCHEIBEL W., DÖLLNER J., SCHRECK T.: Large-scale evaluation of topic models and dimensionality reduction methods for 2D text spatialization. *IEEE Trans. Vis. Comput. Graph.* 30, 1 (2024), 902–912. doi:10.1109/TVCG.2023.3326569. 4
- [AF23] AKKURT E., FOLDE J.: *Visual Analysis of Humor Assessment in Edited News Headlines*. Master's thesis, Linköping University, 2023. 2
- [AL19] ALHARBI M., LARAMEE R.: SoS TextVis: An extended survey of surveys on text visualization. *Computers* 8, 1 (2019). doi:10.3390/computers8010017. 2
- [EMK*21] ESPADOTO M., MARTINS R. M., KERREN A., HIRATA N. S. T., TELEA A. C.: Toward a quantitative survey of dimension reduction techniques. *IEEE Trans. Vis. Comput. Graph.* 27, 3 (2021), 2153–2173. doi:10.1109/TVCG.2019.2944182. 4
- [GPQX07] GÖRG C., POHL M., QELI E., XU K.: Visual representations. In *Human-Centered Visualization Environments*, vol. 4417 of LNCS. Springer, 2007, pp. 163–230. doi:978-3-540-71949-6_4. 3
- [Gro20] GROOTENDORST M.: KeyBERT: Minimal keyword extraction with BERT, 2020. doi:10.5281/zenodo.4461265. 3
- [Gro22] GROOTENDORST M.: BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* (2022). doi:10.48550/arXiv.2203.05794. 3
- [HG14] HUTTO C., GILBERT E.: VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proc. of ICWSM* 8, 1 (2014), 216–225. doi:10.1609/icwsm.v8i1.14550. 3
- [HHSS23] HARTMANN J., HEITMANN M., SIEBERT C., SCHAMP C.: More than a feeling: Accuracy and application of sentiment analysis. *Int. J. Res. Mark.* 40, 1 (2023), 75–87. doi:10.1016/j.ijresmar.2022.05.005. 3
- [HKG19] HOSSAIN N., KRUMM J., GAMON M.: “President vows to cut <taxes> hair”: Dataset and analysis of creative text editing for humorous headlines. In *Proc. of NAACL-HLT* (2019), ACL, pp. 133–142. doi:10.18653/v1/N19-1012. 2, 3
- [HKGK20] HOSSAIN N., KRUMM J., GAMON M., KAUTZ H.: SemEval-2020 Task 7: Assessing humor in edited news headlines. In *Proc. of SemEval* (2020), ICCL, pp. 746–758. doi:10.18653/v1/2020.semeval-1.98. 2, 3
- [HKS20] HOSSAIN N., KRUMM J., SAJED T., KAUTZ H. A.: Stimulating creativity with FunLines: A case study of humor generation in headlines. *arXiv preprint arXiv:2002.02031* (2020). doi:10.48550/arXiv.2002.02031. 2, 3
- [HW13] HEINRICH J., WEISKOPF D.: State of the art of parallel coordinates. In *Eurographics STARS* (2013), The Eurographics Association, pp. 95–116. doi:10.2312/conf/EG2013/stars/095-116. 3
- [JLH*23] JENTNER W., LINDHOLZ G., HAUPTMANN H., EL-ASSADY M., MA K.-L., KEIM D.: Visual analytics of co-occurrences to discover subspaces in structured data. *ACM Trans. Interact. Intell. Syst.* 13, 2 (2023). doi:10.1145/3579031. 2
- [KBR*12] KUKSENOK K., BROOKS M., ROBINSON J. J., PERRY D., TORKILDSON M. K., ARAGON C.: Automating large-scale annotation for analysis of social media content. In *Proc. of IVTA Workshop* (2012). 2
- [KCW*21] KHARTABIL D., COLLINS C., WELLS S., BACH B., KENNEDY J.: Design and evaluation of visualization techniques to facilitate argument exploration. *Comp. Graph. Forum* 40, 6 (2021), 447–465. doi:10.1111/cgf.14389. 6
- [KK15] KUCHER K., KERREN A.: Text visualization techniques: Taxonomy, visual survey, and community insights. In *Proc. of IEEE PacificVis* (2015), IEEE, pp. 117–121. doi:10.1109/PACIFICVIS.2015.7156366. 2
- [KPSK17] KUCHER K., PARADIS C., SAHLGREN M., KERREN A.: Active learning and visual analytics for stance classification with ALVA. *ACM Trans. Interact. Intell. Syst.* 7, 3 (2017), 14:1–14:31. doi:10.1145/3132169. 2, 3
- [KSD*22] KUCHER K., SULTANUM N., DAZA A., SIMAKI V., SKEPPSTEDT M., PLANK B., FEKETE J.-D., MAHYAR N.: An interdisciplinary perspective on evaluation and experimental design for visual text analytics: Position paper. In *Proc. of BELIV* (2022), IEEE. doi:10.1109/BELIV57783.2022.00008. 2, 5, 6
- [LOG*19] LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTMLOYER L., STOYANOV V.: RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692* (2019). doi:10.48550/arXiv.1907.11692. 3
- [MSJF20] MACQUISTEN A., SMITH A. M., JOHANSSON FERNSTAD S.: Evaluation of hierarchical visualization for large and small hierarchies. In *Proc. of IV* (2020), IEEE, pp. 166–173. doi:10.1109/IV51561.2020.00036. 4, 6
- [NA19] NONATO L. G., AUPETIT M.: Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment. *IEEE Trans. Vis. Comput. Graph.* 25, 8 (2019), 2650–2673. doi:10.1109/TVCG.2018.2846735. 4
- [RG19] REIMERS N., GUREVYCH I.: Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proc. of EMNLP-IJCNLP* (2019), ACL, pp. 3982–3992. doi:10.18653/v1/D19-1410. 4
- [Rob07] ROBERTS J. C.: State of the art: Coordinated & multiple views in exploratory visualization. In *Proceedings of the Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization* (2007), CMV 2007, IEEE, pp. 61–71. doi:10.1109/CMV.2007.20. 3
- [SHM14] SMITH A., HAWES T., MYERS M.: Hiérarchie: Visualization for hierarchical topic models. In *Proc. of ILLVI Workshop* (2014), ACL, pp. 71–78. doi:10.3115/v1/W14-3111. 6
- [Shn96] SHNEIDERMAN B.: The eyes have it: A task by data type taxonomy for information visualizations. In *Proc. of IEEE VL* (1996), IEEE, pp. 336–343. doi:10.1109/VL.1996.545307. 3
- [Spa23] SPACY: Industrial-strength natural language processing. <https://spacy.io/>, 2023. 3
- [STLD20] SCHEIBEL W., TRAPP M., LIMBERGER D., DÖLLNER J.: A taxonomy of treemap visualization techniques. In *Proc. of VISIGRAPP* (2020), SciTePress, pp. 273–280. doi:10.5220/0009153902730280. 3
- [VdMH08] VAN DER MAATEN L., HINTON G.: Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 11 (2008). 4
- [WMW*22] WANG X., MING Y., WU T., ZENG H., WANG Y., QU H.: DeHumor: Visual analytics for decomposing humor. *IEEE Trans. Vis. Comput. Graph.* 28, 12 (2022), 4609–4623. doi:10.1109/TVCG.2021.3097709. 2
- [ZYQ*08] ZHOU H., YUAN X., QU H., CUI W., CHEN B.: Visual clustering in parallel coordinates. *Comp. Graph. Forum* 27, 3 (2008), 1047–1054. doi:10.1111/j.1467-8659.2008.01241.x. 3