New perspectives on gathering, vetting and employing Big Data from online social media : An interdisciplinary approach

Schamp-Bjerede, Teri; Paradis, Carita; Kucher, Kostiantyn; Kerren, Andreas; Sahlgren, Magnus

2015

New perspectives on gathering, vetting and employing Big Data from online social media: An interdisciplinary approach

Teri Schamp-Bjerede,[1] Carita Paradis,[1] Kostiantyn Kucher,[2] Andreas Kerren,[2] and Magnus Sahlgren[3]

[1]Centre for Languages and Literature Lund University, Lund, Sweden, [2]Department of Computer Science, Linnaeus University, Växjö, Sweden, and [3]Gavagai AB, Stockholm, Sweden

Accepted abstract for paper presentation at the 36th *International Computer Archive of Modern and Medieval English - ICAME Conference.* 27 – 31 May, 2015, Trier, Germany.

Massive textual data sets available in online social media are valuable resources for research in linguistics compared to static corpora in that they offer the possibility of analyzing the temporal unfolding of communication in addition to a wide variety of forms, styles and topics. By analyzing such data, linguists are able to investigate the ongoing evolution of language use in text and discourse. There are however also disadvantages with big data sets of this kind, such as problems of navigation and selection of sample texts, lack of annotation, abundance of neologisms and ungrammatical expressions. In order to be able to make sensible and optimal use of these data sets, the analysis requires advanced computational methods—both automated and semi-automated (from basic concordances and to more complex natural language processing)—accompanied by visualization techniques that are needed both for the visualization itself and for the interpretation of the results.

The purpose of this paper is to present a use case to demonstrate our approach to Big Data research using insights from linguistics, natural language processing, and information visualization/visual analytics (Thomas & Cook 2006). The use case is concerned with the analysis of social media texts related to the movie series "The Hobbit". The data selected for analysis consists of microblog and forum posts before and after the release of "The Hobbit: The Desolation of Smaug" in December 2013.

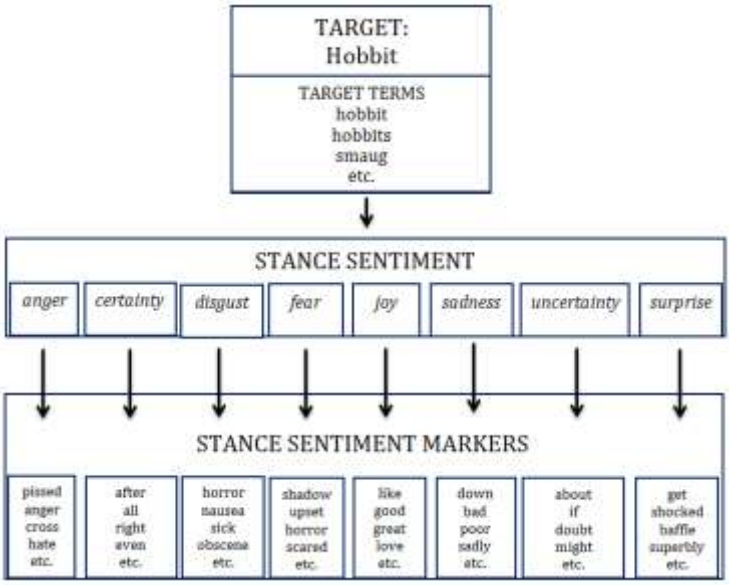

Figure 1 uVSAT logic map

At this point in time, our visual analytics tool uVSAT (Kucher et al. 2014) is designed to use a sentiment analysis approach (Figure 1). We use the sentiment markers (words), expressing six core emotions (Ekman 1992) and two epistemic stance categories, CERTAINTY / UNCERTAINTY, to identify the interlocutors' degree of commitment and attitudes to what they are discussing and in relation to the sentiments expressed. Through the use of such markers, we are able to identify and later analyze when heated discussions occur, how they occur in the flow of communication and in relation to what events in the world, as well as in the discussion itself among the participants. This will give us an idea about exactly what words and constructions are used to express the different sentiments and stances towards what is talked about, when, and where, and by whom (Du Bois 2007). That information can then be used to interact with the data in various ways. By using our visual analytics tool, we are able to analyze the temporal developments in social media, fetch the sets of relevant HTML documents, analyze the distribution of markers in the retrieved data, explore the text content with assistance of multiple interactive techniques, and export new markers as well as processed documents (Figure 2). The temporal trends as well as the distribution of sentiment and stance markers pertaining to the various categories, anger, joy, surprise, or certainty, etc., can be used to make predictions about sentiment and stance expressions in social media that will occur with regard to, in this particular use case, the subsequent movie "The Hobbit: The Battle of the Five Armies". After the premiere in December 2014, we will analyze the actual text data and compare the outcome of that with our predictions.

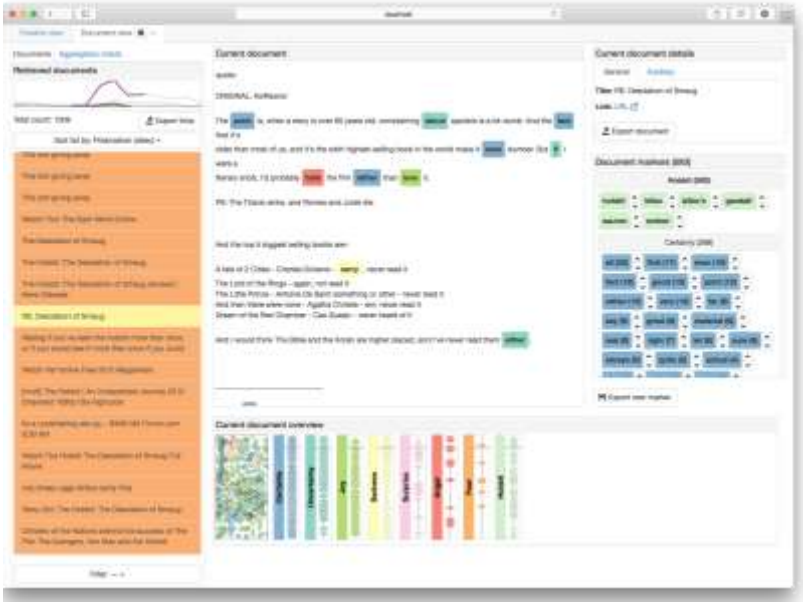

Figure 2 Document view

The contributions of this work to the corpus community include: (i) the introduction of a multidisciplinary approach to sentiment and stance analysis involving social media text data instead of static corpora, (ii) the description of the combination of visualization and interaction techniques that facilitate the linguistic analysis of textual data, and (iii) a use case demonstrating how insights can be gained from large amounts of social media texts through these techniques.

References

Du Bois, J.W. (2007). The Stance Triangle. In R. Englebretson (eds.) *Stancetaking in Discourse: Subjectivity, Evaluation, Interaction*. Amsterdam: Benjamins, 139-182.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, 6(3-4): 169-200.

Kucher, K., Kerren, A., Paradis, C., & Sahlgren, M. (2014, November). Visual Analysis of Stance Markers in Online Social Media. Poster session presented at *IEEE Visual Analytics Science and Technology (VAST'14)*, Paris, France.

Thomas, J. J., & Cook, K.A. (2006). A visual analytics agenda. *Computer Graphics and Applications, IEEE*, 26(1): 10-13.