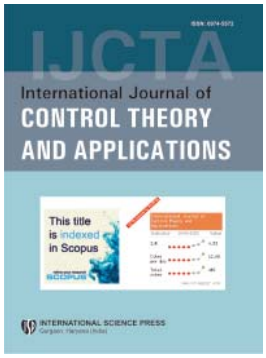

International Journal of Control Theory and Applications

ISSN : 0974–5572

Volume 9 • Number 40 • 2016

Opinion Mining of Twitter Data for Recommending Airlines Services

Pranika Jindal[a] Varun Jaiswal[a] and M. Uma[c]

[a]*B.Tech, Department of Software Engineering , SRM University, Chennai*

E-mail: jindal.pranika13@gmail.com, varun.jaiswal1996@gmail.com

[b]*Assistant Professor, Department of Software Engineering, SRM University, Chennai*

E-mail: uma.m@ktr.srmuniv.ac.in

Abstract: Sentiment Analysis aims to determine the emotions of a writer, speaker , from a piece of text we analyse the opinion of a person. Social Media is generating a large amount of sentiment rich data in the form of tweets, status updates etc. Sentiment Analysis of twitter data is difficult as compared to other media due to the misspellings and usage of slang language. In this research paper we tried to gather data about various airlines and then tried to analyse person's opinion as negative, positive or neutral and also showcase the growth of a particular airlines.

Keywords: *Sentiment Analysis, Polarity based, Vader, Natural language processing, Natural language tool kit.*

1. INTRODUCTION

Sentiment is a perspective or judgement induced by feelings. Sentiment Analysis is also referred to as opinion mining because in this analysation we are extracting a person opinion and then analysing his feelings. Also it helps to know if there has been any change in public opinion about a particular product. The current and main work in sentiment analysis is based on assigning words into categories of negative, positive or neutral. Social media has evolved to become a source of varied kind of information. People on social media discuss their opinions about current issues or the products they are using in their daily life. These opinions and discussions play a major role in knowing the sentiment about a product .Sentiment Analysis is critical because it helps you to know customers opinions about the product.

Sentiment Analysis is not once and done effort. The company needs to have a check regularly in order to know where they lag behind and in which area they need to focus more. Any irregular increase or decrease in the sentiment score of the product will help to know more about the downfall or popularity of a product.

In this paper we looked at one of the most popular social media, Twitter. Data used in this paper are the tweets extracted from twitter about various airlines in USA. The data consists of tweets made by people about Virgin America, Delta, United, Southwest, US Airways, American airlines of USA. We will extract the data from twitter and then analyse the polarity of every tweet made and hence visualize them as negative, positive and neutral. To process the data, use of VADER (Valence Aware Dictionary for sentiment Reasoning) and Python's NLTK (natural language tool kit) library will be made. Vader is the most popularly known rule based model

used for sentiment analysis. The validation, development and evaluation of vader have been briefly described by Hutto and Gilbert in their tremendous work of Vader model. Vader performs exceptionally well in the social media analysis. It requires no training data and is fast enough for online streaming data. Vader is available with python's NLTK library. Natural language tool kit helps to tokenize the sentences into array of words and also converting the text data into the format which can be used for sentiment analysis. Token represents a small unit of text such as a word. It also provides an introduction to programming for language processing. It is written in Python and distributed under GPL license. It supports research and training in natural language processing. It is a set of program modules, data sets, tutorials and exercises which covers symbolic natural language processing.

2. BACKGROUND AND LITERATURE SURVEY

One of the most basic problem in sentiment analysis is categorization of polarity. For example if there are two humans A and B and each of them has formed their own corpus then there may be some words or phrases which may be positive for A and maybe negative or neutral for B. So in such case we can define that there is a problem while categorization of polarity. Moving to the background of sentiment analysis it was seen that sentiment analysis was developed in order to know the customer opinion about a product. In 2011, Nikola Kocic described about the model that dealt with the phenomenon where "technology meets application method". The described method was a combined approach of twitter and the Microsoft Dynamics CRM model. The method was developed so as to analyse the customer opinion by making use of customer relationship management model. Many challenges were still coming in the way of development of sentiment analysis. These challenges were described by Bo Pang and Lillian Lee. They described one of the major challenges in sentiment analysis which was limitation of words in corpus. A writer may have written the words which may not be available in the corpus. This was the major till vader came into play. In 2014, C.J Hutto and Eric Gilbert came with most efficient rule based model called as VADER which overcame all the challenges. The Vader is meant to deal to social media data. Vader now has a corpus of nearly 4500 words, not only that it now consists of various emoticons and the slang words(like LOL).So vader solved all the challenges which were due to limitation of words in corpus.

3. METHODOLOGY

The sentiment analysis can be classified into 3 categories as word level, sentence level and document level. In this paper we used the word and sentence level polarity. We will be calculating the polarity of the words and then the polarity of whole sentence which is known as the compound polarity. Our approach involves the usage of Vader sentiment analysis. Vader is one of the most popularly known library used for social media analysis.

$$S_w = \frac{1}{p} \times \sum_{j=1}^{p} s_{cj} \tag{1}$$

Formula (1) is used to find the polarity of the sentence the word w is the average of the sentiment score of the characters $c_1, c_2, \ldots c_p$. According to this formula a character and a word is given at most the score 1 and at least the score -1. Let us take an example of the words "AB" (good people) and "CB" (bad people). There are 8,024 characters in positive opinion words and 22,024 in negative opinion words. The character "B" (people) appears in positive opinion for 80 times and 266 times for negative opinion. Therefore, the total score of "B" is 0.04, which is neutral. Similarly the total score of "A" and "C" 0.60 and -1, so at last we conclude the opinion score of "AB" is 0.28 and that of "CB" is -0.52, so AB is a positive word and CB is a negative word.

If a word does not appear in the corpus, only the word whose opinion score is above 0.4 or below -0.4 is taken into consideration and treated as a sentimental word.

So, the basic work flow is we extract the data, convert into readable format such as csv, then calculate the polarity of characters followed by word and finally the whole sentence, and finally we conclude the sentiments using polarity.

4. SYSTEM ARCHITECTURE

System architecture is something which defines the behaviour and more the view of the system. The system architecture of our approach is seen shown in fig 1.

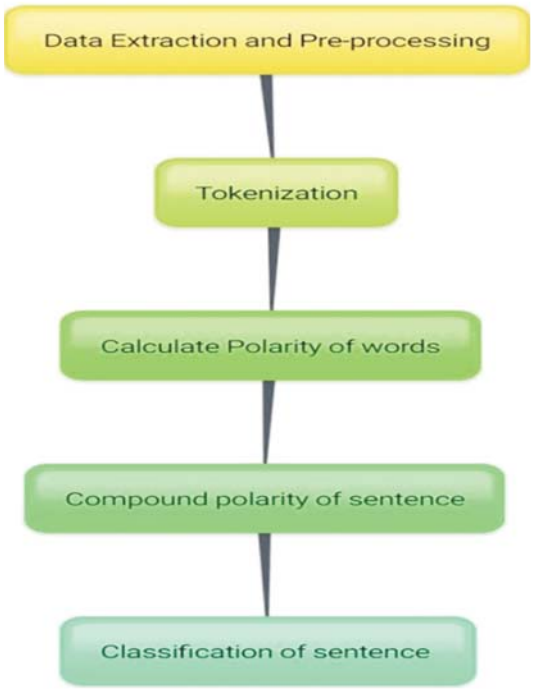

Figure 1: System Architecture

The shown figure displays the various modules possessed in the system. The modules description is described below:

4.1. Data Extraction and Pre-processing

In this paper instead of making our own dataset which will be limited to some thousand words and phrases we fetched the data from twitter which is real time and more reliable. This fetched data contains all the important key fields such as tweets, no of retweets, likes etc. needed for judging the sentiment of people towards a particular subject (in our case flight analysis).

We extracted the data as follows: A) we got all the twitter API keys (API secret, Access token, Access token secret). Access token is used to make authorized calls to twitter's APIs, through this we first obtain an OAuth access token as a twitter user and once we have an access token and token secret we can use different twitter APIs. B) After getting the keys we used twitter streaming API for downloading the data. The streaming API give us the low latency access to twitter's global stream of data. We didn't use the REST API because in REST API functions like parsing, filtering is not available and also connecting to a streaming API requires keeping a continuous HTTP connection open so streaming API is preferred over REST API is also because this gives us a real time extracted data. C) In my source code i.e. in my python script I entered the keys which we got in the previous step and then downloaded the data by executing the program. The main problem with this downloaded data is in its format, the extracted data is in JSON format and unstructured and contains unnecessary words and tags which increases the complexity of analysis and also not needed for sentiment analysis such as hashtags.

In data pre-processing we processed the data into understandable and readable format i.e. we cleaned the data and excluded all the unnecessary data. This is done using pandas package.

Figure 2: Raw Unstructured data

Pandas is a python package that provides fast, flexible to make working with "labelled" data easy. It aims to be the fundamental high-level building block for doing practical, real world data analysis in python

After using pandas library we will get dataset which will be free of unstructured words like hashtags and now using pandas this data can be stored in a csv or JSON format and further we can apply different algorithms as discussed later to do our sentiment analysis.

4.2. Tokenization

Figure 3: Processed data

Tokenization is a process of chopping a given sentence into pieces called as tokens, and also removing the unnecessary characters such as punctuation. For example:

Input: I love to read books, eat food and do shopping.

Output:

I	Love	To	Read	Books
Eat	Food	And	Do	Shopping

Tokenization is done by locating word boundaries. Ending of one word and the beginning of another is called as word boundary. The obtained list of tokens become the input for parsing or text mining. Tokenization is important before any language processing. Tokenization is done using the nltk library of python. Tokenization has various challenges. One of them is linguistic challenge. Tokenization cannot be done for every language. For example, Japanese has small characters and also no white spaces between words and hence becomes very difficult to tokenize. It is difficult mainly for those languages which have no word boundaries. It is kind of pre-processing which identifies the words which needs to be processed. The most basic work of tokenization is to covert a text into words. White space between words makes tokenization easier. If in case there are two adjacent words which have no whitespace but a punctuation symbol, then tokenization assumes the punctuation symbol to be white space and separates the adjacent words. Tokenization is the first step in language processing. So it needs to be done carefully, because any error at this step can cause larger errors at further steps of language processing. The notion of the tokens must be defined before doing any kind of processing. The notion can be linguistic or methodological. So according to the nation any further processing is done.

4.3. Classification of sentence

After calculating the polarity of sentence we can hence classify them as negative, positive or neutral. It is not the way we think it to be. First we calculate the polarity of the word and then calculate the compound polarity. Its not the word polarity that helps us know the sentiment of the writer, the whole sentence polarity help us know the sentiment. The range of polarity can vary from -0.4 to 0.4. All the sentences which have the polarity 0 are considered to be neutral. For example: He told her to run slowly will have 0 polarity, so the sentence can said to be neutral. Any sentence like 'I cannot do that work' will have negative polarity. So according to these we can classify the sentences or we can say the tweets which we extracted.

5. RESULTS

From the research and the sentiment analysis of tweets done we obtained various results. Figure 4 shows the mood summary of various airlines. The mood of various airlines are described as negative, positive or neutral. The vader sentiment analysis helps to calculate the polarities also called as the mood summary. The mood summary is represented using a bar graph. Bar graph has been used because it makes it easy to compare the data and also represents the comparison in a simplified way. The X axis represents the mood ie positive, negative and neutral and the Y axis represents the no of tweets/people who had those moods.

Now we got the know the mood summary of people about various airlines but it is important to know the proportion of tweets of particular airline.

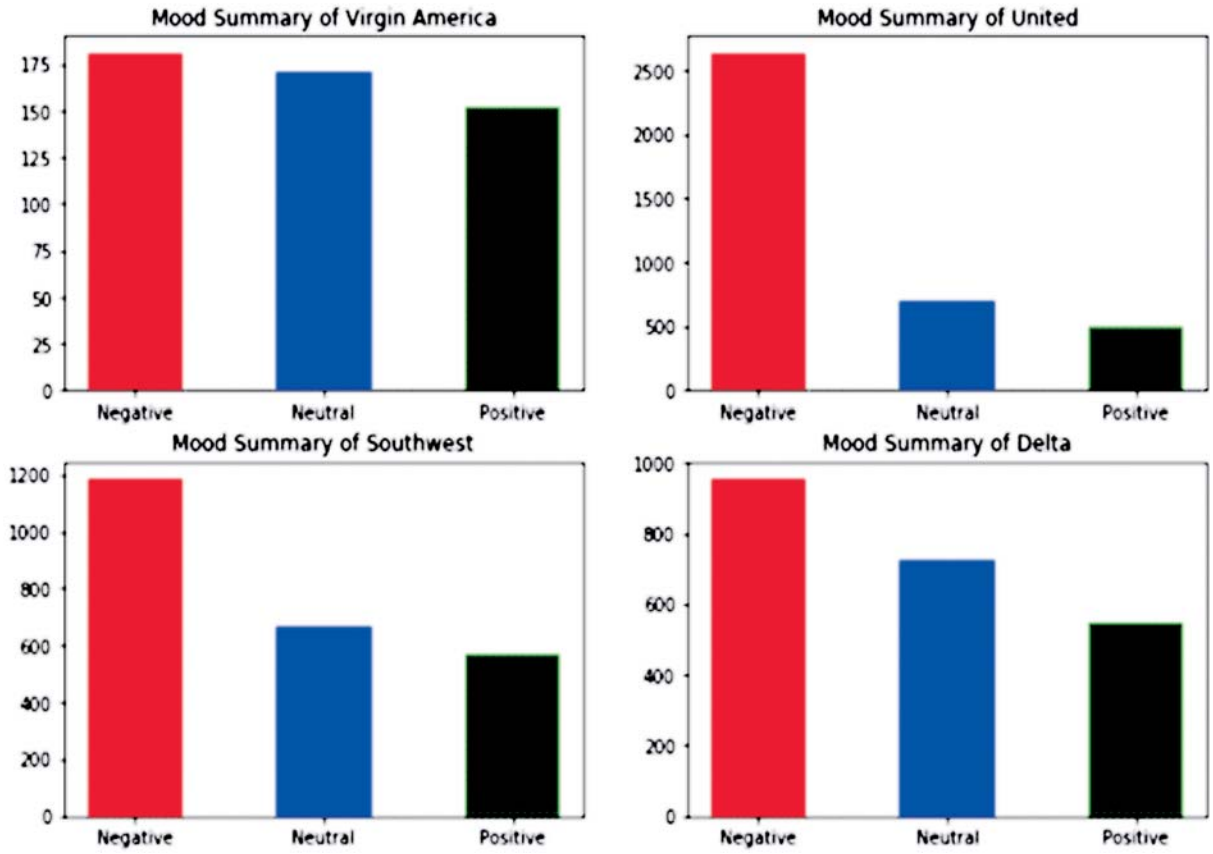

Figure 4: Mood Summary

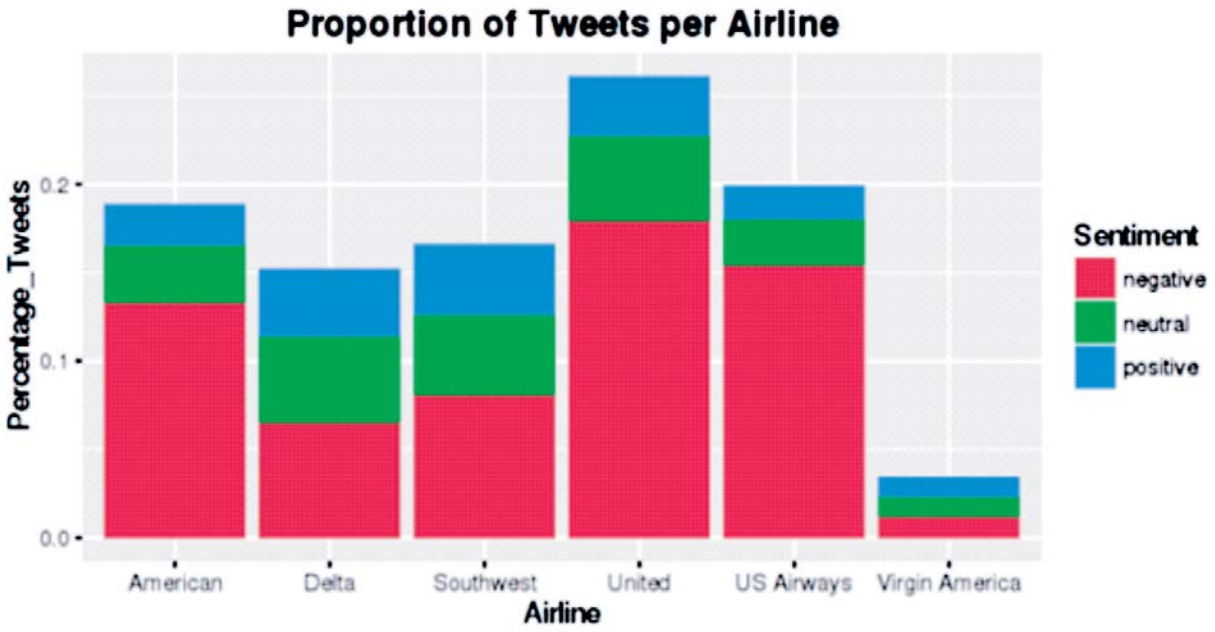

Figure 5: Proportion of tweets per airline

The above figure represents the different airlines on X-axis and percentage of tweets on Y-axis. The proportion represents the percentage of negative, positive or neutral tweets made for particular airline. Every airline would want to know the services where they lag behind so we need to know in which particular area the tweets were made. Figure 5 represents the reasons for the bad sentiment. It represents the frequency of various reasons for bad sentiment. The X-axis represents various negative reasons and the Y-axis represents the frequency of those negative reasons. This graph will help the airline know the various reasons where they lag behind and also where their customers are unhappy.

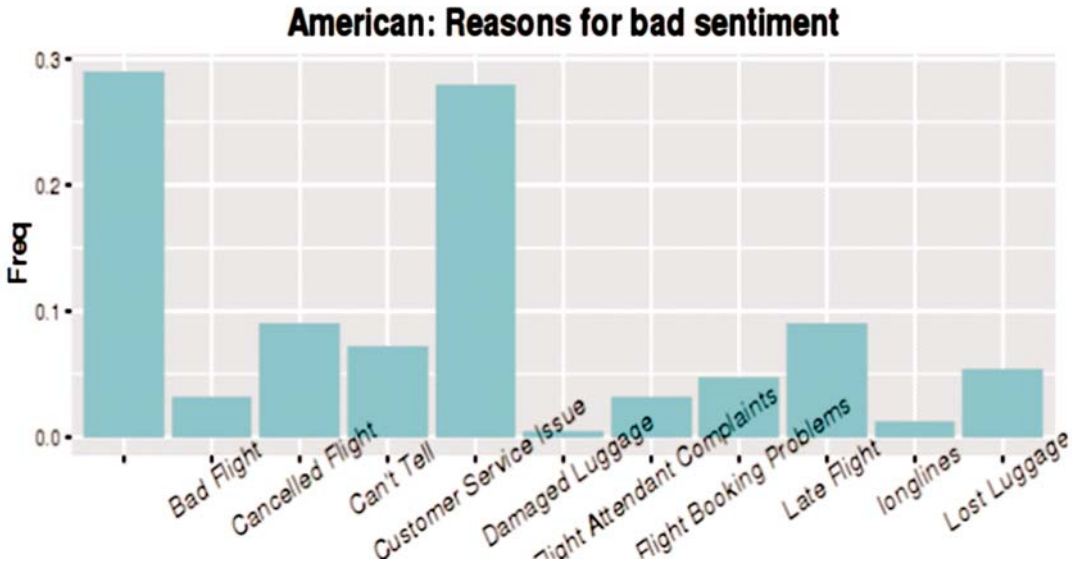

Figure 6: Reasons for bad sentiment

6. CONCLUSION AND FUTURE WORK

This research paper presented an overview of sentiment analysis of Twitter. After this research we may conclude that Vader is still the best option to analyse sentiments of social media. We presented the reviews for all the airlines. This sentiment analysis of airline tweets will help different airlines to know where they lag behind or in which are their customers face problems. Sentiment analysis of airline data will serve as a recommendation to the new users or the existing users. Also sentiment analysis will help a company to boom their business and provide better quality to their customers. The vader still has a limitation over language. It is still made only for English and needs to be developed further for other languages as well. In the future work we look forward to implement software engineering and the software quality attributes in this work. We will try to add the quality attributes to the corpus so that it can help the software companies to know about the reviews of all the software they release and also to know the issues which their customers face and ultimately by knowing the issues they can work on that part and provide the customer with the best quality.

REFERENCES

[1] C.J. Hutto, Eric Gilbert (2014). Vader-A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text.

[2] Agarwal, A. Xie, B. Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of Twitter data. In Proc. WLSM-11s.

[3] Liu, B. (2012). Sentiment Analysis and Opinion Mining. San Rafael, CA: Morgan & Claypool

[4] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations & Trends in Information Retrieval, 2(1), 1–135.

[5] Liu, B. (2010). Sentiment Analysis and Subjectivity. In N. Indurkhya & F. Damerau (Eds.), Handbook of Natural Language Processing (2nd ed.). Boca Raton, FL: Chapman & Hall.

[6] Liu, B., Hu, M., & Cheng, J. (2005). Opinion Observer: Analyzing and Comparing Opinions on the Web. In Proc. WWW-05.

[7] Pang, B., & Lee, L. (2004). A sentimental education: sentiment analysis using subjectivity summarization. In Proc. ACL-04

[8] Nikola Kocic (2011). Performing Sentiment Analysis with Twitter and Microsoft Dynamics CRM 2011

[9] Nielsen, F. A. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In Proc. ESWC-11.

[10] Walaa Medhat**,** Ahmed Hassan, Hoda Korashy (2014). Sentiment analysis algorithms and applications: A survey**.**