Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 3, March 2016, pg.512 – 516

OPINION MINING ON TRAFFIC DATASET USING RULE BASED APPROACH

Gayathiri.R, Arunkumar.A
Department of Computer Science and Engineering, India
14mi003@skcet.ac.in, arunkumara@skcet.ac.in

ABSTRACT:

Due to the development of various social networking sites, data on the internet explored more in recent years. Now days People started to post their views, ideas on social networks. So it is very difficult to analyze those big data manually or using traditional database systems. In order to handle those data, Sentiment Analysis is used. Sentiment Analysis is one kind of approach which is used to analyze the meaning of given contents and to find the polarity based on their structure of public opinion. Sentiment Analysis concentrates on many fields, one such field is Intelligent Transportation System. People share their opinion in internet about the government policies on Transportation system. These opinions can be analyzed using Sentiment Analysis approaches and send the results to the traffic services. This paper analyzes the people opinion using Rule Based approach and identifies positive and negative reviews. Finally, the accuracy of proposed system is calculated using the measures such as Precision and Recall and the results are obtained.

Keyword: Big Data, Sentiment Analysis, Rule Based Approach, Intelligent Transportation System, Accuracy.

I. INTRODUCTION

The rapid growth of internet users and the time users spend on the internet generates more data's in various formats such as audio, video, animations, text, etc which explored into Big Data [1]. The term 'Big Data' itself explains that it is a 'massive volume of data' which is very difficult to analyze both structured and unstructured data using traditional system so it requires the real time techniques to handle those data [2]. Now a Days People used to post their views, opinions, ideas about product, politics, marketing, company strategies on social networking sites. So People opinions can be analyzed from these sites and profitable information can be gathered, conclusions can be suggested and the decisions can be made. People opinions about particular topic can be judged using Sentiment Analysis [3]. Sentiment Analysis or sometimes called as Opinion Mining which targets to understand users mindset about particular topic by evaluating, scrutinizing and extracting subjective texts and discovers whether people have positive, negative or neutral opinion on the topic. The decisions can be made about the topic based on the opinions of the people [4]. Sentiment Analysis concentrates on many fields. Intelligent Transportation System is one of the sub-field where people opinions to be considered in order to get useful information and to take decisions. The Intelligent Transport System used to bring new policies, rules that citizens should follow. So People used to share their views, ideas regarding those rules or policies in social Networking Sites. Those feedbacks can be investigated, evaluated and predicted with the help of Sentiment Analysis.

The main objective of this paper is to analyze people feedback, comments about rules and policies of Intelligent Transportation system in order to find whether people satisfies about the rules and then reports the result to public traffic service section. The opinions are analyzed using Rule Based Approach. Accuracy of Rule Based Approach is evaluated using the measure Precision and Recall.

II. LITERATURE SURVEY

The section demonstrates the various machine learning algorithm used for Sentiment Analysis (SA) in recent years:

A. Machine Learning Approach:

Machine learning (ML) approach depends on the Machine Learning algorithms to solve the opinion mining as a regular text classification problem that makes use of syntactic or linguistic features. If suppose we have a set of training records

$$D = \{p1, p2, \ldots\ldots pn\}$$

Where each record gets labelled to a class. The classification model is related to the distinctive attribute in the underlying record to one of the class labels the model is used to predict a class label for a given instance of unknown class. When only one label is assigned to an instance then it is a hard classification problem. When a probabilistic value of labels is assigned to an instance then it is a soft classification problem.

B. Supervised Learning:

The supervised learning methods rely on the existence of labelled training documents. There are many kinds of supervised classifiers. Following are the brief details of some of the most frequently used classifiers in SA [5].

C. Decision Tree Classifier:

Decision tree classifier breaks the training data space into smaller parts in a hierarchical order at which the condition on the attribute value is used to divide the data. The condition or predicate is the existence or non-existence of one or more words. The data space is divided repeatedly until the leaf nodes contain certain least numbers of records which are used for the purpose of classification. There are other kinds of predicates which depend on the resemblance of documents to correlate sets of terms which may be used to further decomposition of documents. One of the different types of splits are Single Attribute split which use the existence or non-existence of particular words or phrases at a particular node in the tree in order to perform the split [6].

D. Linear Classifier:

Given Y={Y1,Y2,…..Yn} is the normalized document word frequency, vector B= {b1,……bn} is a vector of linear coefficients with the same dimensionality as the feature space, and b is a scalar; the output of the linear predictor is defined as p= B . Y + a, which is the output of the linear classifier. The predictor pr is a separating hyperplane between dissimilar classes. There are many kinds of linear classifiers; one among them is Support Vector Machines (SVM) which is a form of classifiers that try to obtain good linear separators between different classes. Two of the most famous linear classifiers are as follows [5].

1. Support Vector Machines Classifiers (SVM):

The SVMs is used to determine linear separators in the search space which can separate the different classes in best way. There are 2 classes named as x, o and there are 3 hyper planes named as A, B and C. The normal distance of any of the data points is the largest, so Hyperplane A provides the best separation between the classes because it represents the maximum margin of separation. SVM classification is well suited for text data classification because of its sparse nature, in which few features are immaterial, but they tend to be connected with one another and generally ordered into linearly separable categories. SVM can build a nonlinear decision surface in the original feature space by mapping the data instances non-sequentially to an inner product space where the classes can be separated linearly with a hyperplane [7].

E. Rule-Based Classifier:

Rule based classifier use set of rules to model the data space. The left hand side represents a constraints on the feature set expressed in disjunctive normal form and the right hand side represents the class label. The constraints are based on the term existence. If the term does not get exists then the term absence is rarely used because it is not provide useful information in sparse data. There are numbers of principles to generate rules, the training phase construct all the rules depending on these principles. The most occurring two principles are Confidence and Support [8].

F. Probabilistic Classifiers:

Probabilistic classifiers perform classification by using the mixture model. The mixture model pretends that each class is an element of the mixture. Each mixture element is a generative model that provides the probability of sampling a particular term for that element. These kinds of classifiers are also called generative classifiers. There are three types of most famous probabilistic classifiers are as follows:

1. Naive Bayes Classifier:

Bayesian classifiers are based upon the Bayes rule, a way of viewing conditional probabilities that agrees to flip the condition around in a suitable way. A conditional probability is that probably event E will occur, if the evidence EV is given. That is normally written as P(E | EV). The Bayes rule helps us to find this probability when the probability of the opposite result is known [9]:

$$P (E \mid EV) = P (E) P (EV \mid E) /P (EV)$$

2. *Maximum Entropy Classifier (ME):*

The Maximum Entropy Classifier also known as a conditional exponential classifier which converts labelled attribute sets into vectors using encoding. This encoded vector is then utilized to calculate weights for each attribute that can then be merged to determine the most likely label for attribute set. This classifier is parameterized by set of X {weights}, which is used to merge the joint attributes that are created from an attribute-set by an X {encoding}. In particular, the encoding maps each C {(feature set, label)} pair to a vector [10].

3. *Weakly, semi and unsupervised learning:*

The main intention of text classification is to categorize documents into a certain number of predefined divisions. Large number of labelled training documents is used for supervised learning for text classification. Sometimes it is difficult to create these labelled training documents, but it is not difficult to collect the unlabelled documents. The unsupervised learning methods overcome these difficulties which divides the documents into sentences, and each sentence are categorized using keyword lists of each category and measures sentence similarity [11].

III. PROPOSED SYSTEM

The proposed system is used to classify the polarity of public opinion as positive, negative and neutral. The proposed system collects the dataset from the various social media websites. Then the dataset is pre-processed and the sentiment word, negative word and degree word. Then the each opinion will be classified as sentence and the polarity of each sentence is calculated and weight age will be given to each sentence. The polarity and weight age of each sentence is aggregated to each opinion and polarity of sentence is derived. Then the performance of the proposed algorithm is evaluated.

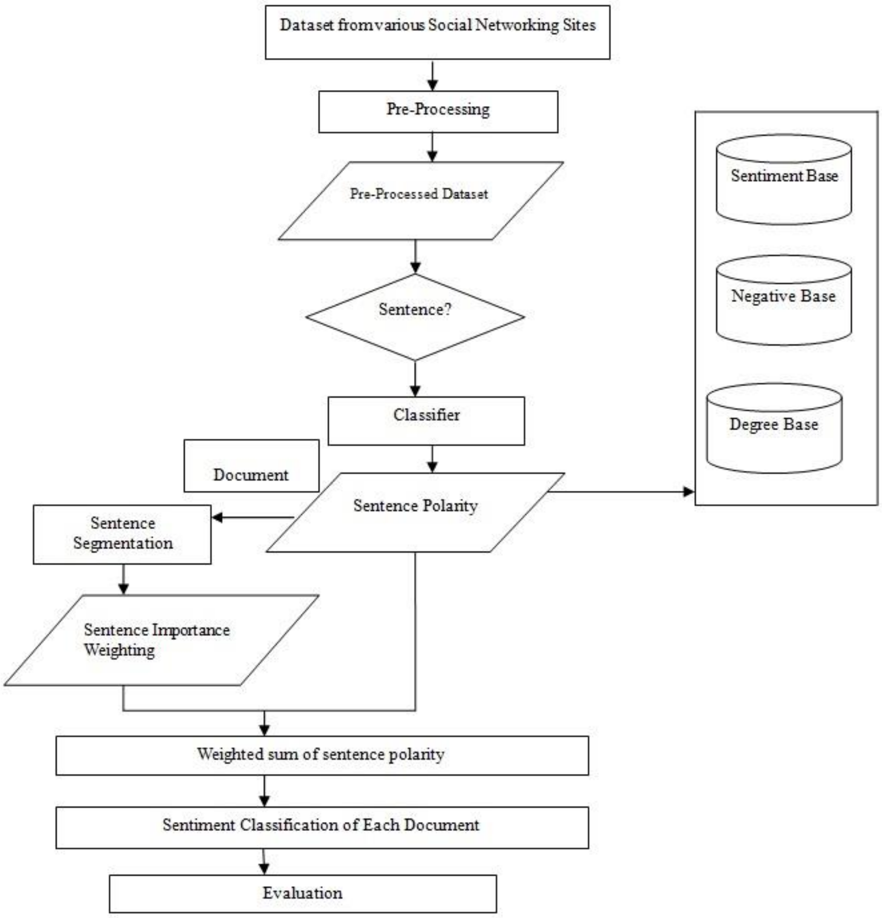

FIG1. ARCHITECTURE DIAGRAM

IV. MODULES

A. Module 1: Pre- Processing

Data is collected from various social networking sites such as facebook, youtube, news article sites, etc regarding people opinion on rule imposed by government on Intelligent Transportation System. The people opinions are then pre-processed to implement the sentiment analysis technique. The unwanted noises are removed from the database that is removing the stop words and the text between the open and close braces and also the special characters are removed from the text except full stop which is used to segment the sentence. Sentiment word base, Negative word base and Degree word base are collected from English club dictionary and the database are created to implement rule based approach and from the dataset collected by using Morpheme. Emotional verbs and emotional negative, degree words that contain polarity are collected from English club dictionary and data base are created.

B. Module 2: Implementing Rule Base Sentiment Analysis

Each comment is taken and checks whether it is a sentence or document. If it is a document, document will splits into sentences. By using the sentiment word base, Negative word base and Degree Word base the polarity of each sentence is calculated based on the rule base. The semantic rule of sentiment is the pattern of the sentiment words (S) and their modifiers negative words (N) and degree words (D), which is expressed by the pattern SND. Among the three factors, S is considered as the most important. Therefore, we first select S from the sentence. The corresponding N and D are placed around S. The SND model is then established. In the N+S rule, negative words are equivalent to a non-negative word and odd negative words are equivalent to one negative word. In the N+D+S rule, N is the modifier of D, and N+D is the modifier of the sentiment word (S). Therefore, the characteristics of N+D+S are the same as those of D+S. However, in the D+N+S rule, the negative word (N) is the modifier of the sentiment word (S), and the degree word (D) is the modifier of N+S. S usually represents a verb or a noun. Features such as position of the sentence, term weight, similarity between sentence and the headline, occurrence of keyword in the sentence and the first-person mode are extracted from each sentence of the document. Based on these features importance of sentences are weighted. Based on the polarity of each sentence and weight age of important sentence, polarity of whole will be calculated and based on this document polarity comments will be classified as positive, negative and neutral.

C. Module 3: Performance Evaluation

The True Positive, True Negative, False Positive and False Negative are calculated for both the approaches such as rule based approach. Precision calculated the percentage of positive results returned that are relevant. Recall is the fraction of relevant instances that are retrieved. Precision, recall and f-measure is evaluated by using the True positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Then the Precision, Recall and f-measure of both the approaches are compared and the best result is obtained.

1. Accuracy

Accuracy can be calculated from formula given as follows:

Accuracy = (True Positive + True Negative)/(True Positive + True Negative + False Positive + False Negative)

2. Precision

Precision value is the fraction of retrieved instances that are relevant. Precision is calculated the percentage of positive results returned that are relevant.

Precision = True Positive / (True Positive + False Negative)

3. Recall

Recall value is calculated is the fraction of relevant instances that are retrieved. Precision is calculated the percentage of positive results returned that are Recall in this context is also referred to as the True Positive Rate. Recall is the fraction of relevant instances that are retrieved,

Recall = False Negative / (True Positive + False Negative)

V. CONCLUSION

We have proposed sentiment analysis on people opinion on traffic rules. The rule based Sentiment Analysis is used to evaluate the people opinion which considers only the positive and negative reviews in which its precision value is high. The future work concentrates on considering the three reviews such as positive, negative and neutral review. The proposed system also improves the recall accuracy which is calculated using true positive, true negative, false positive, false negative.

REFERENCES

1) Xiaoxue Zhang; Feng Xu, "Survey of Research on Big Data Storage," in Distributed Computing and Applications to Business, Engineering & Science (DCABES), 2013 12th International Symposium on , vol., no., pp.76-80, 2-4 Sept.2013.
2) B. Pang and L. Lee, "Opinion mining and sentiment analysis," Found. Trends Inf.Retrieval, vol. 2, no. 1/2, pp. 1–135, Jan. 2008.
3) Han Hu; Yonggang Wen, Tat-Seng Chua and Xuelong Li, "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial," in Access, IEEE, vol.2, no.,pp.652-687, 2014.
4) Zhu Nanli, Zou Ping, Li Weiguo and Cheng Meng, "Sentiment analysis: A literature review," in Management of Technology (ISMOT), 2012 International Symposium on , vol., no., pp.572-576, 8-9 Nov. 2012.
5) Aggarwal Charu C, Zhai Cheng Xiang. Mining Text Data. Springer New York Dordrecht Heidelberg London: Springer Science+Business Media, LLC'12; 2012.
6) Quinlan JR. Induction of decision trees. Machine Learn 1986; 1:81–106.
7) Ng Hwee Tou, Goh Wei, Low Kok,"Feature selection, perceptron learning, and a usability case study for text categorization" In: Presented at the ACM SIGIR conference; 1997.
8) Liu Bing, Hsu Wynne, Ma Yiming, "Integrating classification and association rule mining". Presented at the ACM KDD conference; 1998.
9) Troussas, C.; Virvou, M.; Junshean Espinosa, K.; Llaguno, K.; Caro, J., "Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning," in Information, Intelligence, Systems and Applications (IISA), 2013 Fourth International Conference on , vol., no., pp.1-6, 10-12 July 2013.
10) Kaufmann JM. JMaxAlign,"A Maximum Entropy Parallel Sentence Alignment Tool". In: Proceedings of COLING'12: Demonstration Papers, Mumbai; 2012. p.277–88.
11) Ko Young joong, Seo Jungyun,"Automatic text categorization by unsupervised learning".Proceedings of COLING-00, the 18th international conference on computation al linguistics; 2000.