Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 01, January-2016

An Approach for Identifying Road Traffic Information using Opinion Mining from Twitter Messages

Dr. Ananthi Sheshasaayee
Research Supervisor, PG and Research Department of Computer Science & Applications,
Quaid-E-Millath Government College for Women (Autonomous), Chennai, TamilNadu, India

R. Jayanthi
Research Scholar, PG and Research Department of Computer Science & Applications,
Quaid-E-Millath Government College for Women (Autonomous), Chennai, TamilNadu, India

***Abstract*—In recent years, online conversation has become very popular. Common people share their thoughts, views, ideas and comments on their topic of interest on the social networking sites. The important thing is the content generated from these social media sites remains mostly unused. These contents in terms of text data are preserved like historical data. Although there is a growing research concerned with retrieval of data from social media, the information extraction using social media data is limited because of its unstructured format. This article presents a text mining approach to identify the road traffic information collected from microblogs like twitter messages. The result shows the challenge in interpreting the unstructured text and extracting useful information from it. This approach provides an efficient way to extract opinions using Text mining techniques integrated with supervised learning methods, classification and rated against an evaluation criteria.**

Keywords—Text mining; opinion mining; classification; Information Extraction; Social Media data

I. INTRODUCTION

In the recent years, there is abundant availability of electronic data whereas traditional means of expressing opinions or reviews was through oral or written medium. Subsequently e-mails and other electronic media became popular. Ideas, feelings, opinions are expressed frankly through social networking sites. One's ideas, thoughts, comments, likes, dislikes are shared as their opinions. As these opinions are one's freedom of right to think, they are available in a large volume in social media. These opinions in the form of text can be used to extract valuable information more objectively. The extraction of information hidden inside the unstructured text gives an organization a competitive edge hence the knowledge gained from the social media data contributes for effective decision making[1].

Now-a-days people not only share the information, they also comment on the interesting information like rate the products, important decisions, movies, health care, traffic on roads, weather and what not. The receivers of the information not only reads the information but in turn they actively participate in the social media and also contribute new pieces of information. This in turn produces large data in the social media sites. The opinionated information is an important part of textual data which influences better decision making [2]. Many internet users use microblogging sites such as Twitter, Facebook, LinkedIn, PInterest, etc., to share their information. Those information would be informal descriptions, mostly unstructured and doesn't follow any language grammar. This is an emerging field that attracts the researchers to discover new patterns from human generated content[3].

This article focuses on collecting such user-generated content on road traffic from twitter messages. The focus of this approach is to extract the road traffic information from text classification and the results were registered and the evaluation measures were calculated and compared with the traffic details of Google Maps[4].

The rest of the article is organized as follows, the related work is discussed in Section 2, Section 3 portrays the system architecture of opinion miner and Section 4 presents the Data set and Information Extraction. Experimental Results are discussed in Section 5 and finally Section 6 concludes the paper with the future direction of research in this area.

II. RELATED WORK

First, Microblogs are used to create short messages. They provide light weight, easy and fast way of communication. Twitter is a famous microblog used to send short stream of messages called "tweet". Messages are limited to 140 characters or less, but that’s more than enough to post a link, share an image, or even trade thoughts with favorite celebrity or influencer. The influence of social media data for research and how the content can be used to predict real-world decisions that enhance business intelligence, by applying the text mining is discussed in [5].

The ontology based approach is discussed by Efstratios K,Christos B, et al., that posts are not simply characterized by a sentiment score, as is the case with machine learning-based classifiers, but instead receive a sentiment grade for each distinct notion in the post[6].

In [4], a geographic approach has been proposed which provides a reliable quantitative indicator of the usefulness of messages from social media by leveraging the existing knowledge about natural hazards such as floods, thus being valuable for disaster management in both crisis response and preventive monitoring.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 01, January-2016

Raymondus Kosala, Erwin Adi, Steven discussed the virtually indeterminate sources of data on Twitter, they proposed an algorithm to measure the traffic information confidence level for the real-time traffic monitoring system. Based on the test results of this study, the system could be used to serve its intended purpose[7].

III. SYSTEM ARCHITECTURE

Twitter messages on road traffic details are collected, extracted, classified and evaluated using opinion mining system. This approach consists of different phases like collecting tweets related to road traffic, pre-processing, generate Parse tree using POS tagging with rule-based grammar, feature selection and feature extraction, classification of extracted features using Naive Bayes classifier, finding semantic word similarity and polarity assignment using fuzzy logic and finally evaluation and visualization of opinioned data. The system architecture for the road traffic monitoring opinion mining system in shown in the following Fig.1.

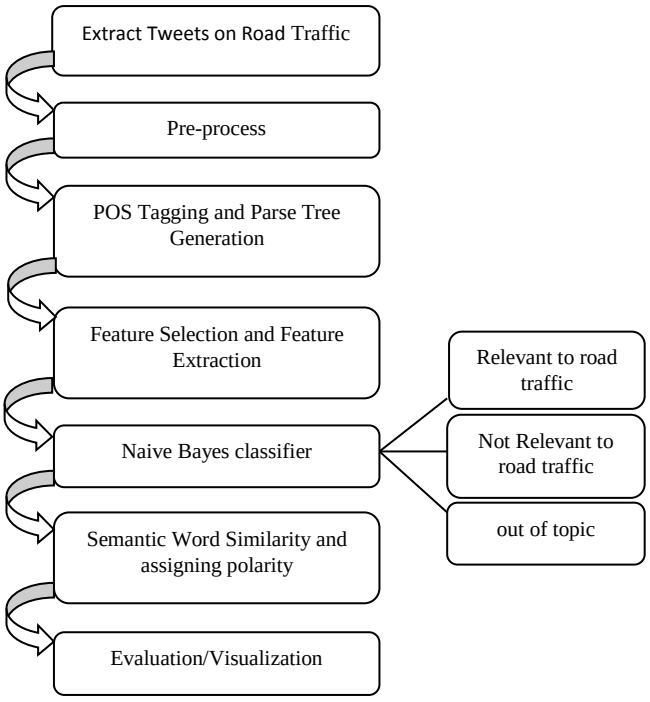

Fig. 1. Road Traffic monitoring opinion mining system

IV. DATASET

Tweets of Chennai city road traffic during flood were collected from twitter collection Chennai Traffic, Chennai City Traffic and from their followers. Initially, all informal user generated traffic content needs to be pre-processed before considering the data for the progress of the research.

The collected data must be classified into following categories as shown in Table-I: *out of topic* - refers to the road traffic tweets not related to Chennai city during flood, *not relevant to the topic* - refers to the road traffic tweets but not relevant to the exact traffic details, *relevant to the topic* - refers to the relevant information which contributes to the exact situation.

TABLE I. CATEGORIZATION OF TWEETS

Out of topic	gavaskee.. @gavaskee · Nov 30 #chennaiweather Too high from morning today #chennaitraffic Interesting is reduced Petrol Diesel Price check
Not relevant to the topic	K Balakumar @kbalakumar · Nov 30 People, braving rains, standing on streets & guiding traffic away from ditches & craters. Kindness of anonymous souls overwhelming. #Chennai Kamalraj Duraiswamy @kamaldurai · Dec 1 #chennairains Rain has started the next innings.. Hope this should not be another traffic day _/_ #chennaitraffic
Relevant to the topic	Sarthak Saraswat @sarthaks007 · Nov 30 Santhome High Road flooded. Better to avoid it. #ChennaiRains #chennaitraffic M prabhu @prabhu_sr · Dec 1 @chennaiweather heavy traffic in gst road #saidapet towards #guindy #rain #chennai M Siva.Ramakrisshna @srkfilmmaker · Dec 1 Chennai 100feet road, Anna salai mount road, like... Almost complete traffic closed... More than 7 feets flood going threw roads....

V. EXPERIMENTAL RESULTS

The experimental setup is divided into following phases and is developed using Python Programming with Natural Language Toolkit.

5.1 Pre-processing

For the better performance of the opinion mining system the collected data need to be preprocessed. Each Twitter post contains properties like *usernames*, *Hash tags*, and *Re-tweet*. The usernames contains '@' symbol which is a de facto standard must be followed. Hast tags(#) are allowed in Twitter data to represent the content of the tweet. Eliminate the tweets that are not in English. Remove the stop-words like digits, special characters, smileys, etc.,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 01, January-2016

This phase is very important, because extra effort had to be applied to establish uniformity of syntax and interpretation of the expressed opinion. The preprocessing of the sample data is shown in Fig.2 with stopword removal and tokenization.

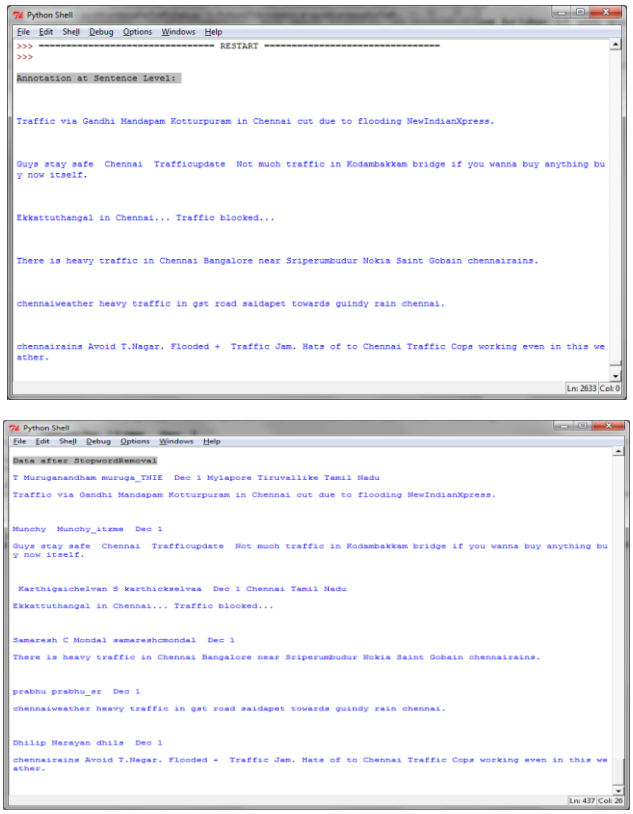

Fig. 2. Pre-processing : Stopword Removal and Tokenization

5.2 POS Tagging and Parse-Tree Generation

After pre-processing the text data, the informal descriptions are converted to formal descriptions and now ready for the next phase. The sentences which follows SVO pattern are extracted by applying Part-of-Speech tagging. Grammar chunks are framed for Noun Phrase and Verb Phrase using bigram Rule based grammar[8]. This phase mainly applies the POS tagging with the focus to recognize nouns, verbs and adjectives and parse tree is generated as shown in Fig.3.

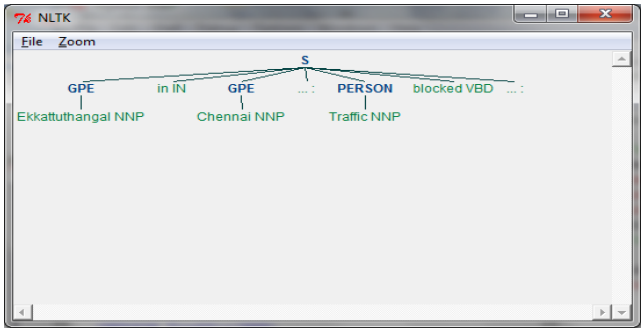

Fig. 3. Parse Tree Generation

5.3 Feature Selection and Feature Extraction

Information Extraction which combines feature selection and feature extraction are found to improve the representation of data from redundancy. The proposed approach uses domain-specific feature knowledge base for feature extraction. Features of interest on traffic data like *out of topic*, *not relevant to the topic* and *relevant to the topic* are collected accordingly. From the parse tree, the Noun, Verb, Adjective and Discriminator tags are classified using Noun chunks and Verb chunks. and it is shown in Fig.4.

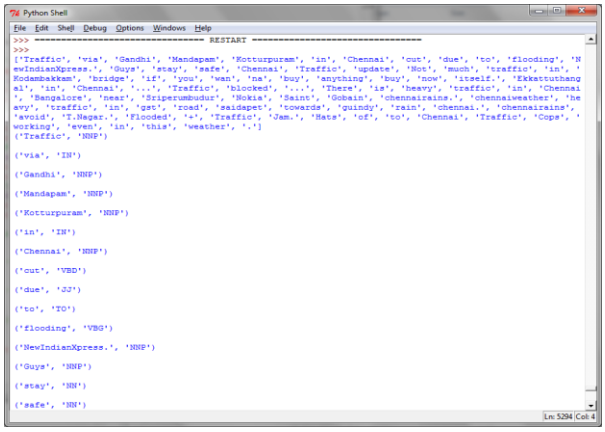

Fig. 4. Feature Generation and Feature Selection

5.5 Classification

Naive Bayes classifier is a supervised learning classification method works well on text categorization. Assume the class attribute C contains the set of tweet value attribute $T(t_1,t_2,.....t_n)$, then the Naive Bayes classifier finds the conditional independence. For the given class attribute value, other feature attributes are conditionally independent and that can be calculated using

$$P(Ci|T) = \frac{(\prod_{j=1}^{m} P(tj|ci))P(ci)}{P(T)}$$

This classifies the tweets into conditional independent categories named official tweets, media reports, public response and volunteer opinions. The total tweets collected on the below mentioned date as shown in Table-II under each categories are further classified as relevant(R), not relevant(NR) and not of topic(N) class as discussed in chapter 4.

TABLE II. CATEGORIZATION OF TWEETS

Date	30 Nov 2015				01 Dec 2015			
	#	%			#	%		
		R	NR	N		R	NR	N
Total tweets	59	36	27	37	126	22	60	18
Official tweets	5	40	60	0	4	25	50	25
Media reports	2	0	100	0	12	25	50	25
Public response	41	30	24	46	102	22	64	14
Volunteer tweets	11	63	10	27	8	12	38	50

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 01, January-2016

5.6 Evaluation and Result Analysis

A set of keywords related to amount of traffic is rated against heavy, average and no traffic and are compared by finding the word-similarity semantically with the help of WordNet dictionary as shown in Fig.5.

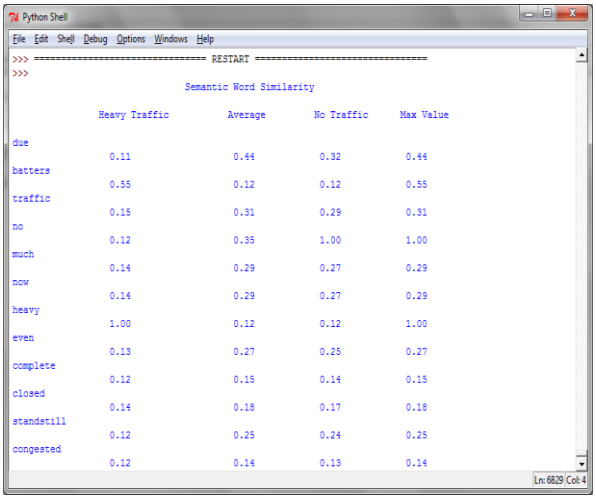

Fig. 5. Semantic Word Similarity using WordNet

The proposed method focus on opinions with heavy traffic, average and no traffic to determine the semantic orientation of a tweet instead of positive, negative and neutral. The opinion on traffic details for the sample data collected is given in the following Fig.6.

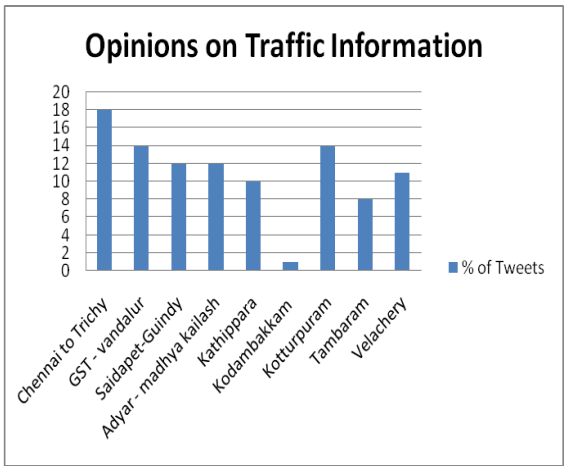

Fig. 6. Traffic Information from Twitter Messages

The extracted word lexicons are difficult to process using computational linguistic methods. This encouraged us to convert the lexicons into discrete values ranging from heavy traffic, average traffic, and no traffic. Fuzzy logic is applied to convert continuous features to distinct values and this is achieved by changing the linguistic variables to numerical values[9][10]. Each lexicon is assigned a numeric value by finding word similarity with the set of polarity words for road traffic {No, Average, High}. From the above graph, the range of high value is assigned from 20-11, the average value is assigned from 10-4 and 0-3 is assigned to no traffic value. Opinions expressed in this manner are easy to understand and facilitate to interpret the result easily.

By comparing the numerical values with the opinions plotted in the graph shown in Fig.6, *Heavy traffic* is in Chennai-Trichy, GST-Vandalur, Kotturpuram, Adyar-Madhya Kailash, Saidapet-Guidy, Velachery, Kathippara. *Normal traffic* in Tambaram and with *Minimum traffic* in Kodambakkam.

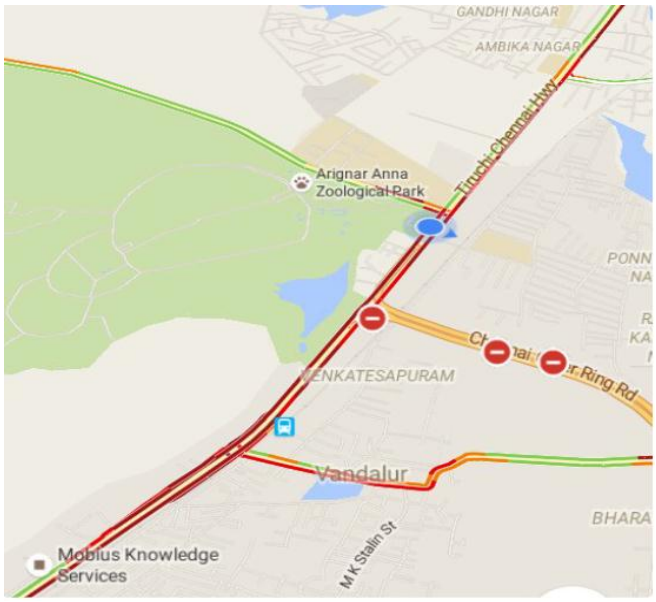

Fig. 7. Partially viewed Traffic detail from Google Map

To further examine the comparison of the opinions from twitter traffic data, the red lines of Google Maps image in Fig.7 confirms the heavy traffic in Chennai-Trichy road and GST-Vandalur. Thus the proposed approach can offer an efficient way to extract opinions using Text mining techniques integrated with supervised learning methods, classification and rated against an evaluation criteria.

VI. CONCLUSION

Opinion mining is a special field of Text mining. The goal is to extract opinions from different forms of social media data and analyze the opinions of the users on a particular topic with the conclusion of relating with criterion values. This article proposed an opinion mining approach to extract opinions from traffic related tweets and using classification methods, a training data set were built to improve the performance of the model and obtained good result compared with the Google Maps traffic images. The proposed model demonstrated with an experiment using sample data collected from twitter microblog and developed a working prototype of opinion mining system pertaining to social concern- monitoring traffic on a rainy day. The future work is to fine tune the model and make it generic opinion mining system so that when any opinion is given as input the influence of the opinion on decision making is measured and interpreted.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 01, January-2016

REFERENCES

[1] Wu Hea, Shenghua Zha, Ling Li, "Social media competitive analysis and text mining: A case study in the pizza industry", International Journal of Information Management 33, pp. 464-472, 2013.

[2] S Padmaja, S Sameen Fatima, "Opinion Mining and Sentiment Analysis –An Assessment of Peoples' Belief: A Survey", International Journal of Ad hoc, Sensor & Ubiquitous Computing (IJASUC) , vol.4, No.1, pp. 21-33, February 2013.

[3] J. H. Kietzmann, K. Hermkens, "Social media? Get serious! Understanding the functional building blocks of social media." Business Horizons, vol.54(3), pp. 241-52, 2011.

[4] João Porto de Albuquerque, B. Herfort, A. Brenning, Z. Alexander,"A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management", International Journal of Geographical Information Science, vol.29, Issue 4, 2015.

[5] S. Ananthi, R. Jayanthi, "Exploring the potential of Social Media Data using Text Mining to augment Business Intelligence", An International Journal of Advanced Computer Technology, vol.3 (4),pp. 738-742, April 2014

[6] K. Efstratios, B. Christos , "Ontology-based sentiment analysis of twitter posts", Elsevier- Expert Systems with Applications vol.40, pp.4065–4074, 2013.

[7] K. Raymondus, A. Erwin, Steven, " Harvesting Real Time Traffic Information from Twitter", International Conference on Advances Science and Contemporary Engineering- Elseview 2012.

[8] A. Erik Cambria, A. Hussain, "Sentic Computing: Techniques, Tools, and Applications", Springer 2012.

[9] J. Shaidah and Hejab M. Alfawareh, "Applying fuzzy sets for Opinion Mining", 978-1-4673-5285-7/13, IEEE 2013.

[10] H. A. Md, T. Rahman, "Sentiment Analysis by using Fuzzy Logic", International Journal of Computer Science, Engineering and Information Technology, Feb., vol.4, Issue 1,pp.33-48, 2014.

[11] Po-Wei Liang and Bi-Ru Dai, "Opinion Mining on Social Media Data", IEEE14th International Conerence on Mobile Data Management, 2013.

[12] S. Poornima, S. Gayatri , "Opinion Mining on Social Media: Based on unstructured Data", International Journal of Computer Science and Mobile Computing, vol. 4, Issue 6, pp. 768-777, 2015.