International Journal of Scientific Engineering and Applied Science (IJSEAS) – Volume-2, Issue-4,April 2016
ISSN: 2395-3470
www.ijseas.com

Sentiment Analysis Of Twitter Generated Data

Ananth Nath[1], Anirudh Sudan[2], Gautam Kumar[3], Saurabh Bhosale[4]
[1] MIT AOE, Pune, India
[2] MIT AOE, Pune, India
[3] MIT AOE, Pune, India
[4] MIT AOE, Pune, India

Abstract

Twitter is a microblogging website which allows users to share their views and opinions. These tweets are displayed on an interface. Users can interact with each other by using the @redirection. The tweets and trending topics can be displayed on the UI. In this day and age nearly a billion people with access to the internet generate huge terabytes of data on a daily basis. This leads to difficulty in Data mining and analyzing data from social networks. Sentiment analysis on a website like twitter can play a vital role in helping advertising and market strategies. Sentiment analysis can be used in prediction of trends in the market when implemented correctly as it paints an accurate picture of the minds of users and potential customers.

Keywords: *Sentiment Analysis, Emoticons, Twitter, hashtag.*

1. Introduction

The advent of social media has seen a drastic change in how content is created and shared on the internet. This has shifted the focus of marketing and advertising agencies from traditional methods to digital marketing. One thing that most of these agencies require is the analysis of the content on such social media websites like twitter. Twitter acts as a platform. for different kinds of users to share their views and sentiment on various kinds of topics in 140 characters or less. We have developed a tool that is able to extract tweets pertaining to a particular topic and analyze them to calculate their polarity i.e. positive, negative or neutral. The advent of social media has seen a drastic change in how content is created and shared on the internet. This has shifted the focus of marketing and advertising agencies from traditional methods to digital marketing. One thing that most of these agencies require is the analysis of the content on such social media websites like twitter. Twitter acts as a platform for different kinds of users to share their views and sentiment on various kinds of topics in 140 characters or less. We have developed a tool that is able to extract tweets pertaining to a particular topic and analyze them to calculate their polarity i.e. positive, negative or neutral.

2. Project Idea

The objective of the project is to successfully implement a sentiment analysis tool that retrieves tweets and is able to infer the sentiment of these tweets.
Sentiment analysis can be classified into 3 types:

1. Sentence Level: Sentence Level sentiment analysis focuses on an entire sentence and calculates the sentiment of the sentence as a whole.

2. Aspect Level: Aspect Level sentiment analysis focuses on certain aspects/characteristics of an entity.

3. Document Level: In a document there may be many opinions expressed throughout the length of the document. These opinions need to be reduced to a single opinion as it is easier to understand the overall opinion or sentiment that is being conveyed through this document using Document level sentiment analysis.

2.1 Project Aim

The aim of the project is to build a sentiment analysis tool that is capable of extracting the polarity of tweets retrieved from the twitter API with as little human interaction required as possible.

3. Literature Survey

The following papers and publications were referred for this project:
1. Correlation Analysis of User Influence and Sentiment on Twitter Data- Uses reciprocity to calculate and do the analysis. Calculates value of influence using Bayes theory, using number of retweets

2. Crime Prediction Using Twitter Sentiment and Weather-We can predict response or display average response to an entity. Paid attention to emoticons.

3. Visual Sentiment Analysis on Twitter Data Streams Sentiment analysis is classified into:
I. Topic-based.
II. Stream analysis.
III. Visual analysis.
Pixel sentiment analysis that comes under visual analysis focuses on geo-maps to show the sentiments.

4. Twitter Sentiment Analysis: Used lexicon based and machine learning approach for sentiment analysis. Output was stored in JSON file and the data was represented using a pie-chart.

5. SASM: A Tool for Sentiment Analysis on Twitter- Analyzes the data from twitter and describes the process to emoticon sentiment analysis resource. Uses lexicon based approach using fuzzy linguistic logic hedges for analysis.

6. Sentiment Analysis on Tweets for Social Events TSAM model had three modules:
I. Sentiment identification.
II. Sentiment aggregation.
III. Scoring module.

7. Twitter for Sentiment Analysis: When Language Resources Are Not Available- Lexicons contain lists of words an-notated with their emotional assessments. Annotated word lists are one of the tools frequently used in sentiment analysis to detect a mood or classify emotions within a text.

4. Proposed Idea

The proposed system consists of three modules:

1. Feature selection module is built for extracting the opinionated words from each sentence.

2. Sentiment identification module that associates expressed opinions with each relevant entity in each sentence level.

3. Sentiment aggregation and scoring module is built for calculating the sentiment scores for each entity. In our proposed system, first the tweets are taken from twitter API and then for each sentence in tweet, POS tagging and stemming is performed. A Part-Of- Speech Tagger (POS Tagger) stemming is used to select one single form of a word instead of different forms. The goal of stemming is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. In this work we use the Stanford software package for both POS tagging and stemming. A typical tweet contains word variations, emoticons, hashtags etc. The objective of the preprocessing step is to normalize the text into an appropriate form to extract the sentiments. Below are the preprocessing steps used:

    a) POS Tagging: POS Tagger gives a part of speech tag associated with the words. POS tagging is done using NLTK.

    b) Stemming: Stemmer gives the stem words; non-stem words are stemmed and replaced with stem words. This would aid the engine to do a word match from

the text to the lexicon. Stemming is done using NLTK.

c) Exaggerated word shortening: Words which have the same letter more than twice and not present in the lexicon are reduced to the word with the repeating letter occurring just once.

d) Emoticon detection: Emoticons have some sentiment associated with them. Twitter NLP is used to extract emoticons along with the sentiments in the Twitter data.

e) Hashtag detection: The hashtag is a topic or a keyword that is marked with a tweet. Hashtag is a phrase starting with no space between them. Hashtags are identified and sentiments are extracted from them.

f) Stop Words: All the stop words (like a, an, the, is etc.) and discourse connectives are discarded.

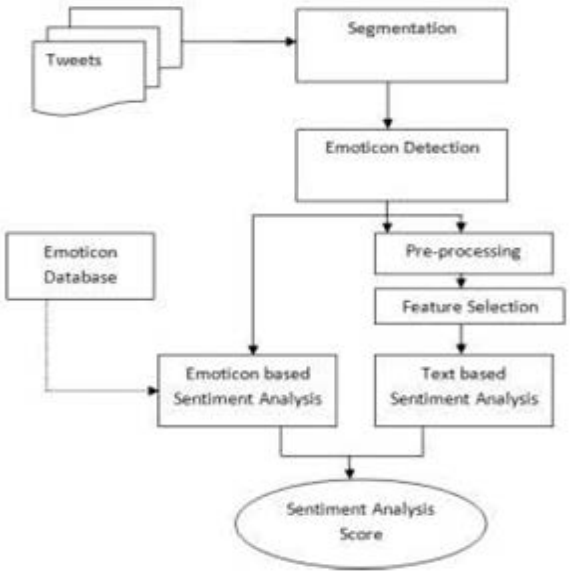

Fig 1: Workflow

5. Twitter API

Twitter provides access to these tweets through a developer account which provides us with client id to access tweets.
Use of Twitter REST API allows developers to access core Twitter data. This includes the possibility to update timelines, status data, and users use OAuth authentication. Most parts of the API use a REST model, Data is returned in both XML and JSON.

6. Accuracy

As we have to prove that the accuracy of emoticon based analysis is more than text based analysis, we first made a text based classification and then an emoticon based analysis. The accuracy of data can be calculated using the formula given below:

$$\text{Accuracy} = \frac{TP+TN+TNL}{TP+TN+TNL+FP+FN+FNL}$$

Where,

- TN / True Negative: case was negative and predicted negative.
- TP / True Positive: case was positive and predicted positive.
- FN / False Negative: case was positive but predicted negative.
- FP / False Positive: case was negative but predicted positive.
- TNL / True Neutral: case was negative and predicted neutral.
- FNL / False Neutral: case was false and predicted neutral.

The graph shown below shows the sentiments of tweets without considering polarities of emoticons:

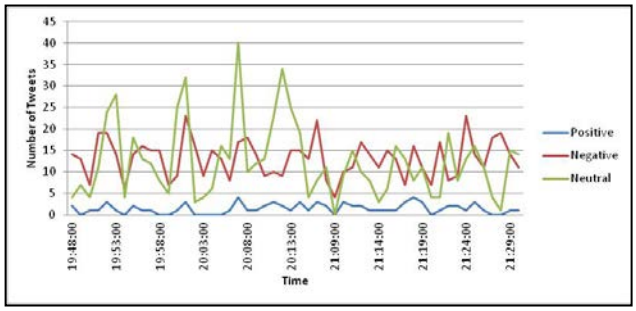

Fig: Sentiments Analysis without Emoticons

The graph shown below shows the sentiment of tweets with emoticons taken into consideration.

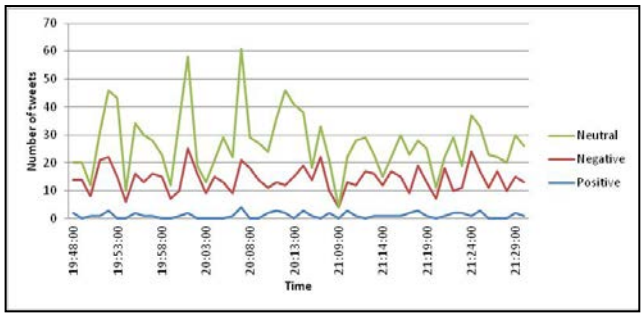

Fig: Sentiment Analysis with Emoticons

When these two graphs are compared, we can find that there's a difference in opinion.

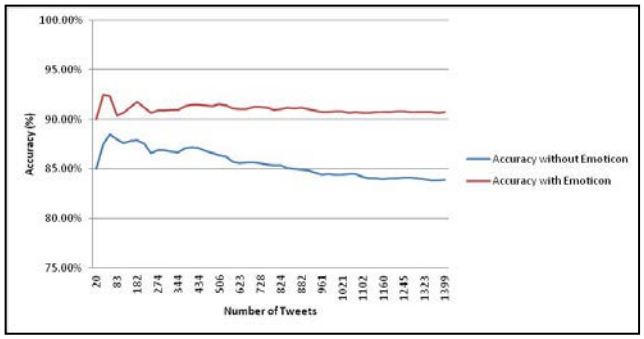

Fig: Accuracy

7. Scope for improvement

Sentiment analysis tools are still in their infancy. This means that most tools that are being developed have a scope for improvement. There may be issues pertaining to the language of the tool, or the accuracy. This is why we have identified the following as the areas where our tool can be improved:

1. Accuracy-Though the tool developed by us will be able to calculate the average sentiment, the system will not be 100% accurate in deciding the exact opinion of the user as simple words may not represent the emotion of the user every time

2. Language Barrier-The machine developed by us uses preexisting databases which understand only English, for analyzing the tokenized tweets and the words present in them. As twitter spans over a large demographic of users, many of them may use languages other than English. Hence tools using databases for every known language or translators can be used in future.

3. Use of abbreviation or Slang-As explained earlier, the tool uses databases for comparing words to those present in the tweets. These databases use the words defined in the dictionary and therefore would not understand any abbreviations or slang words present in them. The dictionary can be expanded and slang words or abbreviations in texts used in daily language can be incorporated in the dictionary.

8. Conclusion

In conclusion the tool developed by us is a simple showcase of a system which will have a number of applications in the near future. With the shift of advertising and marketing from print to digital and social media, sentiment analysis will have a huge role in deciding how to push products to the consumers and how to interact with them and twitter will be one of the main platforms for users to exploit this untapped market.

Acknowledgments

We would like to thank Mrs. Mayura Kulkarni for her efforts and her guidance. We thank her for allowing us access to IEEE publications and for her valuable inputs.

References

[1] Fadhli Mubarak bin Naina Hanif, G. A. Putri Saptawati, ”Correlation analysis of user influence and sentiment on twitter data”.
[2] Xinyu Chen, Youngwoon Cho, and Suk young Jang,” Crime Prediction Using Twitter Sentiment and Weather”ification. J. Mach. Learn. Res. 3 (Mar. 2003), 1289-1305.
[3] Ming Hao, Christian Rohrdantz, Halldr Janetzko,
Umeshwar Dayal Daniel A. Keim,Lars-Erik Haug, MeiChun Hsu, ”Visual Sentiment Analysis on Twitter Data Streams”.
[4] Onifade O.F.W, Malik M.A.,” SASM: A Tool for Sentiment Analysis on Twitter”.
[5] Aliza Sarlan, Chayanit Nadam, Shuib Basri,“Twitter Sentiment Analysis”.
[6] Seyed-Ali Bahrainian, Andreas Dengel,” Sentiment Analysis and Summarization of Twitter Data”.
[7] Meral, Banu Diri,” Sentiment Analysis on Twitter”.
[8] Alexander Pak, Patrick Paroubek,” Twitter for Sentiment Analysis: When Language Resources Are Not Available”.
[9] Xujuan Zhou,Xiaohui Tao, Jianming Yong, ”Sentiment Analysis on Tweets for Social Events”.

Ananth Nath Final Year Computer Engineering student at MITAOE, Pune.

Anirudh Sudan Final Year Computer Engineering student at MITAOE, Pune.

Gautam Kumar Final Year Computer Engineering student at MITAOE, Pune.

Saurabh Bhosale Final Year Computer Engineering student at MITAOE, Pune.