OPINION MINING IN WEB-SCALE CORPORA

Timothy O'Keefe
Irena Koprinska and James Curran
ə-lab, School of Information Technologies
tokeefe@it.usyd.edu.au, irena@it.usyd.edu.au, james@it.usyd.edu.au

1. Introduction

The opinions that people hold about other individuals, products, political parties, businesses and many other entities have always been of interest to us. Over the last decade there has been increasing interest in a new set of techniques which fall under the terms *opinion mining* and *sentiment analysis*. When defined broadly, opinion mining and sentiment analysis refer to the same basic task, which is to automatically find and classify opinions in text as being positive or negative. When combined with the massive amount of opinionated text available on the world wide web, these techniques offer a cheap method of determining the prevailing opinion about a wide range of possible entities.

2. Why is opinion mining important?

Simple narcissism aside, there are a range of cases where the opinions that people have about an entity are extremely important. Businesses spend large sums of money on market research to determine whether or not a given product is popular, while political parties conduct similar research to determine the opinions voters have about particular policies. In both of these situations researchers are heavily dependent on focus groups and surveys, which, relatively speaking, are slow, expensive and usually suffer from a small sample size. This limits their usefulness to products that have a large enough market that the cost of performing the research is actually worth the benefits that would flow from it. This means that for many smaller products or personalities, market research is simply out of the question.

By contrast, opinion mining requires little or no human interaction beyond a search term, and can be used on as much data as is available or that can be searched in the time allowed. This means that it is not only useful in situations where surveys are not, but can also be used to augment the information obtained from surveys, and can be used far more often than surveys. As such, opinion mining looks set to become a valuable tool in market research, political campaigns, product recommendation systems, and potentially other, unknown applications.

3. How does it work?

Opinion mining can be conducted at several levels of granularity. The levels of granularity are:

- **Term-level:**

  The goal of term-level work is typically to build a sentiment-annotated dictionary that can be used as an input to other tasks. These dictionaries have been built using various graph [2, 3, 1] and other methods [6, 4]. In particular it is common to find positive and negative terms by finding words with similar definitions

- **Sentence-level:**

  Sentence-level sentiment analysis is the most difficult level of granularity, as it is attempting to identify sentiment at about the granularity at which it is expressed. Work in identifying opinions at the sentence level is extremely varied, with the simplest methods summing the sentiment of the terms within the sentence.

- **Document-level:**

  At the document-level the most common approach is to treat sentiment analysis as a special case of topic-based categorisation, with the classes *positive* and *negative*. Typically a unigram bag-of-words model is employed [5].

4. Work so far

During the 9-month period between February and October I have worked on four major areas of research, as well as one minor problem, and a reasonably extensive literature review. The first project was a continuation of my honours research into document-level sentiment analysis that involved using ensembles of classifiers. The second project was an alteration to the Naïve Bayes equation, which could allow the Naïve Bayes method to scale up to hundreds of millions or more features. The aim of the third project was to discover cycles in the definitions of terms in WordNet, so that a rigorous term-level sentiment lexicon could be built. The final project uses a method related to the work of Hatzivassiloglou & McKeown [4] to build a sentiment lexicon.

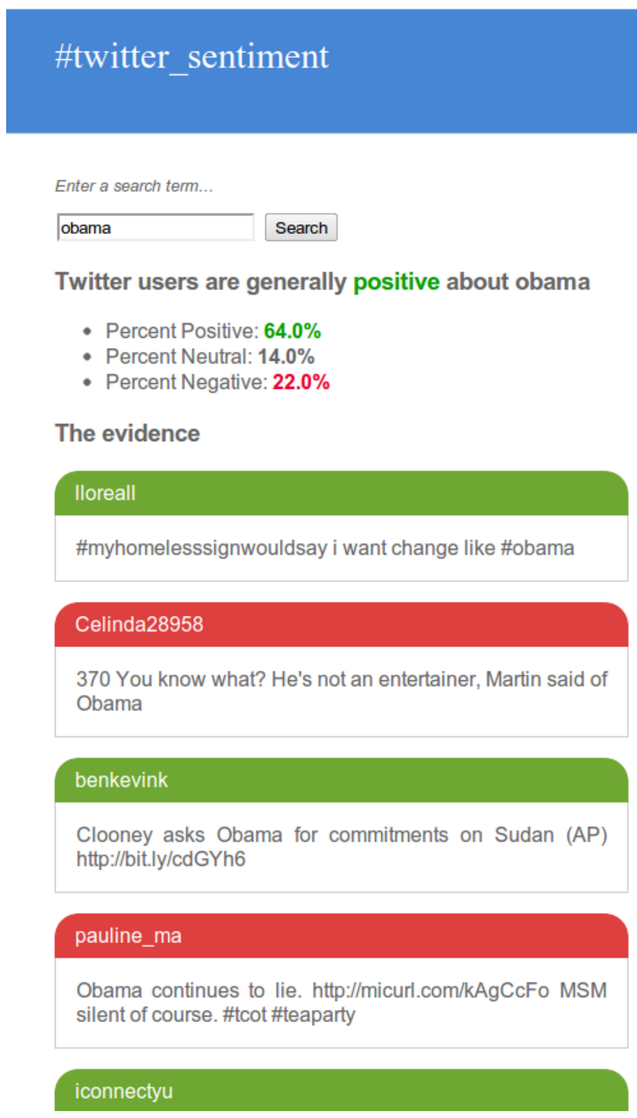

Figure 1: *Small-scale demonstration system using Twitter as a data source.*

5. Future work

The work that I plan to complete can be divided up into four main sections. These are:

- To complete the work that I have already begun on a sentiment-annotated dictionary
- To build a sentence-level sentiment classification system that uses the term-level dictionary
- To use the previous two components in a document-level classification system
- To integrate these components with a data retrieval component, which will allow the system to mine opinions from the web

6. Conclusion

Although opinion mining as a field is relatively new, there has been intense research interest, with a large number of publications discussing a range of approaches. Despite this large body of research, there is a lack of research into web-scale sentiment analysis, as well as a lack of an integrated system that can operate on multiple levels of granularity. I propose to address this shortcoming through a programme of research that will see a web-scale, multiple-granularity system produced by 2013. Such a system will enable fast, timely, and cheap access to the combined opinions of potentially millions of people, which will aid businesses, political parties, journalists, and others in understanding and acting upon those opinions.

References

[1] S. Baccianella, A. Esuli, and F. Sebastiani. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Seventh conference on International Language Resources and Evaluation, Malta. Retrieved May*, volume 25, 2010.

[2] A. Esuli and F. Sebastiani. Determining the semantic orientation of terms through gloss classification. In *Proc. of the ACM SIGIR Conf. on Information and Knowledge Management CIKM 2005*, pages 617–624. ACM, 2005.

[3] Andrea Esuli and Fabrizio Sebastiani. Pageranking wordnet synsets: An application to opinion mining. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 424–431, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[4] V. Hatzivassiloglou and K. R McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, page 181, 1997.

[5] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 79–86, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[6] H. Takamura, T. Inui, and M. Okumura. Extracting semantic orientations of words using spin model. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, page 140, 2005.