Sentimental Analysis of Twitter Data

Usha,

Master's in Information Technology,

University of New Hampshire, Manchester.

Abstract:

Sentimental analysis is a process of determining the opinion or feeling regarding a piece of text. Now-a-days there are numerous social media sites emerging and millions of people use these social media/microblogging sites on day to day basis to express their emotions. This research is about analyzing the sentiments from the text from famous microblogging site named Twitter. Analyzing the emotions, sentiments or contexts behind the 140 character of a single tweet is the main task. This research is about achieving it and providing the most effective and accurate sentimental analyzer. The analyzer aims to categorize the tweets obtained into positive tweets, negative tweets or neutral tweets. The social media is expanding rapidly with over 200 million registered users out of which 100 million are active users and half of them log on to the twitter on very day basis -generating almost 250 million tweets per day. I'm making a research on this data to see how much of effect does it have on the rest of the world, how it is done and what are the methods they use for it to succeed in such a huge scale. Most beneficial sector from this research would be the companies who wants to know the reviews and feed backs from its customers and wants to monitor that. The aim of this research is to identify which among all the previous and present sentimental analyzer is best for the companies, and which of those individually or combinations of such analyzer can come up with the most accurate and effective analyzer.

Introduction:

These social medias provide a platform for millions to express their emotions in real time, which in turns leads to the emergence of abundant data. Among all the social medias this research is centered to Twitter. Twitter is a treasure of all sentiments, people all around the world output thousands of reactions and opinions on each topic every second. So, one could say that twitter is a big phycological database, which is constantly being updated. Twitter audiences vary from celebrities to common people, so the information obtained could be influential, along with an enormous collection of mixed emotions and opinions. Thus, with such a large amount of data freely available, the data scientist could easily an analyze the sentiments of a tweets or series of tweets. The core idea is to analyses millions of texts in seconds with the help of machine learning.

Sentimental analysis has become a major area of analysis, it can analyses and help companies to receive feedback and reviews from its customers. Once one understands the customer's opinion, by analyzing the reviews and feedback the company can identify which features their customers like and dislike thus helping a recommendation system for the

company. Companies across the world are implementing this system by using the machine learning techniques. Currently, there are already hundreds of sentimental analyzers in use. This research will help narrow the number of such analyses and boost the effectiveness and accuracy of those analyzing systems by combining individual sentimental analyzers, in order to achieve the best recommendation system possible.

As we are the audience of the blogging or microblogging platforms and services here, it keeps growing rapidly and the data from these sources gives us the big platform for us to experiment with how the data are effective, how they are affecting the regular day-to-day activities, most of all of them how can we get the machines understand these data and sentiments how it will learn. This would be very important for the big companies when it comes about understanding about their customers reviews for example:

- How to people like the new product
- What are the changes customers need for the recently released product?
- How positive (or negative) are the customers reacting to the releasing of the product

Political people might want to know about what people think about their government, what are the changes they want, what are the problems people are facing and how they expect the government to handle it. Social organization may want to ask for the people opinion, since internet gives the wide range of opinions all over the world, it is one of the best places for asking for ideas, opinions, support and marketing. The wide range of this information can be obtained by the microblogging services. We are using such services for the sentimental analysis and perform data mining, opinion mining and text classification on all of these data.

One of the main questions is why twitter:

- One of the major social media where there are 200 million users registered in it and there is a wide range of data to be experimenting with, also we have the data stating that around 250 million tweets are being tweeted on everyday basis which is enough amount of data for one to perform the research on sentimental analysis. This is the valuable platform for obtaining wide range of opinions.
- Twitter contains a huge amount of the data posts and it grows rapidly every day. There will be different opinions of people all over the globe, different culture different opinions, ideas, solutions basically sentiments.
- One of the major problems also raisers as well, since there are wide range of the people, there will be different languages as well, analysis these sentiments would be one of the challenges as well when we must analyses the sentiments in multiple languages.

Problem statement:

- Tweets will be having sentiments hidden in them. Deciphering should be always right, otherwise the sentiments might be affected.
- The given tweet would be in different languages, trying to understand all of them.
- Given a tweet, the analyzer should be able to decide whether the message is positive or negative. If the message is conveying both, then the analyzer should choose a single polarity, which it inclines towards the most.

Methods:

Given a tweet, a sentimental analyzer should be able to decide whether the message is a positive tweet, a negative tweet or a neutral tweet. The case where the message is conveying both, then the analyzer should notify the maximum polarity which the tweet inclines to. There are three machine learning algorithms used in the data mining, opinion mining and text categorization, **Naive Bayes, Maximum Entropy, Support Vector Machine**.

Naive Bayes is used for the categorization of the text. Naive Bayes classifiers assigns values to each word in the tweets, which later is helps in scoring of the tweets.

Maximum Entropy, this helps in the overlapping problems of the tweets and helps to analyze the tweets more efficiently. Finally, Support Vector Machine is used to support vector machines, these are used for pattern reorganizations, which plays a prominent role in sentimental analysis. Addition to these, one could use the annotated predefined dictionary and acronym dictionary for comparison of emoticons and acronyms used in the tweets. To improvise one can also use the POS (Parts of speech tags) tags here, which helps in recognizing the sentiments in the tweets, concentrating solely on the parts of speech used in the tweets.

Furthermore, predefined data sets are also included for the comparison of the data available from tweets. Some of the data sets will compare the characters, while others will compare the phrases, emotions and hashtags.

Here the analyses help in recognizing the public sentiments for the trending topics all over the internet, the main idea behind this is based on Data Extractions, pre-processing of extracted data and categorizing them in different emotions.

Irrespective of individual methods, almost 90% of the sentimental analyzers use the following process in their algorithms to process and analyze the sentiments from tweets.

1. Retrieval of tweets.
2. Pre-processing of extracted data:
    - Filtering- cleaning up the basic tweets and categorizing them into positive negative and neutral emotions
    - Tokenization- the whole sentence is broken into small tokens, which will help analyze the whole sentence based on the punctuations and other special characters.
    - Removal of extra words – Unwanted extra words are chopped off in this step, and only the meaningful tokens are taken in considerations.

3. Parallel programming.
4. Sentimental analysis (polarity identification).
5. Output statement.

Applications:

One can get the idea on how well the public is taking the opinion of someone like the President or some other Public figure. Product managers can analyze how their product is viewed by the public by reading the reviews and analyzing the reviews. Some of the main uses are social media consultations, audience responses, social media alerts, trend tracking, crisis management etc.

Microblogging is one of the major types of the communications in recent days, it's said to be the recent brand-mouthing online. The large amount of data which we obtain from the Twitter data can be used for all the sentimental analysis of the data. In majority of the research said abut automatically obtaining the data in a huge collection which can be used for the training of the sentimental classifier. Here are some of the major applications of the Sentimental analysis of data:

- Voice of the customers and Customer experience management
- People analytics and voice of employee
- Social media monitoring
- Regulatory Compliance
- Robotic Process Automation
- Enhancing the customers experience through sentimental analysis in business
- Business for brand marketing
- Building up the intelligence buildup for the business
- Reputation management of the company
- Competitor monitoring
- Information systems
- Machine Learning
- Predictive modeling

Conclusions:

Analyzing millions of texts was successfully executed. Addition of the POS tags resulted in effective evaluation of sentimental analysis of twitter data. Combination of analyzers resulted in significant higher performance than the individual analyzers. The overall performance of the system was improvised. Challenges like highly unstructured, non-grammatical, out of vocabulary words, use of acronyms were also minimized and analyzed, which resulted in an effective and accurate sentimental analyzer.

References:

[1] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau, "Sentiment Analysis of Twitter Data" Department of Computer Science, Columbia University, New York, 2009.

[2] Akshi Kumar and Teeja Mary Sebastian, "Sentiment Analysis on Twitter" department of Computer Engineering, Delhi Technological University, Delhi, India, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3, July 2012

[3] G. Vinodhini, R. M. Chandrasekaran "Sentiment Analysis and Opinion Mining: A Survey" Assistant Professor, Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar-608002, Volume 2, Issue 6, June 2012, IEEE paper

[4] Varsha Sahayak Vijaya Shete Apashabi Pathan BE (IT) BE (IT) ME (Computer) Department of Information Technology International Journal of Innovative Research in Advanced Engineering (IJIRAE), Savitribai Phule Pune University, Pune, India, Issue 1, Volume 2 (January 2015).

[5] Bo Pang, L.L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval January Volume 2 Issue 1-2, 1–94 (2008), http://www.cs.

[6] Sentimentor: Sentiment Classification of Twitter Data.Cornell.edu/home/llee/omsa/omsa. (peer reviwed).

[7] Chuan-Ju Wangz, Ming-Feng Tsaiy, Tse Liuy, Chin-Ting Changzy, "*Financial Sentiment Analysis for Risk*

Prediction" Department of Computer Science & Program in Digital Content and Technology National Chengchi

University Taipei 116, 2013.

[8] Changbo Wang, Zhao Xiao, Yuhua Liu, Yanru Xu, Aoying Zhou, and Kang Zhang, *"SentiView: Sentiment Analysis*

and Visualization for Internet Popular Topics" IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS, VOL.43, NO. 6, NOVEMBER 2013.