Contrastive Co-occurrence Analysis on Twitter for the German Election 2013

Uli Fahrer
uli.fahrer@googlemail.com

Supervised by: Prof. Dr. Chris Biemann
Kind of work: Bachelor Thesis

Abstract: This paper describes an approach based on word co-occurrence that contrasts two separate keywords regarding their strongly associated words. This approach is used to investigate how real-world events are reflected in Twitter. Furthermore we present an HTML Dashboard that allows real-time interaction to explore and visualize the data for analyses. Based on a case study about the German election, we perform a contrastive analysis that shows differences and commonalities between two politicians. Results show that the overlap in our analysis is an indicator for key political events. We also found that the Twitter stream reflects real-world events well, and is in high accordance with the daily press.

1 Introduction

The Internet changed over the years in terms of how people interact with each other. Online social networking and microblogging services like Twitter were key factors in this transition. Recent uprisings in the Ukraine or the Arab Spring have shown that Twitter is a tool for sharing information about the protests with the rest of the world. This raises the question of how real-world events are reflected in Twitter. We present an approach based on word co-occurrence that enables us to retrieve words that are strongly associated with a particular keyword and to perform a contrastive analysis on two separate keywords. Besides the investigation of only one keyword, the contrastive analysis can be used to show the differences and commonalities between two keywords and how they are reflected in Twitter. The paper tackles the following research questions:

(1) How do two given keywords differ with respect to their strongly associated words?
(2) How does Twitter reflect real-world events?

We are conducting our research on a case study about the German federal election, which took place on 22nd September 2013.

2 Related Work

Much research has already been done on word association measures in natural language texts. Dunning [Dun93] introduced the log-likelihood measure and Church and Hanks [CH89] the point mutual information measure. Also see Evert and Krenn [EK01] for an overview and an evaluation of co-occurrence measures. Tumasjan et al. [TSSW10] performed a sentiment analysis on the German election 2009 and found that the sentiment of Twitter messages closely corresponds to candidate profiles. Blenn et al. [BCD12] presented a polarization analysis of commonly used words with two keywords based on general Twitter messages. Their system arranges the resulting associated words according to their overall polarity strength. We combine statistical significance measures with this approach to contrast two keywords with respect to their strongly associated words.

3 Data Acquisition and Preprocessing

The data we use in our study was collected from Twitter between August 2, 2013 and October 9, 2013, which includes the German election 2013. We used the Python Tweepy module[1] as an interface for the Twitter Search API[2], where we set the language parameter to German. Further, we defined the 6 parties represented in the German parliament with their top candidates as search terms. Overall, we collected a corpus of 6,163,367 tweets. For the tokenization, we employed the Twitter tokenizer from Owoputi et al. [OOD+13]. We also removed function words, as well as punctuation from the output. In addition we employed the unsupervised POS tagging system from Biemann [Bie06] to annotate the tokens with word classes. Based on a word list consisting of all electable German politicians we also tag words as named entities. To determine the words strongly co-occurring with a given word, we use the log-likelihood measure [Dun93]. We apply this measure to rank the vocabulary according to descending values. For the representation we use a weighted and undirected co-occurrence graph $G(V, E)$, where each vertex $v(\mathrm{id}, \mathrm{freq}, \mathrm{word\_class}) \in V$ is a triple. The triples consist of a unique identifier used to represent the word, the frequency of the word in the corpus and its word class respectively. Each edge $e(w_1, w_2, \mathrm{likelihood}, \mathrm{weight}) \in E$ is a 4-tuple consisting of the two connected words, the significance measure and the edge weight. In addition, we index tweets by words and their co-occurrences to be able to display the reason for the association.

Figure 1: Context: Showing tweets that contain keywords Brüderle, Trittin and #dreikampf, as queried from Figure 2.

[1] http://pythonhosted.org/tweepy/html/
[2] https://dev.twitter.com/docs/api/1/get/search

4 Visualization

We designed an HTML dashboard, which allows real-time user interaction and offers several parameters to affect the generation of the co-occurrence chart:

- **Keywords:** One or two keywords that are used to query the system.
- **Measure:** Variation of the statistical significance measures e.g likelihood.
- **Minimal edge weight:** A threshold on the edge weight that specifies how often the given keyword and its co-occurring words have to occur together in the corpus.
- **Display limit:** Sets the number of displayed words associated with the keywords.
- **Include tweets after date:** Check to include Twitter posts after the election day.
- **Named entities only:** Check to visualize only named entities.
- **Part-of-speech option:** Select word classes that should be included in the chart.

Figure 2 shows a chart for the keywords *Brüderle* and *Trittin* , which are both top candidates for minor parties. We removed user names and excluded tweets after the election day. The left side of the graph shows words only co-occurring with the keyword *Brüderle* and the right side only co-occurring words with *Trittin* . The overlap in the middle indicates words that are co-occurring with both terms. We call this a contrastive analysis. To obtain the context related to the given keyword and a particular co-occurring word the user can select the context view. Figure 1 represents a context for the co-occurring word #dreikampf.

5 Case Study

The fact that Twitter is used as a platform for political deliberation does not necessarily mean that meaningful information can be extracted, or that the distribution of opinion tweets reflects the distribution of opinions in the population. To investigate how Twitter reflects real-world events, we show a contrastive analysis with the keywords *Brüderle* and *Trittin* to examplify the capabilities of our software (see Figure 2). The words in the overlap are related to the clash between both politicians during a TV duel that was being discussed on Twitter under the hashtag #dreikampf. The third politician in the three-way-battle was *Gregor Gysi*, whose name is also found in the overlap. This indicates that the overlap might be a reflection for key political events. In order to compare the results to the daily press, we use the "Wörter des Tages"[3] platform from the University of Leipzig, which shows terms that are particularly relevant for a day with respect to different daily newspapers. From these terms we identified relevant topics that match our keywords. We found out that about 60% of the co-occurring words related to *Brüderle* are reflected in these topics. For the keyword *Trittin* we even found an overlap of about 70%. As this small example demonstrates, the Twitter stream reflects real-world events well, and is in high accordance with the daily press. We found similar ranges of overlap for other events such as the election, the Stinkefinger affair, the pedophilia discussion, and many more.

[3] http://wortschatz.uni-leipzig.de/wort-des-tages/

6 Conclusion and further research

We presented an approach that visualizes co-occurring words in Twitter messages in an intuitive and flexible way that allows contrastive analysis. With this approach we analyzed over 6 million tweets mentioning parties or politicians in the weeks leading up to the German election 2013. Overall, we found out that the overlap in our contrast analysis gives a sensible reflection of key political events, and that associations characterize single words very well. In addition most of the relevant newspaper topics regarding our contrastive analysis are reflected in Twitter. In our further work, we aim to extend our system to investigate how particular opinions (word-associations) change over time. In addition, it would be a natural extension to also include documents that are linked to from tweets into our corpus.

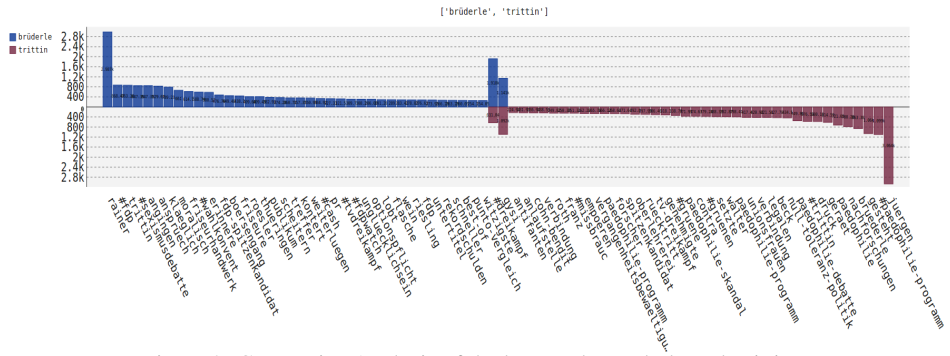

Figure 2: Contrastive Analysis of the keywords Brüderle and Trittin.

References

[BCD12] Norbert Blenn, Kassandra Charalampidou, and Christian Doerr. Context-Sensitive Sentiment Classification of Short Colloquial Text. In *Proc. IFIP'12*, pages 97–108, Prague, Czech Republic, 2012.

[Bie06] Chris Biemann. Unsupervised Part-of-speech Tagging Employing Efficient Graph Clustering. In *Proc. ACL/COLING-2006 SRW*, pages 7–12, Sydney, Australia, 2006.

[CH89] Kenneth W. Church and Patrick Hanks. Word Association Norms, Mutual Information, and Lexicography. In *Proc. ACL-1989*, pages 76–83, Vancouver, Canada, 1989.

[Dun93] Ted Dunning. Accurate Methods for the Statistics of Surprise and Coincidence. *Comp. Ling.*, 19(1):61–74, 1993.

[EK01] Stefan Evert and Brigitte Krenn. Methods for the Qualitative Evaluation of Lexical Association Measures. In *Proc. EACL-2001*, pages 188–195, Toulouse, France, 2001.

[OOD+13] Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL-HLT*, pages 380–390, 2013.

[TSSW10] Andranik Tumasjan, Timm Sprenger, Philipp Sandner, and Isabell Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proc. AAAI-2010*, pages 178–185, Atlanta, GA, USA, 2010.