Visual Analysis of Stance Markers in Online Social Media

Kostiantyn Kucher*
Linnaeus University, Sweden

Andreas Kerren†
Linnaeus University, Sweden

Carita Paradis‡
Lund University, Sweden

Magnus Sahlgren§
Gavagai AB, Sweden

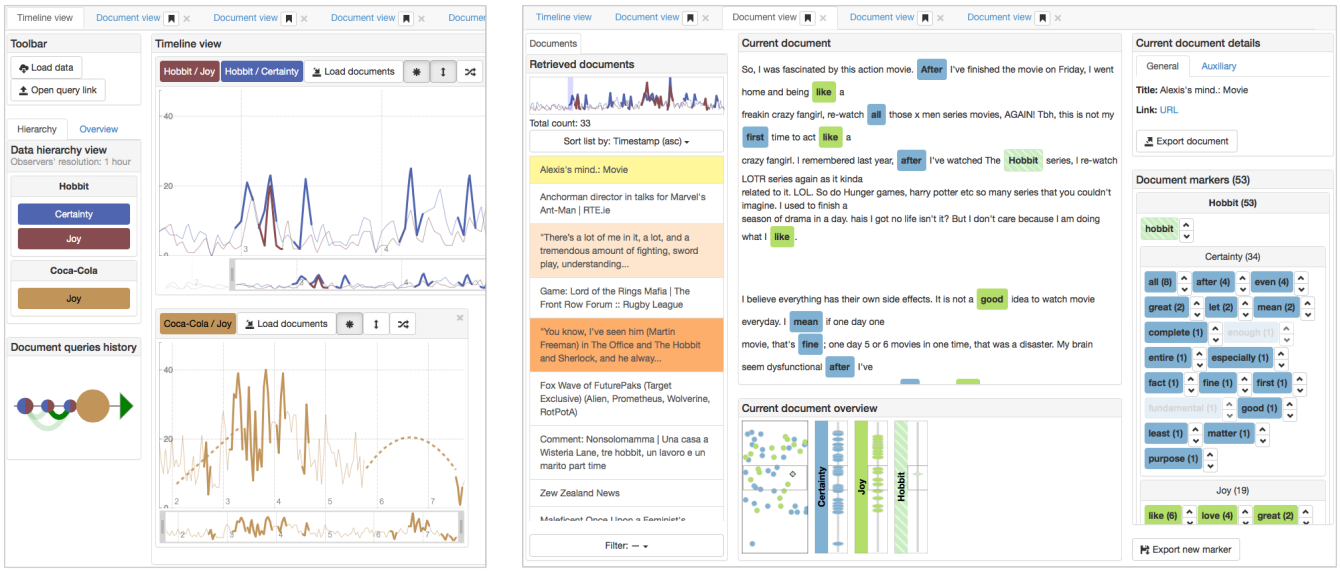

Figure 1: The left screenshot of our tool shows the *timeline view*: users start by loading time-series data associated with certain targets (here 'Hobbit' and 'Coca-Cola') and observers ('certainty' and 'joy'), explore the data and select time intervals for specific observer combinations to fetch URI links for. The right screenshot displays our *document view* for a selected time range and the observer combinations 'Hobbit/Joy' and 'Hobbit/Certainty': here users can explore the resulting set of documents and are able to analyze the contents of specific documents with regard to the occurrence of stance markers summarized on the right hand side.

Abstract

Stance in human communication is a linguistic concept relating to expressions of subjectivity such as the speakers' attitudes and emotions. Taking stance is crucial for the social construction of meaning and can be useful for many application fields such as business intelligence, security analytics, or social media monitoring. In order to process large amounts of text data for stance analyses, linguists need interactive tools to explore the textual sources as well as the results of computational linguistics techniques. Both aspects are important for refining the analyses iteratively. In this work, we present a visual analytics tool for online social media text data and corresponding time-series that can be used to investigate stance phenomena and to refine the so-called stance markers collection.

Keywords: Visualization, text visualization, interaction, time-series, stance analysis, sentiment analysis, NLP, text analytics

1 Introduction

The vast amount of digital data available nowadays on the Internet provides unprecedented opportunities for automated analyses. For example, text data of all kinds makes it possible for researchers in the field of linguistics to employ a bottom-up approach to understand various aspects of language. The research on specific phenomena benefits from text data collected from certain domains such as online social media (Twitter, Facebook, blogs, forums, etc.). Those texts are typically created by multiple authors engaged in discussions in which they express their thoughts and opinions.

This presents an opportunity for researchers interested in stance analysis. *Stance* is a relatively ambiguous concept in linguistics [1] that might be operationalized as subjectivity expressed in human communication (e.g., attitudes, feelings, perspectives, or judgements). Research on stance includes both theoretical efforts (in order to crystallize the definition and the knowledge about the nature of this phenomenon) and practical efforts (in order to collect evidence and to explain the means of stance being expressed) and can lead to various text analytics applications. The practical tasks require processing large quantities of textual data that is unsuitable for manual investigation. Therefore, stance researchers are interested in taking advantage of the automated ways of text processing that can be offered by researchers from the field of natural language processing (NLP).

However, many linguists face difficulties when trying to interpret the output of NLP algorithms. Equally challenging for them is how to gain insight into the data and provide useful feedback in order to refine their analyses. In fact, NLP researchers would also benefit from a technique that could improve their understanding of the computational processes associated with the state-of-the-art NLP algorithms. This predicament can be resolved by introducing a Visual Analytics (VA) approach to provide linguistics researchers with interactive visualizations for analyzing large text data and for presenting the NLP experts with feedback at the same time. Our research project StaViCTA (Advances in the description and explanation of Stance in discourse using Visual and Computational Text

*e-mail: kostiantyn.kucher@lnu.se
†e-mail: andreas.kerren@lnu.se
‡e-mail: carita.paradis@englund.lu.se
§e-mail: mange@gavagai.se

IEEE Symposium on Visual Analytics Science and Technology 2014
November 9-14, Paris, France
978-1-4799-6227-3/14/$31.00 ©2014 IEEE

Analytics) [6] addresses this research challenge and aims to produce a refined theory of stance, efficient interactive visualization and computational techniques for its analysis, as well as solutions for specific applications.

One of the steps towards the refined theory and sophisticated NLP models for stance analysis is a training data set consisting of text fragments (n-grams) that are associated with specific expressions of stance. We refer to these n-grams as *stance markers* and collect our training data set by analyzing online social media texts (in English). In the rest of this poster abstract, we present our tool called *uVSAT* that is designed to help stance researchers to identify candidate documents that may contain stance expressions, analyze the document texts and export the new stance markers.

Background and Related Work Since our tool focuses on the analysis and collection of specific stance markers, we limit the scope of our understanding of stance to a set of categories that describe the attitudinal, emotional or epistemic saturation of text. This approach can be described as multidimensional sentiment analysis.

While there have not been many efforts of automated stance analysis in the field of computational linguistics, sentiment analysis is well-described, and corresponding techniques have been applied for analyses involving multiple sentiment categories. Our approach is based on a particular semantic analysis technique, *random indexing* [5], that allows us to construct a vector space model for large document collections with "seed words" (the markers in our terminology) used as features. We combine this technique with simple lexical analysis to identify documents relevant to specific stance categories and to localize the markers in the text.

From the VA perspective, our tool has been designed to visualize and interact with actual text data as well as text processing results which include time-series. There have recently been multiple works dedicated to text visualization and analytics of social media that address the temporal aspect and even sentiment [4], though not stance analysis yet. With respect to the main purpose of our tool, i.e., the refinement of stance markers collections, we should note the works [2, 3] that facilitate the sentiment lexicon collection process.

2 Data Model

Our web-based VA tool is currently designed to use time-series data provided by our collaboration partners at the company Gavagai who monitor and process online social media documents with a computational framework based on random indexing. To detect the documents associated with stance, we are currently using seed words relevant to stance (i.e., stance markers) from several available sources (WordNet, GeneralInquirer, and CompassDeRose). The data consists of timestamped numerical values that describe the aggregated polarity of specific *observers* (i.e., kinds of stance, like 'joy' or 'aggression') for specific *targets* (i.e., monitoring/search entries, like 'weapons' or 'diet') with a time resolution of one hour. These values do not directly correspond to the amount of our expected markers, but they describe the tendencies, so further exploration of original text data is required from the users. We are also provided with the URIs of documents used to calculate the polarity scores, though the corresponding HTML content has to be downloaded and processed on our side.

3 Visualization and Interaction Approaches

Fig. 1 displays our tab-oriented user interface with two types of tabs: a single timeline view tab that is used to work with an arbitrary number of timeline plots, and multiple document view tabs that are opened by the user when fetching the document URIs.

uVSAT uses a standard line plot representation for time-series data and supports usual interaction techniques for such plots (overview, scroll&zoom, filtering). It supports multiple plots displayed on the same canvas or separately. For the comparison of several plots displayed side by side, the user can control the automatic vertical scaling. To facilitate the user with the search for *regions of interest (ROI)*, the tool supports automatic ROI highlighting based on outliers/differential analysis. We also support trend analyses by displaying linear or quadratic trend lines as overlay plots or instead of selected plot segments. To enable efficient data exploration and record the analysis process, we have implemented a history control with an interactive diagram that provides an overview of the document URI queries sequence, their results and relations. Finally, users can select time intervals for specific sets of observers and load the corresponding URI links. In this case, a new document view tab is created, and a thumbnail of the line plot used for the query is displayed in order to preserve the mental map.

A document view tab presents the user with the list of URI links. By selecting a link, the corresponding document content is fetched, processed and displayed including its text representation with highlighted stance markers, the hierarchical marker representation (used for statistics, filtering, navigation and brushing), and the document overview (including a 2D overview of marker distributions and overviews for each kind of markers with respect to its distribution over the entire document length) coordinated with the other views. The user can explore the content of the document, export the whole document with highlighted markers as a static HTML file, and export new stance markers by selecting text fragments and assigning them with arbitrary tags. uVSAT even provides an opportunity to copy the query link for a given document view tab and to use it in later analysis sessions by opening a tab with identical contents.

4 Conclusion

uVSAT contains multiple approaches for analyzing temporal and textual data as well as exporting stance markers in order to prepare a stance-oriented training data set. We already use this tool in our interdisciplinary StaViCTA project, and we hope that it will also be useful for other researchers. Future work includes additional overview and navigation techniques for document sets, support for local database caching, streaming data, uncertainty tackling (with regard to missing time-series data as well as unavailable web documents), and arbitrary time-series data sources.

Acknowledgements This research was funded by the framework grant "The Digitized Society – Past, Present, and Future" with No. 2012-5659 from the Swedish Research Council. We also want to thank Teri Schamp-Bjerede and Afshin Rahimi for their valuable comments and input.

References

[1] R. Englebretson. Stancetaking in discourse: An introduction. *Stancetaking in Discourse: Subjectivity, Evaluation, Interaction*, pages 1–25, 2007.

[2] M. L. Gregory, N. Chinchor, P. Whitney, R. Carter, E. Hetzler, and A. Turner. User-directed sentiment analysis: Visualizing the affective content of documents. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, SST '06, pages 23–30, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[3] R. Makki, S. Brooks, and E. E. Milios. Context-specific sentiment lexicon expansion via minimal user interaction. In *Proceedings of the International Conference on Information Visualization Theory and Applications (IVAPP)*, pages 178–186. SciTePress, 2014.

[4] C. Rohrdantz, M. C. Hao, U. Dayal, L.-E. Haug, and D. A. Keim. Feature-based visual sentiment analysis of text document streams. *ACM Trans. Intell. Syst. Technol.*, 3(2):26:1–26:25, Feb. 2012.

[5] M. Sahlgren. An introduction to random indexing. In *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE*, volume 5, 2005.

[6] StaViCTA. Project homepage. `http://cs.lnu.se/stavicta/`, 2014.