International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 04, April 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

TRAFFIC SENTIMENT ANALYSIS ABOUT TRAFFIC REVIEW

Mayuri Kinge, Prof. Sugandha Nandedkar, Prof. Gaurav Narkhede.

[1]PG Scholar, [2]Assistant Professor, Dept. Of CSE, DIEMS, Aurangabad.

[3]Assistant Professor, Dept. of E & TC, MITCOE, Pune

ABSTRACT

Sentiment analysis or Opinion mining is a field of text classification which is very active research field nowadays. There are very researches almost in every area of sentiment analysis but there are not more studies on the domain of traffic hence there is lack of safety. Hence to reduce the traffic related problems this paper proposes the sentiment analysis about the traffic. Twitter incorporates the world knowledge of traffic-specific features in the system which is used to obtain a collection of the review, consisting of the reviewer's opinions about the specific aspects of the traffic in specific area. In this, we implemented the rule-based algorithm to deal with real traffic issues, presented system architecture, building of related bases and overall processing of system. The objective is to classify traffic related review into a sentiment polarity class, positive, negative or neutral, based on those sentences bearing opinion on the traffic alone, leaving out other irrelevant text.

Keywords: Sentiment Analysis, Opinion mining, Reviews, sentiment polarity, Twitter, Traffic.

I. INTRODUCTION

There are many social networking sites are available for interaction between people. And by using social networking sites, much beneficial and significant information can be extracted. So, SA is a method that uses machine learning algorithms by acquiring this information. For understanding people opinions about the specific topic sentiment analysis plays an important role [1]. It helps us to extract what people want, what they think and what are their reactions [3].

So, the main goal is achieving the answer of following questions that are:

- What people give the response?
- How many feedbacks are positive and how many feedbacks are negative?
- Are customer's satisfied or not?

It can also be used as a very useful tool to make decisions considering reviews of users. People are always interested in getting to know the feedback of other users before purchasing and using a particular product. Because this way they are able to find the reliability of specific thing. And at that time sentiment analysis can provide useful information from those comments and reviews of users [2]. Sentiment analysis has a wide range of applications for businesses, organizations, governments and individuals. For instance, a business would want to know customer's opinion about its products/services and that of its competitors. Likewise, governments

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

would want to know how their policies and decisions are received by the people. Similarly, individuals would want make use of other people's opinion (reviews or comments) to make decisions. Also, applications of sentiment analysis have been established in the areas of politics, stock markets, economic systems and security concerns [4] among others but there is not much applications concerned with traffic. To minimize traffic related problem, it is very necessary to monitor traffic in real time so that we can recognize the regions and places that are risky to safety of the people. Hence to reduce the problems sentiment analysis about the traffic is proposed. It is not practically possible to monitor each corner of the transportations network by deploying and maintaining the sensor network. Facebook, twitter, MySpace etc., blogs, forums just because of that social networking sites it is possible to get information of every corner of transportation network. Social media play the important role in collecting the information and recent statuses about traffic flow and transportation system which improves the public safety.

The remaining paper is arranged as follows. Section 2 explains existing work in the field of sentiment analysis. Section 3 explains the detail architecture of proposed system. Section 4 presents the results and comparison between the systems. Section 5 presents the conclusion and future work of the system.

II. LITERATURE SURVEY

Unsupervised learning technique uses some fixed syntactic pattern for classification of reviews. Sentiment word plays an important role in unsupervised learning technique. By using Part of speech tags rules get formed. Part of speech tags are given to the single word in the sentence depending on role of the word in the sentence. A single word can play different roles like noun or adverb or adjective. Adjective plays an important role in sentiment analysis. In this we are using pre-existing corpus called as Penn Treebank corpus which are mostly used standard POS tags [1].

Bhoir and kolte [5] proposed sentiment analysis at aspect level for movie review to find out which aspects of movie are liked and disliked by user. They used SentiWordNet approach to find out orientation of extracted opinion. They proposed two different methods are implemented for finding subjectivity of sentences and then rule based system is used to find feature-opinion pair and finally the orientation of extracted opinion is revealed.

Gann et al. [6] uses twitter tweets and from that he separated 6800 tokens, to each token he applied sentiment value, this value is called as total sentiment value and according to that value he classified positive and negative tokens. TSI for a particular token is calculated as:

$$TSI = (p - tp/tn * n) / (p + tp/tn * n)$$

Where p is the number of times a token appears in positive tweets and n is the number of times a token appears in negative tweets. And tp/tn is the fraction of total no. of positive tweets present and total no. of negative tweets present. Turney [8] presented unsupervised learning technique for classification of review as recommended or not recommended. At starting turney used two seed words excellent and poor. For classification of review he used semantic orientation of phrases. The sentiment of a document is calculated by using average semantic orientation of all such phrases. This approach was able to achieve 66% accuracy for the Similar to Zagibalov and Carroll [9]. Harb et al. [7] used Google's web search engine for making association rule. Beginning with

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

two set of positive and negative seed words he perform blog classification. He counted the no. of positive versus no. of negative adjective in a document for classification of the documents.

III. METHODOLOGY

It includes the following components

- Data collection
- Preprocessing
- Tokenization
- Construction of Bases
- Calculate sentiment polarity

3.1 Data collection

Designing a dataset is the primary step of sentiment Analysis. Social media are playing a very important role in exchanging the information. Nowadays facebook, twitter, MySpace etc., are largely used for collection of data. We used tweets from twitter to carry our research work. Because twitter restricted to use 140 words in tweets, whereas other sites contain large no. of words in review.

- Twitter contains a large number of tweets posts and every day it grows with a large amount. And hence the collected corpus can be huge.
- From regular users to politicians, celebrities, company representatives and even country presidents has an account on twitter. Hence the twitter audience varies. Therefore, it can be possible to collect large tweets from interests groups and different social sites [10]-[12].

Table 1. Examples of Typical Posts from Twitter.

RT @Joydas: Massive Traffic Jams all over India.
RT@DeepikaBhardwaj Bombay gets bad name for traffic. But Bombay traffic is much better than Bangalore and even Delhi.
RT@BavishiChintan Traffic Department here in Rajkot is like rest of India, corrupt and inefficient

Twitter data retrieving

A lot of tweets posted on twitter every day. But traffic related tweets are not available easily in on twitter in large number. Twitter gives us two types of API's, when user enter some keyword about specific topic it search or download tweets.

1. REST API
2. Streaming API

In first type REST indicate Representational State Transfer, in this when user enter the specific keyword, then according to that specific topic twitter retrieves much recent tweets. REST API uses set of logical operators like OR, AND, NOT (e.g., "accident OR vehicle"). REST API limits it can download 3500 tweets per query [15].

In second type, this API requires to put HTTP connection anytime ON, so that it receives the much recent tweets from twitter.

International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 04, April 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

To access the twitter data first we require the authentication of twitter, so for that purpose we have make one client application with twitter. After creating client application with twitter, it gives four keys to user as follows:

1) consumer_key

2) consumer_secrete_key

3) Access_token_key

4) Access_token_secrete_key

By using these four keys twitter gives authentication to user and user can access twitter data. The detail process is as shown in figure 1.

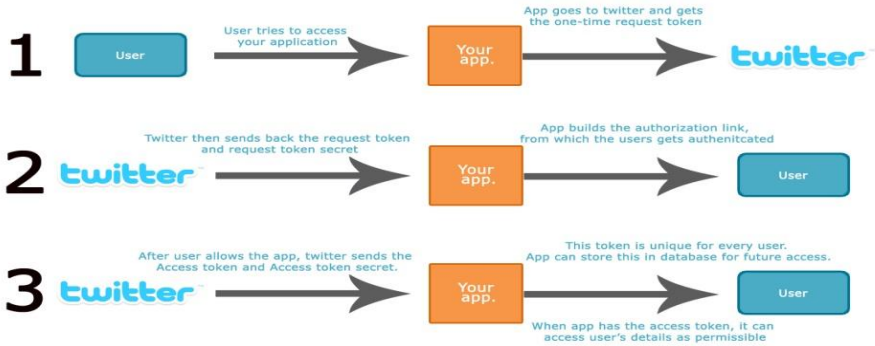

Fig. 1. Twitter Authentication System

As our project is domain based so we extract the traffic related tweets from twitter as shown in figure 3.3

It can filter the delivered tweets according to three criteria:

• Specific keyword(s) to track/search for in the tweets.

• Specific Twitter user(s) according to their user-id's.

• Tweets originating from specific location(s) (only for geo-tagged tweets).

3.2 Data Preprocessing

It is very important step before further processing; it filters the reviews so that it improves accuracy and also removes unnecessary disturbances. It includes elimination of stopwords. Special characters and also Unicode characters like ❤☀★☂♟☯etc. are removed before sentiment analysis. Also front and back slang, wrong spelling, URL's, RT, @, hash tags # are all removed. The questions such as what, which, how etc., are not playing a role for calculating the polarity of sentence hence these are removed. When data from twitter is retrieved then various tweets retrieved more than one time at a time, so it is necessary to remove duplicate line for time complexity. Tweets that are in upper case convert to that in lower case so it make easy for comparison with seed words.

3.3 Sentence Segmentation

Sentence segmentation is very important step usually perform on the document before moving forward. It is also known as sentence boundary detection. It is the process of dividing the document in to number of sentences. This commonly uses punctuation mark such as full stop as a boundary of the sentence.

The example of sentence segmentation is as follows:

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Today's traffic was horrible. People stuck in traffic. People having trouble in getting their vehicles out of the crowd.

This paragraph is segmented by sentence segmenter as follows:

Today's traffic was horrible.

People stuck in traffic.

People having trouble in getting their vehicles out of the crowd.

3.4 Tokenization

Tokenization is the process of dividing the given text or sentence in to tokens and tokens are may be word, phrase, characters or other important unit. The example of tokenization is as follows:

Input: Heavy traffic jam at JM road due to accident.

Output: [Heavy | traffic | jam | at | JM | road | due | to | accident]

3.5 POS Tagging

Training data generally takes a lot of work to create, so a pre-existing corpus is typically used. These usually use the Penn Treebank**.** By recognizing part of word in the sentence it specifies part of speech tag to every word in the sentence. Traditional grammar classifies words based on eight parts of speech: the verb (VB), the noun (NN), the pronoun (PR+DT), the adjective (JJ), the adverb (RB), the preposition (IN), the conjunction (CC), and the interjection (UH), but now there are addition in it. E.g. We saw the yellow dog

"We_PRP saw_VBD the_DT yellow_JJ dog_NN"

3.6 Construction of Related Bases

For this project we required to build related bases. Hence we build sentiment base, modifier base and rule bases.

Sentiment base

Sentiment analysis contains two tightly connected modules sentiment lexicon and its words sentiment polarity. In sentiment analysis our primary task is to establish positive and negative seed set. Examples of positive and negative seed set are p = [beautiful, wonderful, amazing…] and seed n = [bad, awful, poor…]. We manually added various traffic related words, such as accident, U-turn, overload, traffic jam etc. Finally, we constructed positive seed set with 2005 and negative seed set with 4739 words.

A morpheme is the undistinguishable smallest meaningful word in the sentence [15]. It is used for calculating the sentiment polarity. If the morphemes of the word finds in the positive seed set then we consider morpheme of the word is positive and if morpheme of the word finds in negative seed set then morpheme of the word considered as negative. And if the morphemes of the word is not appear in both positive and negative seed set then for that purpose we use the wordnet as a lexical resource. To measure the positive and negative tendencies of the morphemes, we assign the positive and negative weights to the morphemes as follows:

$$\text{weight } p_{c_i} = \frac{fp_{c_i}/\sum_{i=1}^{n} fp_{c_i}}{fp_{c_i}/\sum_{i=1}^{n} fp_{c_i} + fn_{c_i}/\sum_{i=1}^{n} fn_{c_i}}$$

$$\text{weight } N_{c_i} = \frac{fn_{c_i}/\sum_{i=1}^{n} fn_{c_i}}{fn/\sum_{i=1}^{n} fn_{c_i} + fp_{c_i}/\sum_{i=1}^{n} fp_{c_i}}$$

$$S_{c_i} = Weight\ p_{c_i} - Weight\ N_{c_i}$$

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

In formula (3), the polarity S_{c_i} depends on morphemes C_i, and the absolute value of S_{c_i} is the degree of tendency of morphemes C_i .

Steps for calculating the sentiment polarity of the word are as follows:

Scan the positive and negative word lexicons; if the word appears in the positive word lexicon then sentiment polarity is 1 and if the word appears in the negative word lexicon then sentiment polarity is −1. Otherwise sentiment polarity is calculated using weight formulae.

Modifier base

Sentiment of a sentence is determined by the sentiment words. The sentiment of the word is changed by adverbs [14]. Sentiment polarity get opposite if it preceded by negative word. For Example good is positive, but it gets reverse if preceded by the word "not". There are various negation adverbs such as never, no, nobody, none, not. In the same way, We consider the strength to calculate sentiment polarity of degree words. We give the score to the degree words such as very, more, most, much, fully, extremely, less, least, little etc.

Rule base

In rule based approach our first task to build rule set. The rule is formed by using three factors such as sentiment word [S], Negation Word [N], Degree words [D].

Table 2. Construction of various rules

Rule	Example	Characteristics	Formula
S+D	good+very safe+more	S is usually an adjective	
D+S	very+cool high+temperature	adverb+adjective adverb+verb adjective+noun	$p = p_s * p_d$
N+S	not+safe	Negation of adjective	$p = -p_s$
N+D+S	not+very+safe	Similar as D+S	$p = -(1/3) * p_s * p_d$

In this paper, we use a sentiment polarity score to express the sentiment of a text. We give the score to each sentiment word in our dataset of seed word. Consider p value of sentiment polarity of the sentence. p_s is the value of sentiment word. p_d is the score of degree word. We calculate the sentiment polarity of sentence by using rule is as follows: For Example value of sentiment word "expensive" is 2, i.e. p_s is 2. Value of degree word "more" is 3. i.e. p_d is 3. So according to the rule stated in table 1, the value of more expensive is 6. Value of not expensive is -2 and value of not more expensive is -2.

3.7 Calculate Sentiment Polarity

Classification of text can be done on three different levels such as word, sentence and document level. Figure 2 shows the complete working of proposed work. It consists of mainly two steps document sentiment aggregation and sentiment polarity calculation. We first retrieve the data from twitter, if it in document level we consider that whole document as one atomic unit and then do further processing. Then After that we decompose the document and divide into sentence level. Then calculate the polarity of every sentence. Then finally we calculate the overall polarity of all sentences by using aggregation polarity calculation.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

The polarity of each sentence s_i is calculated by using SND rule such as p_s ,p_d. Practically for calculating the overall polarity of the document weight is considered. We define overall problem explanation is: consider for a given document d contains the number of sentences $s_1, \ldots, s_n$ as an input the system. Then calculate the sentiment polarity P_t of each sentence. Then calculate the overall polarity score of document. If $P_t > 0$, then we consider the final polarity is positive otherwise document is negative.

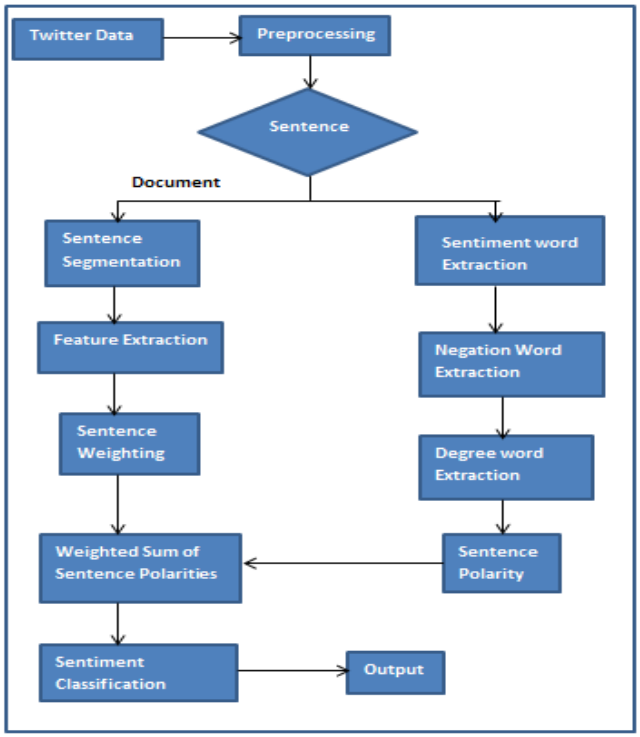

Fig. 2. Overall process of rule based algorithm.

IV. EXPERIMENT AND RESULTS

In sentiment polarity evaluation, we use the confusion matrix to solve the evaluation problem [16]. To measure the performance of algorithm we require to consider four factors these are accuracy, recall, precision and F1-measure. Table 3 shows the confusion matrix. We require four values to measure performance are as follows:

Table 3. Confusion Matrix

#	Predicted Positive	Predicted Negative
Actual positive cases	Number of True positive cases(TP)	Number of False Negative cases(FN)
Actual Negative cases	Number of False positive cases(FP)	Number of True Negative cases(TN)

TP- True Positive: Instance was positive and system classified it as positive.

TN- True Negative: Instance was Negative and system classified it as Negative.

FP- False Positive: Instance was Negative and system classified it as positive.

FN- False Negative: Instance was Positive and system classified it as Negative.

The document classification accuracy is stated as:

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

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

The recall and precision rates in positive case computed as follows:

Precision = $\frac{TP}{TP+FP}$, Recall = $\frac{TP}{TP+FN}$, F1 = $\frac{2\times Precision\times Recall}{Precision+Recall}$

The recall and precision rates in negative case as follows:

Precision = $\frac{TN}{FP+TN}$, Recall = $\frac{TN}{FN+TN}$, F1 = $\frac{2\times Precision\times Recall}{Precision+Recall}$

We calculate the mean of all these 4 factors i.e. accuracy, precision, recall, F1_measure by using:

$$mean_s = 1/n \sum_{i=1}^{n} x_{si}$$

Dataset:

We collected data from different cities and for different time period rather than collecting the dataset at the same time because of that we collected a lot of generalization of data and are from different cities. So we collect the data from five different months 1st to 30th September 2016, 1st to 31th October 2016, 1st to 30th November 2016, 1st to 31th December 2016, and 1st to 31th January 2017 and from three different cities Mumbai, Pune, Aurangabad.

Table 4. Comparison between both algorithms

Algorithm	Previous(Chinese) Algorithm	Our Algorithm
Accuracy	82.45	86.0
Precision positive	30.52	85.48
Recall positive	84.64	91.37
F1-measure positive	253.92	274.13
Precision Negative	98.31	86.14
Recall Negative	82.25	78.57
F1-measure Negative	246.75	235.71

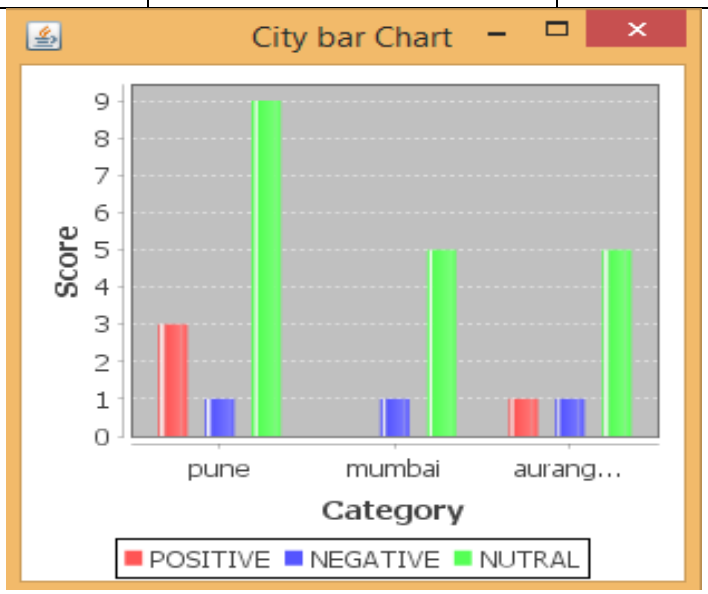

Fig. 3. Sentiment Analysis For Different Cities

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

V. CONCLUSIONS AND FUTURE SCOPE

We have proposed sentiment analysis about traffic to analyze the traffic related issues in new perspective. In proposed system we have used rule based algorithm for English language. We used both adverb and adjective instead of using adjective only to obtain the accurate results in system. System performs very well if we collect polarity bearing words about specific topic.

In future scope this system can work for complex sentences. Adjective and adverb score files can be improved by adding more sentiment bearing words to it. By combining the twitter, GPS, loop detector and camera data it can analyze real time traffic flow. Analyzing the traffic accidents and congestion by using heterogeneous data sources. Calculating time required for traveling.

BIBLIOGRAPHY

1. Bing Liu, Lei Zhang. A survey of Opinion Mining and SentimentAnalysis. Mining Text Data: 415-463.
2. Pang B, Lee L 2008 Opinion mining and sentiment analysis. Foundations and trends in information retrieval, 2(1-2):1-13.
3. Karamibekr M, Ghorbani A A 2012 Sentiment analysis of social issues. Int. Conf. on Social Informatics, 215-221.
4. Vohra S, Teraiya J 2013 Applications and Challenges for Sentiment Analysis: A Survey. Int. Journal of Engineering Research & Technology (IJERT), 2(2):1-5.
5. Purtata Bhoir, Shilpa Kolte, "Sentiment analysis of movie reviews using lexicon approach" IEEE Intell. Syst. March 2016.
6. Gann W-JK, Day J, Zhou S (2014) Twitter analytics for insider trading fraud detection system In: Proceedings of the sencond ASE international conference on Big Data.. ASE.
7. A. Harb, M. Planti, M. Roche, A. Harb, M. Planti, M. Roche, N. Cedex, and A. Harb, "Web Opinion Mining How to extract opinions from blogs, Categories and Subject Descriptors."
8. P. D. Turney,"Thumbsup or thumbsdown?: Semantic orientation applied to unsupervised classification of reviews," in Proc. 40th Annu. Meet. Assoc. Comput. Linguist., 2002, pp. 417–424
9. T.Zagibalov,J.A.Carroll, "Automatic Seed Word Selection for Unsupervised Sentiment Classification of Chinese Text".Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008),Manchester, August 20081
10. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of LREC 2010 (2010)
11. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford (2009).
12. Gebremeskel G 2011 Sentiment Analysis of Twitter posts about news. Master's Thesis, University of Malta.
13. Twitter Search APIs, https://dev.twitter.com/docs/using--- search.4. Surve M, Singh S, Kagathara S, Venkatasivaramasastry K, Dubey S, Rane G, Saraswati J, Badodekar S, Iyer A, Almeida A, Nikam R,

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Perez C G, Bhattacharyya P 2004 AgroExplorer: a meaning based multilingual search engine. Proc. Int. Conf. on Digital Libraries (ICDL), Delhi, India, 1-13.

15. G. Li, C. Wan, H. Bian, L. Yang, and M. Zhong, "Emotional detection of text in the financial domain based-morpheme," J. Comput. Res. Develop., vol. 48, no. z2, pp. 54–59, 2011.
16. N. Kobayashi, K. Inui, Y. Matsumoto, K. Tateishi, and T. Fukushima, "Collecting evaluative expressions for opinion extraction," in Natural Language Processing—IJCNLP 2004, K. Y. Su, J. Tsujii, J. H. Lee, and O.Y.Kwong,Eds. Berlin,Germany:Springer- Verlag,2005