Sentiment Analysis and Visualization in a Micro-Blogging Service

A Thesis Proposal
Presented to
the Faculty of the College of Computer Studies
De La Salle University Manila

In Partial Fulfillment
of the Requirements for the Degree of
Master of Science in Computer Science

by

Tensuan, Juan Paolo

Arnulfo AZCARRAGA
Adviser

April 11, 2012

Abstract

Automated Sentiment Analysis is a study in the field of Computer Science that has gained much traction because of its utility value in the social sciences, marketing, and commerce. While the use of Sentiment Analysis techniques has long been popular and effective with static content like product reviews, the usage of such in micro-blogging sites have proven to be a different challenge altogether. With the popularization of Twitter and other micro-blogging sites, challenges in changing semantics and the usage of social data have become an interesting discourse for many. Likewise, the high volume context-rich data that micro-blogging sites present have made the study even more relevant. This research work aims to look at how Sentiment Analysis can be done in a micro-blogging service like Twitter and at the same time present the results in an easily comprehensible manner. Particularly, different algorithms for sentiment analysis will be used, from which visualization techniques will also be implemented and modified to make easily comprehensible graphic data.

Keywords: Sentiment Analysis, Visualization, Data Mining, Micro-blogging, Twitter

Table of Contents

1 Research Description

In this chapter, the researcher will give an overview of the research to be done. This includes the research objectives, methodology, scope, and limitations.

1.1 Overview of the Current State of Technology

Micro-blogging has been gaining popularity as a medium to share information through websites like Twitter, Plurk, and Facebook. Particularly, it differentiates itself from traditional web blogging services because people are forced to keep their posts concise, under a certain character limit, making posts easier for others to read. Likewise, the short-post limitation has encouraged people to post ideas in short posts frequently as it is the norm for micro-blogging services as opposed to traditional blogs where posts can be as long as a thousand words or more.

With the popularization of Internet-enabled mobile devices and operator-powered micro-blogging services, people have become generally much more engaged in micro-blogging. In fact, 37 percent of Twitter users have tweeted with their phones (Bifet & Frank, 2010). People can now not only post their ideas through desktop computers, but they can now also post anywhere through mobile devices.

Because of the ease and convenience provided by Twitter, Twitter now contains millions of thoughts encapsulated in short posts. Every single day, Twitter would naturally receive hundreds and thousands of these posts which contain many ideas which range from people's thoughts on current events, certain products, politics, and more.

Many research works have capitalized on the sheer amount of data made available through micro-blogging sites like Twitter. Twitter has been the focus of many researches in detecting emerging topics of interest(Michelson & Macskassy, 2010), events sparking controversy (Popescu & Pennacchiotti, 2010), sentiment-analysis (Bifet & Frank, 2010)(Tan et al., 2011)(Calais Guerra, Veloso, Meira, & Almeida, 2011), and even language learning (Michelson & Macskassy, 2010) because of its huge user base and simplicity. Aside from being just a micro-blogging service, Twitter has useful data with regards to user's social relationships with other users as it doubles as a Social Networking site. Particularly, there has been a lot of interest in sentiment-analysis with Twitter because of the sheer volume of opinion-data from millions of users (Bifet & Frank, 2010) in the Micro-blogging service.

Sentiment-analysis in the Micro-blogging service proves to be a different scenario. Twitter's integrated social data gives way to novel approaches (Calais Guerra et al., 2011)(Bifet & Frank, 2010) that take advantage of this. Likewise, Twitter's real-time data is by nature unbalanced (Bifet & Frank, 2010), and it's changing vocabulary over time renders Natural Language Processing techniques less powerful (Calais Guerra et al., 2011).

1.2 Research Objectives

1.2.1 General Objective

To be able to automatically detect and visualize the overall sentiment of topics from a real-time text-stream like Twitter.

1.2.2 Specific Objectives

1. To detect the sentiment of people on certain topics.
2. To evaluate the effectiveness of different sentiment analysis algorithms with a real-time text stream
3. To visualize the sentiment of people on certain topics throughout a time period
4. To compare different sentiment analysis algorithms and visualization techniques

1.3 Scope and Limitations of the Research

This research will specifically be working with Twitter data primarily because of the ease of retrieval of such data and the contextual information available in the data. Because of this, the volume of posts in the dataset will also be limited to the volume that Twitter would allow. Likewise, data will be limited to public tweets in Twitter.

The research be using different algorithms for sentiment analysis on a real-time stream like Twitter's. After which, these algorithms will be evaluated through various Artificial Intelligence measures. Twitter data to be used will be the posts themselves along with the Re-Tweets, Hash Tags, User Mentions, and URLs. User profiles along with the relationships between users may also be used.

Although the creation of a crawler for Twitter may be necessitated by the research, it will not be the focus of this research.

1.4 Significance of the Research

Although there have been many researches done on sentiment-analysis on real-time text streams like Twitter, there is still much room for improvement in the specific field. There are still various modifications that can be done to these algorithms to improve the quality of their results.

Another equally important aspect that this research looks at is the effective visualization of data from these algorithms. Visualization of the results can prove helpful in analyzing the results for both those in the field of computing and those in the social sciences.

Besides being beneficial to the sciences, this research could also prove to be commercially beneficial. The research would be able to aid marketing firms by providing them with a more comprehensive view of specific topics along with their associated sentiments along a given time period in sites like Twitter.

2 Review of Related Literature

In this chapter, the researcher will discuss various research works that are useful references to the task to be done in this research work. Topics ranging from Sentiment Analysis and Topic Detection, particularly in real-time text streams such as Micro-blogging Sites (like Twitter) will be discussed. These topics will be discussed in brief with the purpose of seeing how these research works may be used in the research work.

2.1 Sentiment Analysis in Social Networks

According to Bifet and Frank (2010), there are two methods of sentiment analysis that can be used in a social network like Twitter: analysing sentiment through the text data itself (e.g. using the sentiment associated with specific keywords in the posts) and analysing sentiment through the context of the posts (e.g. the links of these messages to other messages). This section will give a brief overview of both approaches. Likewise, the advantages of each approach will be discussed.

2.2 Sentiment Analysis using Actual Text Data

There are many methods by which a post's textual content can be used to analyse sentiment. Most, however, has to do with using a sentiment lexicon. Such lexicon should contain information such as the keywords that the algorithm would look out for and the polarity of these words(Popescu & Pennacchiotti, 2010).

There are many ways by which researchers make sentiment lexicons. Various methods of making a sentiment lexicon would include: the usage of machine learning algorithms and examining relationships among adjectives, clustering of the adjectives used in posts, using web-based information on specific adjectives used in posts and many more. (Neviarouskaya, Prendinger, & Ishizuka, 2011)

Using actual text data is more flexible as it is more readily available than context data.

2.2.1 Sentiment Knowledge Discovery using Various MLP Techniques

In a work done by Bifet and Frank (2010), text data was analyzed specifically by the presence of specific keywords (as opposed to the frequency of keywords) in a text stream. Bifet and Frank (2010) used Multinomial Naive Bayes, Stochastic Gradient Descent, and Hoeffding Tree on 2 different corpora: a corpora from twittersentiment.appspot.com and the Edinburgh Corpus.

The twittersentiment.appspot.com corpus training data was labelled through the emoticons found in the posts, whereas the test data was composed of posts with and without emoticons manually labelled by annotators. It has 1.6 million text posts, equally distributed to positive and negative sentiment. It contains the post sentiment, date, query used, user, and the actual text data of each post.

The Edinburgh Corpus, meanwhile, is a corpus with 97 million tweets. It contains the timestamp, user name (anonymized), actual text, and posting method of each post.

Bifet and Frank (2010) recommended that Stochastic Gradient Descent be used for Twitter data and warned against the usage of tree-based learning algorithms such as the Hoeffding Tree. While the Naive Bayes performed the best, Bifet and Frank (2010) argued that the Stochastic Gradient Descent method is better because of interpretability of data. Bifet and Frank (2010) was able to how the sentiment of people on different brands changed through time using the Stochastic Gradient Descent.

2.3 Sentiment Analysis using Context Data

Using text-based analysis can be problematic with dynamically changing content such as Twitter. Labelled data can be limited and the dynamic nature of sites like Twitter can easily make these labels or associated values outdated. The most important advantage of context anaysis, however, is that it is independent from the changing use of terms over time that may occur in sites like Twitter. (Calais Guerra et al., 2011)

2.3.1 Sentiment Knowledge Discovery through User Bias Prediction and Transfer-Learning

Because of the changing content and definitions in text streams like Twitter, Calais Guerra et al. (2011) suggest's a different method in predicting user sentiment in such system: bias prediction and transfer-learning. Calais Guerra et al. (2011) suggests that instead of looking at the actual texts of the data itself, it is better to look at the relationship of these posts to other posts and to other users.

Calais Guerra et al. (2011) works in the assumption that like-users would have like-opinions. Hence, Calais Guerra et al. (2011) made use of the concept of *attractors* who are users that are obviously biased towards one opinion. Using this, a graph can be made to see how other users agree with these different *attractors*. The graph is then used to see how biased each user is. Depending on where the user is biased, his or her sentiment towards a topic can be measured. Terms used by the users are then evaluated with new measures of bias depending on the users who use them.

In Calais Guerra et al. (2011)'s work, their methodology was used on two datasets: Twitter data during the presidential elections in Brazil and Twitter data during a soccer game. The research showed that their methodology can be 80% to 90% accurate in classifying tweets knowing only the bias knowledge of 10% of the users.

2.3.2 User-Level Sentiment Analysis using the Social Network Aspect of Twitter

In a similar work by Tan et al. (2011), user sentiment is measured through the social relationships between users (*follower* and *following* information). Like the previous work, Tan et al. (2011) also works on the assumption that similar users would have similar opinions or sentiments towards topics.

Unlike Calais Guerra et al. (2011)'s work, the work of Tan et al. (2011) focuses on getting the sentiment of the users themselves and not on a per-post basis. Tan et al. (2011)'s research showed an accuracy of 70% on some topics (Obama, Sarah Palin, Fox News) and as high as 90% on one (Lakers).

3 Research Methodology

3.1 Review of Related Literature

Other research works particularly on sentiment analysis and visualization will be further reviewed. The different algorithms presented in these papers may be considered for usage in the research work.

3.2 Data Collection

Data will be mined from Twitter using the Twitter API. To be mined are the text content of the tweets and the contextual data such as the relationship of each post from the other and the relationships of users who post with other users.

Initial data collection will be done in a short span of time in order to test algorithms and how they may work with the data. Data collection will continue as the research is done in order to have a richer dataset.

3.3 Determining Keywords and Topics To Be Used

Using Twitter's data on trending topics, the researcher will choose which topics could be used for Sentiment Analysis.

3.4 Sentiment Analysis on the Topics

The researcher will determine the appropriate sentiment analysis techniques to be used on the chosen topics. The sentiment on each topic they are discussing will be recorded. Historical information such as the change in sentiment over time will be recorded.

3.5 Visualization

Various visualization techniques will be utilized to be able to show the historical sentiment information in a more easily comprehensible graphic.

3.6 Documentation

In this phase of the research work, the researcher will document the different activities done all throughout the research process. Documentation will be done gradually throughout the whole process.

3.7 Calendar of Activities

The research process will be mostly following the waterfall process.

Table 3.1 shows a Gantt chart of the activities. Each bullet represents approximately one week worth of activity.

Table 3.1: Timetable of Activities

Activities (2012)	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Review of Related Literature	●●	●●●●	●●●●	●●			●●●●	●●				
Data Collection			●●●	●●●●	●●●●	●●●●	●●●●					
Determining Keywords and Topics to be used			●●	●				●●	●●			
Sentiment Analysis on Topics			●	●●●					●●	●●●●		
Visualization			●	●●●						●●●	●●	
Documentation		●●●●	●●●●	●●●●	●●●●	●●●●	●●●●	●●●●	●●●●	●●●●	●●●●	●

References

Bifet, A., & Frank, E. (2010). Sentiment knowledge discovery in twitter streaming data. In *Proceedings of the 13th international conference on discovery science* (pp. 1–15). Berlin, Heidelberg: Springer-Verlag. Available from `http://dl.acm.org/citation.cfm?id=1927300.1927301`

Calais Guerra, P. H., Veloso, A., Meira, W., Jr., & Almeida, V. (2011). From bias to opinion: a transfer-learning approach to real-time sentiment analysis. In *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 150–158). New York, NY, USA: ACM. Available from `http://doi.acm.org/10.1145/2020408.2020438`

Michelson, M., & Macskassy, S. A. (2010). Discovering users' topics of interest on twitter: a first look. In *Proceedings of the fourth workshop on analytics for noisy unstructured text data* (pp. 73–80). New York, NY, USA: ACM. Available from `http://doi.acm.org/10.1145/1871840.1871852`

Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2011). Sentiful: A lexicon for sentiment analysis. *IEEE Transactions on Affective Computing*, *2*, 22-36.

Popescu, A.-M., & Pennacchiotti, M. (2010). Detecting controversial events from twitter. In *Proceedings of the 19th ACM international conference on information and knowledge management* (pp. 1873–1876). New York, NY, USA: ACM. Available from `http://doi.acm.org/10.1145/1871437.1871751`

Tan, C., Lee, L., Tang, J., Jiang, L., Zhou, M., & Li, P. (2011). User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1397–1405). New York, NY, USA: ACM. Available from `http://doi.acm.org/10.1145/2020408.2020614`