A Survey: Sentiment Analysis on Social Media Data (Twitter)

Udit Hazarika[1], Rashmi Singha[1], Prof. Ratan Kumar Saha[2]

1. Student of Integrated MSc. (Information Technology), Semester -X
School of Computing Sciences, The Assam Kaziranga University, Jorhat, Assam, India

2. Professor and Associate Dean (PG)
School of Computing Sciences, The Assam Kaziranga University, Jorhat, Assam, India

Abstract

Social media have received more attention nowadays. Public and private opinion about a wide variety of subjects are expressed and spread continually via numerous social media. Twitter is one of the social media that is gaining popularity. Twitter offers a platform and capability to business people, wholesalers, dealers and traders to sell products and put their advertisements directly for the consumers, take feedback and opinion in a fast and effective way as per the consumer's perspectives toward the critical success in the market place. Developing a program for sentiment analysis is an approach to be used to computationally measure customers' perceptions. Sentiment analysis of Twitter data became a research tread in this decade. Among popular social network portals, Twitter data is the point of attraction to researchers to predict the democratic events, customer brands, popular movie box-office, stock market, reputation of personalities etc. The term sentiment refers to the feelings or opinion of person towards some particular domain. This paper on the design of a sentiment analysis, extracting a vast number of tweets. Prototyping is used in this development. Results classify customers' perspective via tweets into positive, negative and neutral, which is represented in a pie chart.

Keywords: Sentiment analysis, Opinion mining, Twitter, Social Media, Stop Words

1. Introduction

Nowadays, the age of internet has changed the way people express their views, opinions. It is now mainly done through blog posts, online forums, product review websites, social media, etc. Nowadays, millions of people are using social network sites like Facebook, Twitter, Google Plus, etc. to express their emotions, opinion and share views about their daily lives. Through the online communities, we get an interactive media where consumers inform and influence others through forums. Social media is generating a large volume of sentiment rich data in the form of tweets, status updates, blog posts, comments, reviews, etc.

Moreover, social media provides an opportunity for the business communities by giving a platform to connect with their customers using advertisements of their product and service offerings. Mostly individuals taking their decision to buy goods and services depend upon the opinion of masses, tweets and media generated content over online to a great extent. For e.g. if someone wants to buy a product or wants to use any service, then they firstly look up its reviews online, discuss about it on social media before taking a decision. The amount of online content generated by tweets and social media are too vast for a normal person to analyse. So, there is a need to automate process to find out various opinion by doing widely used sentiment analysis techniques.

Sentiment analysis (SA) tells user whether the information about the product is satisfactory or not before they buy the product and service. Consumers, producers and traders of the product and services will use this analysis data to understand about their products or services in such a way that it can be offered as textual information retrieval techniques mainly focus on processing, searching or analysing the factual data present. Facts have an objective component but, there are some other textual contents which express subjective characteristics. These contents are mainly opinions, feedback, sentiments, appraisals, attitudes and emotions, which form the core of Sentiment Analysis (SA). It offers many challenging opportunities to develop new applications, mainly due to the huge growth of available information on online source like blogs and social network. For example, Recommendations of items proposed by a recommendation system can be predicted by taking into account considerations such as positive or negative and neutral opinions about those items by making use of Sentiment Analysis.

In this paper, we will discuss social network analysis and the importance of it, then we discuss Twitter as a rich resource for sentimental analysis. In the following sections, we show the high-level abstract of our implementation. We will show some queries on different topics and show the polarity of tweet.

2. Literature Review

2.1Literature Review#1

Title: Social Media Data Mining For Sentiment Analysis.

Author:

K.C.Khatib C.S.E,A/P-Herle, Tal-Hatkangle, Dist-Kolhapur

Tanuja D. Kamble C.S.E,A/P-Kavathe Mahankal Tal-K.M, Dist-Sangli

Bhagyashri R.Chendake C.S.E,S/P-Rui,Tal-Hatkangle,Dist-Kohalpur

Gitam N.Sonavane C.S.E,A/P-Mhaswad, Tal-Man, Dist-Satara

Prof. P.S. Kulkarni, Department of Computer Science Engineering

Sou.Sushila Danchand Ghodawat Trust's Group of Institution (Atigre), Maharastra, Kolhapur.

Publication: International Research Journal of Engineering and Technology (IRJET)

Volume:03 Issue:04 April-2016

Objective

Recently, social media is playing a vital role in social networking and sharing of data. Social Media is favoured by many users as it is available to millions of people without any limitations to share their opinions, educational learning experience and concerns via their status. Students post on social network gives us a better concern to take decision about the particular education system learning process of the system. Evaluating such data in social network is quite a challenging process. In the processed system, there will be a workflow to mine the data which integrates both qualitative analysis and large scale data mining technique.

Conclusion

Overall, we conclude that social network based behavioural analysis parameters can increase the prediction accuracy. It is beneficial for the researchers in learning analytics, educational data mining, and learning technologies. It provides a workflow for analysing social media data for educational purposes that overcomes the major limitations of both manual qualitative analysis and large scale computational analysis of user-generated textual content. Our study can inform educational administrators, practitioners and other relevant decision makers to gain further understanding of student's college experiences. However, presence of all the entities in unbiased and equal manner is necessary to provide accurate results. To understand the influential parameters that effects the result, semantic features are also very useful from point view of the entity itself.

2.2 Literature Review#2

Title: : Social Media Analysis for Product Safety using Text Mining and Sentiment Analysis

Author:

Haruna Isah, Paul Trundle, Daniel Neagu, Artificial Intelligence Research (AIRe) Group, School of Electrical Engineering and Computer Science, University of Bradford Bradford, UK

H.Isah@student.bradford.ac.uk, .R.Trundle@bradford.ac.uk, D.Neagu@bradford.ac.uk

Objective

The growing incidents of counterfeiting and associated economic and health consequences necessitate the development of active surveillance systems capable of producing timely and reliable information for all stake holders in the anti-counterfeiting fight. User generated content from social media platforms can provide early clues about product allergies, adverse events and product counterfeiting. This paper reports a work in progress with contributions including: the development of a framework for gathering and analyzing the views and experiences of users of drug and cosmetic products using machine learning, text mining and sentiment analysis; the application of the proposed framework on Facebook comments and data from Twitter for brand analysis, and the description of how to develop a product safety lexicon and training data for modelling a machine learning classifier for drug and cosmetic product sentiment prediction. The initial brand and product comparison results signify the usefulness of text mining and sentiment analysis on social media data while the use of machine learning classifier for predicting the sentiment orientation provides a useful tool for users, product manufacturers, regulatory and enforcement agencies to monitor brand or product sentiment trends in order to act in the event of sudden or significant rise in negative sentiment.

Conclusion

We have demonstrated how machine learning techniques can be used to infer sentiments over social media data suggesting the views and experiences of drug and cosmetic product users. First a framework is developed for harnessing and tracking these views and experiences using text mining and sentiment analysis, we then conducted two case studies using the framework for comparison of sentiment analysis over three cosmetic brands coded as: Brand X, Brand Y and Brand Z and also over three products: Soap, Cream and Deodorant. A Naïve Bayes classifier was used to obtain a baseline result for assessing other classifiers. This paper reports a work in progress and the initial brand and product comparison results signify the usefulness of text mining and sentiment analysis on social media data while the use of machine learning classifiers for predicting the sentiment orientation provide a useful tool for users, product manufacturers, regulatory and enforcement agencies to monitor brand or product sentiment trends in order to act in the event of sudden or significant rise in negative sentiments.

3. Social Network Analysis

Social Network Analysis is the study of data analysis of people’s interactions using internet-based social media sites as a platform on different topics of their purpose amongst their parents, friends, classmates, customers or consumers. Now-a-days it has received more attention for social and business needs through the social media sites such as Facebook, Twitter, LinkedIn, and Instagram, among others. Millions of people give their opinion of different topics and purpose on a daily basis on social medias. It has many applications in different areas of research from social science to business.

Twitter now-a-days is one of the popular social media which according to the statistician currently has over 300 million Twitter accounts. Twitter data is the rich source to learn about people’s opinion and sentimental analysis. As we know, it is important to determine the sentiment of each tweet, whether is it positive, negative, or neutral.

Another challenge with twitter data is only 140 characters is the limitation of each tweet which cause people to use phrases and words, which are not in language processing. Recently twitter has extended the text limitations to 200 characters per each tweet.

4. Twitter Sentiment Analysis

As we know **“Sentiment analysis**: the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer’s attitude towards a particular topic, product, etc. is positive, negative, or neutral.”— Oxford English Dictionary

To analyse sentiments with Twitter text datasets “tweets” is known as Twitters sentiment analysis to help decision making process of either an organization or an individual. The sentiment-aware systems these days have many applications from business to social sciences.

Since social networks, especially Twitter as platform, contains small texts and people may use different words and abbreviations which are difficult to extract their sentiment by current Natural Language processing systems easily, therefore some researchers have used deep learning and machine learning techniques to extract and mine the polarity of the text. Some of the top abbreviations are FB for Facebook, B4 for before, OMG for oh my god and so on. Therefore sentimental analysis for short texts like Twitter’s posts is challenging.

5. Methods/Approaches

Sentiment Analysis (SA) is the process of computationally identifying and categorizing opinions expressed in a piece of text. In our work, sentiment analysis is defined as to identify the polarity of customer behaviour and the emotions of particular consumer by using the Lexicon based sentiment classification method. This holistic lexicon method will classify polarity as positive, negative and neutral from the text with dictionary of words annotated with their semantic orientations. So, in this method we are going to adopt the methodology as follows:

The architecture of the proposed methodology for Sentiment Analysis process flow

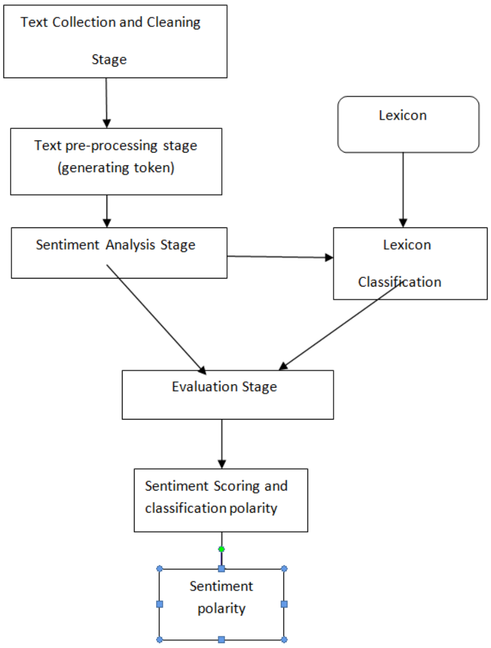

Text Collection and Cleaning Stage
Text pre-processing stage (generating token)
Lexicon
Sentiment Analysis Stage
Lexicon Classification
Evaluation Stage
Sentiment Scoring and classification polarity
Sentiment polarity

Fig.5.1. Architecture of the proposed process flow

The architecture comprises four stages:

- Text collection and cleaning
- Pre-processing
- Sentiment analysis
- Evaluation stage.

5.1 Text Collection and Cleaning Stage

As we know a tweet contains lot of opinions about the data which are expressed in various ways by number of people. Twitter data is used in this work to be classify as positive, negative and neutral polarity to do the sentiment analysis.

In our work, we will be doing sentiment analysis on twitter data by using twitter Application Programming Interfaces (APIs) to share the data. At the text collection and cleaning stage, an API call provided by twitter for data extraction is invoked on Twitter APIs. The Twitter API consists of the Representational State Transfer (REST) and Streaming APIs as follows:

The REST API provides methods for authenticating applications, processing requests, handling imposed limits, etc.

The Streaming API provides client applications with Twitter's global stream (public, user and site) of data.

We will use both the REST and Streaming APIs for searching and fetching tweets.

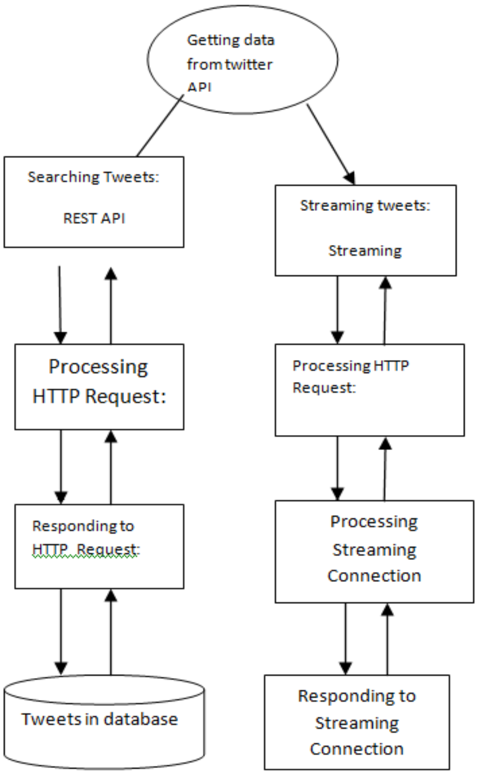

Getting data from twitter API
Searching Tweets: REST API
Streaming tweets: Streaming
Processing HTTP Request:
Processing HTTP Request:
Responding to HTTP Request:
Processing Streaming Connection
Tweets in database
Responding to Streaming Connection

Fig. 5.2. Data Collection Workflow

5.2 Pre-processing Stage:

The raw data of tweet having hidden or direct polarity is highly prone to inconsistency and redundancy. The pre-processing of tweet data include the following steps to transform tweet data into tokenized classification from the stream of text to words:

Step1: Removing delimiters (“,” “.” “;” “{ }”).

Step2: Removing numbers and stop words (and, the, a, it, you, may, that, I, an, of).

Step3: Converting all words to lower or upper case.

Step4: Removing all URLs (e.g. www.xyz.com), hash tags (e.g. #topic), targets (@username).

Step5: Correct the spellings; sequence of repeated characters is to be handled.

Step6: Replace all emoticons with their sentiments, remove all punctuations, symbols and numbers.

Step7: Expand Acronyms (abbreviated formed from the initial letters of other words and pronounced as a word e.g. ASCII, KU)

5.3 Sentiment Analysis Stage

This stage handles the polarity measurement, sentiment classification and clusterization with respect to entire dictionary and some of the targeted entities. We approach these tasks by lexicon analysis. In this process it determines whether a piece of writing (product/movie review, tweet, etc.) is positive, negative or neutral.

Lexical analysis is the first phase of a compiler. It takes the modified source code from language pre-processors that are written in the form of sentences. The lexical analyzer breaks these syntaxes into a series of tokens, by removing any whitespace or comments in the source code.

If the lexical analyzer finds a token invalid, it generates an error. The lexical analyzer works closely with the syntax analyzer. It reads character streams from the source code, checks for legal tokens, and passes the data to the syntax analyzer when it demands.

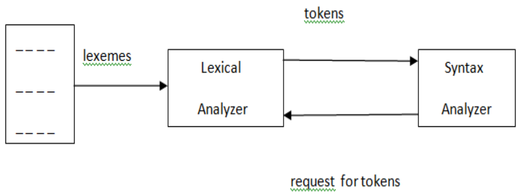


Fig.5.3.lexical analysis

Lexemes are said to be a sequence of characters (alphanumeric) in a token. There are some predefined rules for every lexeme to be identified as a valid token. These rules are defined by grammar rules, by means of a pattern. A pattern explains what can be a token, and these patterns are defined by means of regular expressions.

5.4 Evaluation stage

In this process it evaluates the performance of the sentiment analysis model. It helps in classifying the users comments to either positive or negative or neutral sentiment (Sentiment Polarity).

6. Design and Implementation

Over view

This is a web application which is used to analyse the tweets. We will be performing sentiment analysis in tweets and determine where it is positive, negative and neutral. This web application can be used by any organization office to review their works or by political leaders or by any other company to review about their products and brands.

The main feature of our project is that it helps to determine the opinion about the peoples on products, government work, politics or any other by analysing the tweets. Our system is capable of training the new tweets taking reference to previously trained tweets.

The computer or analysed data will be represented in diagram such as Pie-chart.

This technical paper reports the implementation of the Twitter sentiment analysis, by utilizing the APIs provided by Twitter itself.

6.1Design

The approach to extract sentiment from tweets is as follows:

Step1:Start with downloading and caching the sentiment dictionary

Step2: Download twitter testing data sets, input it in to the program.

Step3: Clean the tweets by removing the stop words.

Step4: Tokenize each word from the tweets dataset and feed in to the program.

Step5: For each word, compare it with positive sentiments, negative sentiments and neutral word in the dictionary. Then increment positive count or negative count, and neutral count.

Step6: Finally, based on the positive count, negative count, and neutral count. we can get result percentage about sentiment to decide the polarity.

we have different procedures to connect the twitter API, fetch the tweets, tweet cleaning or remove stop words, classify tweets which means get the polarity of the tweet, and finally return the results.

6.2 Implementation

a) Installation of "tweepy" and "textblob"

In this paper, we used python to implement sentimental analysis, we have utilized packages ***"tweepy"*** and ***textblob***. We can install the required libraries by following commands:

- pip install **tweepy**
- pip install **textblob**

b) Downloading the dictionary

In this step we are downloading the dictionary by running the following command:

python -m textblob.download_corpora.

The text blob is a python library for text processing and it uses NLTK for natural language processing [6].

Corpora is a large and structured set of texts which we need for analysing tweets.

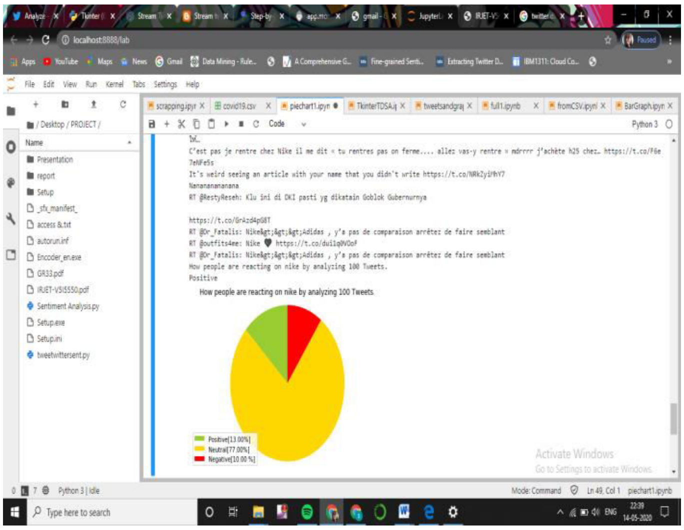
c) *Connect to Twitter using APIs*

To connect to Twitter and query latest tweets, we need to create an account on twitter and define an application. Users need to go to the apps.twitter.com/app/new and generate the API keys.

d) Sample Results

Following shows the sample output of the program for the 'fake news' as a query based on the last 300 tweets from Twitter.

- Positive tweets percentage: 13.00%
- Negative tweets percentage:10.00%
- Neutral tweets percentage: 77.00%



```
1W…
C'est pas je rentre chez Nike il me dit « tu rentres pas on ferme.... allez vas-y rentre » mdrrrr j'achète h25 chez… https://t.co/F6e
7eNFe5s
It's weird seeing an article with your name that you didn't write https://t.co/NRkZyiMhY7
Nananananana
RT @RestyReseh: Klu ini di DKI pasti yg dikatain Goblok Gubernurnya

https://t.co/GrAzd4pG8T
RT @Dr_Fatalis: Nike&gt;&gt;&gt;Adidas , y'a pas de comparaison arrêtez de faire semblant
RT @outfits4me: Nike 🖤 https://t.co/dui1q0VOof
RT @Dr_Fatalis: Nike&gt;&gt;&gt;Adidas , y'a pas de comparaison arrêtez de faire semblant
How people are reacting on nike by analyzing 100 Tweets.
Positive
```

Fig6.1.sample figure of pie chart

7. Future Work

Our future work includes extensive comparison of text mining of different text and sentiment analysis approach on different data sets acquired from multiple resources and multiple languages. We would also work towards finding most computationally inexpensive algorithms for various tasks and sub tasks. We will try to develop sentiment analysis based on opinion summarization technique (survey) using iterative learning tree algorithm.

It uses an iterative practice to categorize the given inputs for the assessment of sentiment by online survey method. This survey method will make decision tree to start comparison from the root node and then compares it with every instance of data. Labels are assigned to the leaf nodes. Every node in the tree represents sentimental knowledge with combine word-similarity knowledge and word-sentiment knowledge with the help of contextual knowledge to define the polarity.

8. Conclusion

In this project we started with very basis of twitter data analysis. We explained for twitter app authentication using OAuth and **Tweepy**.

Then we explained steps to collect historical data as well as streaming data.

We then pre-processed the data using tokenizers. In the final step, we tried to execute a number of the use cases to analyse the stored data.

We represented results of analysing most used terms for a data set, most used **hasgtags**, most used mentions of user account on twitters.

This project is introductory in nature and hence deals with the basic of twitter data analysis using python. In further work, we will try to represent more advanced data analysis patterns decision making more accurate results.

Acknowledgement

We take this golden opportunity too we our deep sense of gratitude to our project guide Prof. Ratan Kumar Saha for his instinct help and valuable guidance with a lot of encouragement throughout this project work, right from selection of topic work up to its completion. Our sincere thanks to Project Co-ordinator Dr. Purnendu Bikash Acharjee who continuously motivated and guided us for completion of this project. We are greatly indebted to our Dean, Dr. Sajal Saha for his valuable guidance and encouragement which have been really helpful for our work. We also especially thanks to those who helped us directly-indirectly in completion of this work successfully.

References

[1] Boiy, Erik, Hens, Pieter, Deschacht, Koen, and Moens, Marie Francine.(2007). —"Automatic Sentiment Analysis in On-line Text." In Proceedings of Conference on Electronic Publishing, pp. 349-360, Vienna, Austria, June 2007.

[2] Cogburn, Derrick, L., Hanson, Mary E., and Wozniak, Amy.(2012). —"Accelerating Social Sciences for the New Age. Moving from Traditional Methods for Analyzing Large Scale Textual Data to Socially High Performance Computational Methods." Paper presented at the CSCW'J2, February 11-15. Seattle. Washington. USA.

[3] Creswell, John. (2007). Qualitative Inquiry and Research Design. Choosing Among Five Approaches. 2nd ed. Sage Publications Inc: California.

[4] Kaplan, Andreas, M., and Haenlein, Michael. (2010). —"Users of the World, Unite! The Challenges and Opportunities of Social Media." Business Horizons, 53 (1): 59-68.

[5] Kim, Soo-Min, and Hovy. Eduard (2006). —"Identifying and Analyzing Judgment Opinions." In Proceedings of the Human Language Technology Conference - North American chapter of the Association for Computational Linguistics annual meeting, New York City, NY.

[6] Kumar, Akshi, and Sebastian, Teja, Mary. (2012). —"Sentiment analysis.A perspective on its past present and future."International Journal of Intelligent Systems and Applications, 4 (10): 1-14.

[7] Kushal, Dave, Lawrence, Steve, and Pennock, David, M. (2003): —"Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews." In Proceedings of the Twelfth International World Wide Web Conference, pp. 519 – 528.