International Journal Of Advanced Research In Engineering Technology & Sciences ISSN: 2394-2819

Email: editor@ijarets.org June- 2015 Volume 2, Issue-6 www.ijarets.org

A Survey on Traffic Sentiment Analysis

Mayuri Kinge
PG Scholar
Dept. of CSE
DIEMS, Aurangabad

Prof.Sugandha Nandedkar
Assistant Professor
Dept. of CSE
DIEMS, Aurangabad

Prof. Gaurav Narkhede
Assistant Professor
Dept. of E & TC
MITCOE, Pune

ABSTRACT:

Sentiment analysis or opinion mining is a machine learning approach in which machines analyzes and classifies the human's sentiments, emotions, opinions etc in the form of positive, negative or neutral comments underlying the text. The internet technology continuously growing technology which adding large number of users everyday .The users not only use web but more often they give their responses or rather feedbacks, which can be help in decision making.

There are various researches in sentiment analysis and its related areas but there are no more studies on transportations, hence there is lack of efficiency and safety. Hence to reduce the traffic related problems, this paper proposes the traffic sentiment analysis (TSA).This survey will try to focus on challenges in SA, related work for automated web data crawling, different levels of SA, subjectivity classification, some machine learning techniques on the basis of their usage and importance for the analysis, evaluation of Sentiment classifications and its recent advancements and the future research directions in the field of traffic Sentiment Analysis.

KEYWORDS: Sentiment Analysis, Opinion Mining, Machine learning.

INTRODUCTION:

Traffic injuries and fatalities are an enormous public health problem. To reduce traffic related injuries and fatalities, it would be helpful to monitor traffic in real time in order to quickly identify regions and activities that have the potential to become a risk to public safety. Hence traffic sentiment analysis is proposed by this paper.

Sentiment analysis concerns itself about the issues of traffic in particular transportation systems, and fairly enough traffic sentiment analysis can be considered as a subset of it. It is clearly impractical to deploy and maintain a large sensor network capable of monitoring every corner of the transportation network, but thanks to the explosion of social media in all forms, including blogs, online forums, face book, and twitter, it should be possible to treat social media as a human sensor network [1]-[3]. This would enable us to collect timely and comprehensive information about the current status of the transportation network and traffic flow to support advanced safety enhancement. To develop a web monitoring system to automatically retrieve tweets related to transportation safety, extract the potential safety topics (e.g., traffic accidents, road flooding), calculate public sentiments was the primary objective of this survey.

Manual training can solve the sentiment analysis problem to a satisfactory level. But an introduction of fully automated system for sentiment analysis which needs no manual intervention has not been done as of yet. Sentiment analysis has various challenges. The first challenge is that the same sentiment is assumed positive as well as negative in different situations. People express their opinions in the different ways is the second challenge of sentiment analysis. On twitter or facebook, some users combine their sentence which contains different opinions which human can easily understand but very difficult to computer to understand or parse.

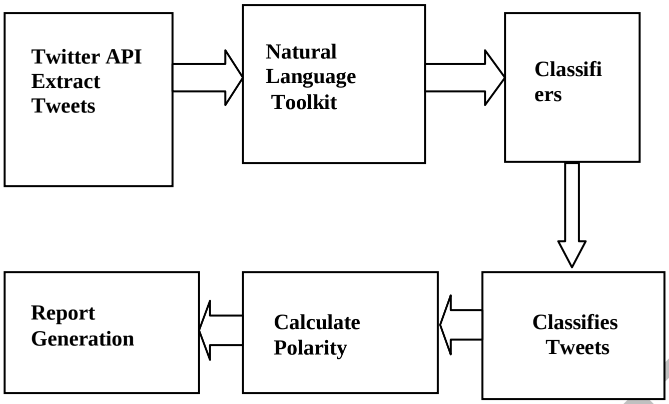

Figure 1: Block diagram of sentiment analysis

AUTOMATIC TWITTER DATA CRAWLING:

Twitter provides two categories of search APIs that allow users to search and download tweets based on search keywords and geographic constraints.

1. REST API

2. Streaming API

In first category Twitter REST APIs, when user enters the queries, it retrieves the recent and most popular tweets according to the indices of the queries entered by the user. The format of the REST query includes a centroid, the radius and a set of keywords with support of operators, also it contain logical operators NOT, AND, OR and EXCLUDE (e.g., "accident OR vehicle").

In second category Twitter streaming APIs, it allows users to keep a continuous HTTP connection to open, it retrieve the most recent public tweets. The format of streaming query includes the combination of the radius and a centroid or a set of keywords with the operators.

These API responds in the form of JSON (JavaScript Object Notation), it uses human readable text to transmit data object. It consists of attribute value pair.

e.g. {"id":"24"}.

DIFFERENT LEVELS OF SENTIMENT ANALYSIS:

A. DOCUMENTLEVEL:

In this document level, it recognizes whole polarity of the document means whether the polarity of the document is positive or negative. For example, for a product review, whether the review gives an overall positive or negative opinion about the product is determined by the system and is commonly known as document-level sentiment classification. Each document expresses opinions on a single entity are the assumption made by this level. Thus the document levelsentiment classification has some advantages and some disadvantages. Advantage is that we can get an overall polarity of an entity from a document. Disadvantage is that it could not extract the different opinions about different featuresof an entity.

B. SENTENCE LEVEL:

Whether each sentence expressed a positive, negative, or neutral opinion is decided by this level. Neutral usually does not give any opinion. This level of analysis contains the subjectivity classification, which distinguishes objective sentences from the subjective sentences. Objective sentence express factual information about an entity and subjective sentences that express emotions and opinions about an entity. Only single opinion contain in simple sentence hence sentiment analysis is easy for simple sentence level. But in complex sentence containsvarious opinions hence sentiment classification is not done [4].

C. ENTITY AND ASPECT LEVEL:

Because of some disadvantages of document level and the sentence level analyses do not recognize peoples thinking about an entity. So finer-grained analysis performed by aspect level analysis. Aspect level is also called as feature-based opinion mining. It does not consider structure of language like sentences, documents, paragraphs or phrases; instead it directly looks at the opinion. It only considers sentiment of the opinion whether it is positive or negative and a target of opinion. An opinion without its target being identified is of limited use. Realizing the importance of opinion targets also help understand the sentiment analysis problem better.

SUBJECTIVITY/ OBJECTIVITY CLASSIFICATION:

There is important challenge of sentiment analysis is subjectivity/objectivity classification. Some text may contain useful information and some text may contain information which is either redundant or not useful for sentiment analysis. Relevant data is contained in a subjective sentence. It includes emotions, opinions, evaluations, beliefs, judgments etc so this type of information is useful for sentiment analysis. While an objective sentence talking about factual information for any topic [5]. Hence while deciding the polarity of sentence the objective sentences can be considered redundant and must be filtered out

NLP TOOLS FOR SENTIMENT ANALYSIS:

D. OPENNLP:

It is a machine learning based toolkit. It is used for the processing of natural language text. It performs the common NLP like tokenization, part-of-speech tagging, named entity extraction, sentence segmentation, parsing, chunking, and co-reference resolution. For performing these tasks more advanced text processing services are required. OpenNLP also contain maximum entropy and perceptron based machine learning. http://opennlp.apache.org/

E. STANFORD CORENLP:

Stanford CoreNLP is present in Java. It contain many NLP tools like the part-of-speech (POS) tagger, the parser, the named entity recognizer (NER), the co-reference resolution system, the sentiment analysis, and the bootstrapped pattern learningtools. It provides model files for the analysis of English, but also for other languages the engine is compatible. This tool is extensible and highly flexible .The goal of this tool is that it should be very easy to apply a bunch of piece of text. It has only two lines of code hence all tools run on it. http://nlp.stanford.edu/software/tagger.shtm/

F. NLTK:

NLTK is used to process human language data. It is a free, open source, community-driven tool. It provides easy interfaces to use. It provides libraries for text processing such as classification, tokenization, tagging, stemming, and semantic reasoning, parsing and an active discussion forum and lexical resources like WordNet. NLTK is suitable for linguists, students, engineers, researchers, educators, and industry users. NLTK is available for different OS like Windows, Linux and Mac OS X. NLTK has been called “a wonderful tool in python for teaching and computational linguistics”. http://www.nltk.org/

SENTIMENT CLASSIFICATION TECHNIQUES:

The sentiment classification is broadly divided into the machine learning approach, lexicon based approach and hybrid approach [6]. The machine learning approach (ML) uses the ML algorithms and linguistic features. The lexicon-based approach is based on a sentiment lexicon, which is the collection of known sentiment terms. The Hybrid Approach combines both the machine learning approach and lexicon based approach.

Machine learning approach of the text classification can be broadly classified into supervised and unsupervised learning methods. The supervised methods use a large number of labeled training documents, but sometimes it is difficult to find these labeled training documents then unsupervised learning methods are used [8]. The syntactic patterns are composed based on part-of-speech (POS) tags. But, the machine learning approach applicable to sentiment analysis mostly belongs to supervised classification in general and text classification techniques in particular [7]. In this supervised classification, two sets of documents are required: first is training set and second is test set. A training set is used to train classifier, and a test set is used to check the performance of the classifier. A number of machine learning techniques have been adopted to classify the comments. Maximum entropy (ME), Naive Bayes (NB), and support vector machines (SVM) have achieved great success in text categorization.

G. NAIVE BAYES CLASSIFICATION :

Naive Bayes method is one of the popular techniques for text classification. Many researchers has proved that it perform extremely well in practice. It is an approach to text classification that assigns the class c* = argmax$_c$ P(c | d), to a given document d. It is a simple probabilistic model. Naive base classifier uses Bayes' theorem. Probability model of naive bayes can be called as "independent feature model". The Bayes’ rule of the Naive Bayes (NB) classifier is eq. (1),

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)} \quad (1)$$

Where, P(d) plays no role in selecting c*. To estimate the term P(d|c), Naive Bayes decomposes it by assuming the $f_i{}'s$ are conditionally independentgiven d’s class as in Eq.(2)

$$P_{NB}\,(c|d) = \frac{P(c)(\prod_{i=1}^{m} P(f_i|c)^{n_i(d)})}{P(d)}(2)$$

Where, m is the no of features and $f_i$is the feature vector.

Text categorization using naive bayes perform very well although its conditional independence assumption clearly does not tend in real-world situations [9]. With highly dependent features Naive Bayes is optimal for certain problem classes [10].

H. SUPPORT VECTOR MACHINES :

At traditional text classification Support vector machines (SVMs) have been very effective. As compared to Naive Bayes and MaximumEntropysupport vector machine have large margin. In the two-category case, the basic idea behind the training procedure is to find a maximum margin hyperplane, represented by $\vec{w}$, that vector separates the document in two classes, and also it finds which margin, is as large as possible. This corresponds to a constrained optimization problem; letting $c_j \in \{1, -1\}$ (corresponding to positive and negative) be the correct class of document d_j, the solution can be written as in Eq.(3),

$$\vec{w} = \sum_j a_j\, c_j \vec{d_j}, \alpha_j \geq 0 \qquad (3)$$

Where, theα_j’s (Lagrangian multipliers) are obtained by solving a dual optimization problem. Those $\vec{d_j}$such thatα_j is greater than zero are called support vectors, since they are the only document vectors contributing to$\vec{w}$. Classification of test instances consists simply of determining which side of $\vec{w}$’s hyperplane they fall on. There are many applications in which SVM are used. To categorize reviews Chen and Tseng [11] have proposed two SVM-based approaches, first is One-versus-All SVM and second is Single-Machine Multiclass SVM. They proposed a classification problem in they calculate the quality of information. To find information oriented feature set they also adopted an information quality (IQ) frame work. SVM worked on various applications like digital cameras and MP3 reviews. The results of SVM showed that this method performs very well and in termsof their quality it accurately classifies reviews.

SVMs were used by Li and Li [12] as a sentiment polarity classifier. By establishing a monitoring system they proved that their mechanism can effectively discover market intelligence (MI) for supporting decision-makers to collect opinions of a business on different aspects in real time.

MAXIMUM ENTROPY :

In natural language processing applications Maximum Entropy (ME) classification is one of the technique, which has also proven to be effective [13]. Sometimes, it gives more accurate result than Naive Bayes at text classification [14]. Its estimate of P(c | d) give the form as in Eq.(3),

$$P_{ME}(c|d) = \frac{1}{Z(d)} \exp \sum_i \lambda_{i,c} F_{i,c}(d,c)) \quad (3)$$

Where, Z(d) is a normalization function. $F_{i,c}$is a feature/class function for feature $f_i$and class c, as in Eq.(4),

$$F_{i,c}(d,c') = \begin{cases} 1 & n_i(d) > 0 \text{ and} c' = c \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

If the bigram appears and the sentiment of the document is considered to be negative a particular feature function might fire. Importantly, the classification techniques like Naive Bayes, Maximum Entropy perform better when conditional independence assumptions are not met,as it does not consider any assumptions about the relationships between features.

TABLE-1: COMPARATIVE STUDY OF DIFFERENT MACHINE LEARNING ALGORITHM FOR CLASSIFICATION

Machine Learning Algorithm	Type	Study	Year	Accuracy
Naive Bayes	Supervised	Pang et al.	2002	81.5
		Dave et al.	2003	81.9-87.0
		Chen et al.	2006	77.5
		gindl and Liegl	2008	66.0
		Go et al.	2009	82.7
		Bifet and Frank	2010	82.5
		Zhang et al.	2011	84.5
Support Vector Machines	Supervised	Pang et al.	2002	82.9
		Chen et al.	2006	84.6
		Annett and Kondrak	2008	77.4
		Go et al.	2009	82.2
Maximum Entropy	Supervised	Pang et al.	2002	81.0
		Gindl and Liegl	2008	83.3
		Go et al.	2009	83
K-nearest Neighbor	UnSupervised	Davidov et al.	2010	66.0-87.0

K-NEAREST NEIGHBOR (K-NN):

To test the degree of similarity between k training data and documents k-nearest neighbor algorithm (k-NN) is used. Then this classification data is stored, for determining the category of test documents. It uses closest feature space in the training set and categorizes objects [15]. Then these training sets are mapped into multi-dimensional feature space. Based on the category of the training set the feature space is partitioned into regions. In the k-nearest training data if the category is the most frequent category in training data, one point in the feature space is get assigned to a that category. For calculating distance between the vectors Euclidean Distance is used.

$$\mathrm{argmax}_i \sum_{j=1}^{k} \mathrm{sim}(D_j\,|D) * \delta(C(D_j), i)$$

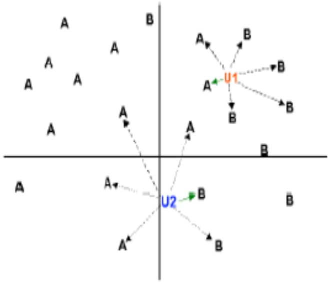

Figure 2: K-Nearest Neighbor

EVALUATION OF SENTIMENT CLASSIFICATION:

In general, the performance of sentiment classification is evaluated by using Accuracy, Precision, Recall and F1-score these four indexes. Confusion matrix is the way for computing these indexes is as shown below

TABLE 2: CONFUSION MATRIX:

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

Where-
TP - Document being classified correctly as relating to a topic.
FP - Document that is said to be related to the topic incorrectly.
TN - Documents that should not be marked as being in a particular topic and are not.
FN - Document that is not marked as related to a topic but should be.
These indexes can be stated in equations:

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

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Accuracy is the portion of all true predicted instances against all predicted instances. An accuracy of 100% means that the predicted instances are exactly the same as the actual instances. Precision is the portion of true positive predicted instances against all positive predicted instances. Recall is the portion of true positive predicted instances against all actual positive instances. F1 is an average of precision and recall.

CONCLUSION:

In this literature we have observed and underlined that sentiment analysis or opinion mining plays a vital role in decision making. Concepts of text mining but also the concepts of information retrieval are encompassed within opinion mining. This paper gives their contribution to the real-world application. We have proposed Web-based TSA to recognize the traffic related problems in a humanizer way. This paper presents the real-time web monitoring system for the detection of safety related patterns from web data. Mine the big unstructured datahas become an important research problem .Many of the techniques are fused together so as to optimally use advantages of all the techniques and also to collectively overcome disadvantages posed by others.

In future, more work is needed on further improving the performance measures which requires the further research. It is extended to Estimation of real-time traffic flows by fusing Twitter, Foursquare, and traditional traffic sensors data, such as GPS, loop detector, and camera data; Detection of traffic accidents and traffic congestions by fusing the precede heterogeneous data sources; and prediction of travel times or origination-destination times.

REFERENCES:

1. B. Pang and L. Lee, “Opinion mining and sentiment analysis,” Found. Trends Inf. Retrieval, vol. 2, no. 1/2, pp. 1–135, Jan. 2008.
2. Jianping Cao, Ke Zeng, Hui Wang, “Web-Based Traffic Sentiment Analysis: Methods and Applications”, IEEE Transactions on Intelligent Transportation Systems, Vol. 15, No. 2, April 2014.
3. F. Y. Wang, “Social computing: Concepts, contents, and methods,” Int. J. Intell. Control Syst., vol. 9, no. 2, pp. 91–96, 2004.
4. B. B. Khairullah Khan, Aurangzeb Khan, “Sentence based sentiment classification from online customer reviews,” ACM, 2010.
5. B. Pang and L. Lee, “A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts,” Proceedings of the42nd Annual Meeting on Association for Computational Linguistics, ACL, 2004.
6. Diana Maynard, Adam Funk,” Automatic detection of political opinions in tweets.”In: Proceedings of the 8th international conference on the semantic web, ESWC’11; p. 88-99, 2011.
7. Mining Hu and Bing Liu, “Mining and Summarizing Customer Reviews”, Proceedings of the tenth ACM SIGKDD International conference on knowledge discovery in data mining (KDD-2004), August 22-25.
8. Read J, carol J.,” Weakly supervised techniques for domain independent sentiment classification”, In: Proceeding of the 1st international CIKM workshop on topic-sentiment analysis for mass opinion; P.45-52, 2009.
9. P. D. Turney, “Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews,” in Proc. 40th Annu. Meet. Assoc. Comput. Linguist., 2002, pp. 417–424.
10. Kang Hanhoon, YooSeongJoon, Han Donglil,”Senti-lexicon and improved Naive Bayes algorithms for sentiment analysis of restaurant reviews” Expert SystAppl ,39:6000-10,2012.
11. Chin Chen Chien, Tseng You-De,”Quality evaluation of product reviews using an information quality framework”,Decis Support Syst;50:755-68,2011.
12. Li Yung-Ming, Li Tsung-Ying,”Deriving market intelligence from microblogs”,Decis Support Syst,2013.
13. Nigam K., Lafferty J., and McCallum A, (1990), “using maximum entropy for Text Classification”. In Proc of the IJCAI-99 Workshop on Machine Learning for Information Filtering.
14. Berger A., A Brief Maximum Entropy Tutorial.
15. Gongde Guo, Hui Wang, David Bell, Yaxin Bi and Kieran Greer, “KNN Model-Based Approach in Classification”, Proc. ODBASE pp- 986 – 996, 2003.
16. Ms Kranti Ghag and Dr. Ketan Shah,Comparative Analysis of the Techniques for Sentiment Analysis, ICATE 2013 Paper Identification Number-124
17. M. Annett, G. Kondrak, “Acomparison of sentiment analysis techniques: Polarizing movie Blogs”, In Canadian Conference on AI, pp. 25-35, 2008.