Sentiment analysis: The need for understanding current trends and its improvisation

SYED MOHAMMAD ARQAM (MCA)[1]*, Dr. Nishat Fatima (Ph.D.)[2], Prabhat Chandra Gupta(Assistant Professor)[1]

[1]College of Computing Science and Information Technology, Teerthankar Mahaveer University, Moradabad, India
[2]Department of Biochemistry, All India Institute of Medical Sciences, Ansari Nagar, New Delhi
*Corresponding author- arqamsyed@gmail.com

Abstract— **Web-blogging is the latest online publishing trend used for feedbacks and both kind of formal and informal knowledge sharing. Internet being a huge landscape has provided a multi-dimensional perspective to monitor sentiments related to various trends. We have used twitter for monitoring the sentiments of people engaged in different perspective throughout the world with respect to various current flashing trends like Politics, Movies, Science, etc. We describe several groups on new data sets based on group postings indicating a strong representation of the viewpoint to the original post. We would be enlightening requirements of more rigorous tools to be employed for improvising the current tool.**

Keywords— **Sentiment, Twitter, Blogging, Computing, Internet, Analysis**

I. INTRODUCTION

Twitter is the most prominent microblogging website where web user express and share their messages("i.e. tweets") on all kind of topics and event every day. Twitter, 320 monthly active users (December 31, 2015)[1] and over 350,000 tweets sent per minute[2], which leads to huge information for an organisation to analyse their customer sentiment on their reputation and brands.

User share their real opinion freely on Twitter which ideally become a great source for extracting user sentiment on various interesting topic such as entertainment gossips, movie review, discussion on the latest product in the market, politics, etc.

Sentiment analysis over Twitter data suffer from several new challenges due to the short length and unstructured data. Two main research guideline can be identified in the literature of sentiment analysis on microblogs. First guideline is to find the new methods to run such analysis, such as performing sentiment label propagation on Twitter follower graphs [1], and employing social relations for user-level sentiment analysis [2,]. The second guideline is focused on identifying new sets of features to add to the trained model for sentiment identification, such as microblogging features including hashtags, emoticons [3], the presence of intensifiers such as all-caps and character repetitions [4] etc., and sentiment-topic features [5].

This paper falls into the second guideline, by investigating a novel set of features derived from the semantic conceptual representation of the entities that appear in tweets. The semantic features consist of the semantic concepts (e.g. "person", "company","city") that represent the entities (e.g. "Steve Jobs", "Vodafone", "London") extracted from tweets. The rationale behind introducing these features is that certain entities and concepts tend to have a more consistent correlation with positive or negative sentiment. Knowing these correlations can help determining the sentiment of semantically relevant or similar entities, and thus increasing accuracy of sentiment analysis. To the best of our knowledge, using these semantic features in the model training for sentiment analysis has not been explored before. We evaluated three popular tools for entity extraction and concept identification; AlchemyAPI,[1] Zemanta,[2] and OpenCalais,[3] and used the one that performed best in terms of quantity and accuracy of the identified concepts.

[1]www.alchemyapi.com
[2] www.zemanta.com
[3] www.opencalais.com

Related Work

Sentiment analysis of tweets data is considered as a much complex than that of conventional text due to the short length of tweets (140), the frequent use of informal and irregular words like "whaaatuu-up", "tmrw",etc and the change of language in Twitter. The large amount of work has been conducted in "Twitter sentiment analysis" following the feature-based approaches.

Go et al. [6] explored expand different n-gram features in the union with POS tags into the training of supervised classifiers including Naive Bayes (NB), Maximum Entropy (MaxEnt) and Support Vector Machines (SVMs). They found that MaxEnt trained from a combination of unigrams and bigrams outperforms other models trained from a combination of POS tags and unigrams by almost 3%. However, a contrary finding was reported in [9] that adding POS tag features into n-grams improves the sentiment classification accuracy on tweets.

Barbosa and Feng [3] confront that the using n-grams on tweets, data may obstruct the classification performance because of the infrequent words in Twitter. Instead, they proposed using microblogging features such as re-tweets, hashtags, replies, punctuations, and emoticons. They found that using these features to train the SVMs enhances the sentiment classification accuracy by 2.2% compared to SVMs trained from unigrams only.

A similar finding was reported by Kouloumpis et al. [6]. They explored the microblogging features including emoticons, abbreviations and the presence of intensifiers such as all-caps and character repetitions for Twitter sentiment classification. Their results show that the best performance comes from using the n-grams together with the microblogging features and the lexicon features where words tagged with their prior polarity. However, including the POS features produced a drop in performance. Agarwal et al. [7] also explored the POS features, the lexicon features and the microblogging features. Apart from simply combining various features, they also designed a tree representation of tweets to combine many categories of features in one succinct representation. A partial tree kernel [8] was used to calculate the similarity between two trees. They found that the most important features are those that combine prior polarity of words with their POS tags. All other features only play a marginal role. Furthermore, they also showed that combining unigrams with the best set of features outperforms the tree kernel-based model and gives about 4% absolute gain over a unigram baseline. Rather than directly incorporating the microblogging features into sentiment classifier training, Speriosu et al. [1] constructed a graph that has some of the microblogging features such as hashtags and emoticons together with users, tweets, word unigrams and bigrams as its nodes which are connected based on the link existence among them (e.g., users are connected to tweets they created; tweets are connected to word unigrams that they contain etc.). They then applied a label propagation method where sentiment labels were propagated from a small set of nodes seeded with some initial label information throughout the graph. They claimed that their label propagation method outperforms MaxEnt trained from noisy labels and obtained an accuracy of 84.7% on the subset of the Twitter sentiment test set from [4]. Existing work mainly concentrates on the use of three types of features; lexicon features, POS features, and microblogging features for sentiment analysis. Mixed findings have been reported. Some [9] argued the importance of POS tags with or without word prior polarity involved, while others emphasised the use of microblogging features [3].

In this paper, we present a method to collect a corpus with positive and negative sentiments, and a corpus of objective texts. Our method allows to collect negative and positive sentiments such that no human effort is needed for classifying the documents. Objective texts are also collected automatically. The size of the collected corpora can be arbitrarily large. We perform statistical linguistic analysis of the collected corpus. We use the collected corpora to build a sentiment classification system for microblogging. We conduct experimental evaluations on a set of real

microblogging posts to prove that our presented technique is efficient and performs better than previously proposed methods.

We collected a corpus of 300000 text posts from Twitter evenly split automatically between three sets of texts:

1. texts containing positive emotions, such as happiness, amusement or joy
2. texts containing negative emotions, such as sadness, anger or disappointment
3. objective texts that only state a fact or do not express any emotions

We perform a linguistic analysis of our corpus and we show how to build a sentiment classifier that uses the collected corpus as training data.

Corpus collection

Using Twitter API we collected a corpus of text posts and formed a dataset of three classes: positive sentiments, negative sentiments, and a set of objective texts (no sentiments). To collect negative and positive sentiments, we followed the same procedure as in (Read, 2005; Go et al., 2009). We queried Twitter for two types of emoticons:

• Happy emoticons: ":-)", ":)", "=)", ":D" etc.
• Sad emoticons: ":-(", ":(", "=(", ";(" etc.

The two types of collected corpora will be used to train a classifier to recognize positive and negative sentiments. In order to collect a corpus of objective posts, we retrieved text messages from Twitter accounts of popular presidential candidate, such as "Donald Trump", "Hillary Rodham Clinton" and latest movie like Superman Vs Batman and latest technology in the market.

Tweets cannot exceed 140 characters by the rules of the microblogging platform, it is usually composed of a single sentence. Therefore, we assume that an emoticon within a message represents an emotion for the whole message and all the words of the message are related to this emotion. In our research, we use English language. However, our method can be adapted easily to other languages since Twitter API allows to specify the language of the retrieved posts.

	tweet	date	lat	lon
21483	Donald Trump at Trump Rally @ Peabody Opera Hous...	2016-03-23 21:14:54	38.62777214	-90.20180254
21484	At the Donald Trump rally for less than five minutes, ...	2016-03-29 20:55:14	42.7145805	-88.9827423
21485	Mexicans burn Donald Trump effigies in Easter ritual ...	2016-03-27 20:31:08	41.88804564	-87.62626724
21486	Donald the Hutt or Jabba the Trump? You decide. #ho...	2016-03-26 21:41:04	34.8009682	-87.6762009
21487	RT @caragsdale: Blistering! A Trump supporter wakes...	2016-03-28 18:57:38	NA	NA
21488	@elizabethforma @realDonaldTrump Did Elizabeth ju...	2016-03-22 20:31:59	36.7842542	-76.2678703
21489	RT @pbump: ATTN TRUMP FANS: Here is your new st...	2016-03-29 22:32:30	NA	NA
21490	Donald Trump says that he is the least racist person ...	2016-03-22 04:12:36	41.2407188	-95.9760462
21491	'That's the argument of a 5-year-old' #trump #gold @...	2016-03-30 03:03:51	45.390907	-122.727061
21492	#ICantBelieveIJustSaw Donald Trump in a porn ...it wa...	2016-03-23 02:28:30	45.8571115	-95.3756431

Showing 21,473 to 21,492 of 21,492 entries

Examples of Twitter posts with expressed users' opinions

WordCloud package

A word cloud is a text mining method that allows us to highlight the most frequently used keywords in a paragraph of texts. It is also referred to as a text cloud or tag cloud. The procedure of creating word cloud is very simple in R software if you know the different steps to execute. A text mining package (tm) and word cloud generator package (wordcloud) are available in R for helping us to analyze texts and to quickly visualize the keywords words as a word cloud.

We collected 21492 recent tweet from Twitter to analyse the most frequently used keywords.

We see from the word cloud that among the most frequent words in the tweets are 'hillary', 'clinton', 'attacked'. This suggests that most tweets were on Trump's recent election campaign.

We also highlight the most frequently used keywords related to Hillary Clinton

We see from the word cloud that among the most frequent words in the tweets are 'donald', 'elect', 'lead', 'poll'. This also suggests that most tweets were on Hillary Clinton's recent election campaign.

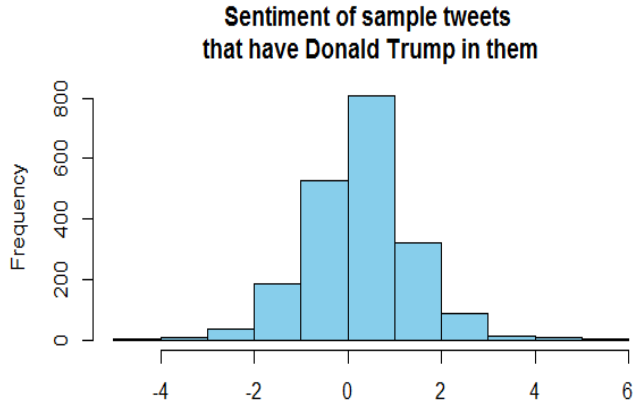
Defining Sentiment

For the purposes of our research, we define sentiment to be "a personal positive or negative feeling." Here are some examples:

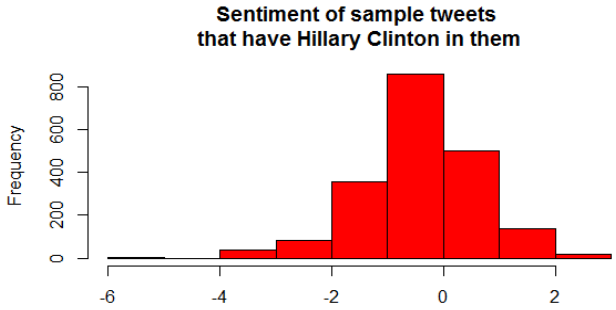
Sentiment	Query	Tweet
Positive	jquery	dcostalis: Jquery is my new best friend.
Neutral	San Francisco	schuyler: just landed at San Francisco
Negative	exam	jvici0us: History exam studying ugh.

We will perform sentiment analysis on latest movie Super Vs Batman, and on politician.

First we will analyse the Sentiment of sample tweets that have Donald Trump in them

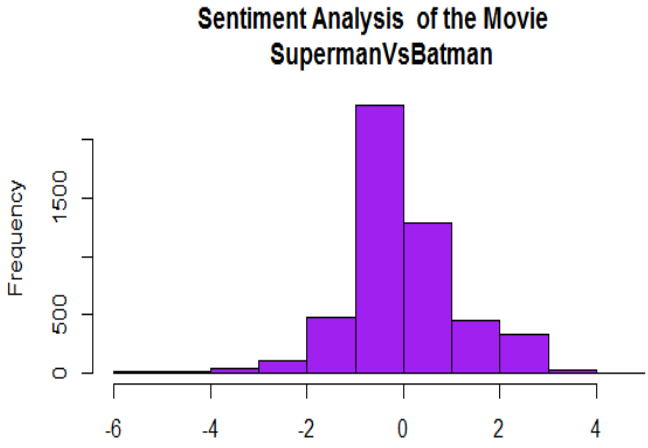


Second we will analyse the Sentiment of sample tweets \n that have Hillary Clinton in them



We see from the histogram that the sentiment is slightly positive.

Analysis on latest movie(Superman Vs Batman)



We see from the histogram that the sentiment is more positive. Hence revenue will more because the sentiment of positive is double of the sentiment of the Negative.

Conclusion

From the above result we can say that Hillary Clinton is a powerful candidate as compared to Donald Trump for upcoming election. This is how we extract the sentiment of web user from their tweets. We perform same sentiment function to analyze the sentiment of user for other purpose like the company want to know reputation in the market.

REFERENCES

[1] *Speriosu, M., Sudan, N., Upadhyay, S., Baldridge, J.: Twitter polarity classification with label propagation over lexical links and the follower graph. Proceedings of the EMNLP First workshop on Unsupervised Learning in NLP pp. 53–63 (2011)

[2] Tan, C., Lee, L., Tang, J., Jiang, L., Zhou, M., Li, P.: User-level sentiment analysis incorporating social networks. In: Proceedings of the 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) (2011)

[3] Barbosa, L., Feng, J.: Robust sentiment detection on twitter from biased and noisy data. In: Proceedings of COLING. pp. 36–44 (2010)

[4] Kouloumpis, E., Wilson, T., Moore, J.: Twitter sentiment analysis: The good the bad and the omg! In: Proceedings of the ICWSM (2011)

[5] Saif, H., He, Y., Alani, H.: Semantic Smoothing for Twitter Sentiment Analysis. In: Proceeding of the 10th International Semantic Web Conference (ISWC) (2011)

[6] Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision.

[7] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R.: Sentiment analysis of twitter data. In: Proc. ACL 2011 Workshop on Languages in Social Media. pp. 30–38 (2011)

[8] Moschitti, A.: Efficient convolution kernels for dependency and constituent syntactic trees. In: Proceedings of the European Conference on Machine Learning. pp. 318–329 (2006)

[9] Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of LREC 2010 (2010)