A Literature Survey on Web-Based Traffic Sentiment Analysis: Methods and Applications

Anju Murali J, Varghese S Chooralil

Abstract— In the advancement of modern intelligent transportation system (ITS), the need of public opinions is very important. For processing traffic information from websites, we propose Traffic Sentiment Analysis (TSA). TSA is a subfield of Sentiment Analysis which gives focus to the issues of the traffic. Here in this paper, we use rule-based approach rather than learning-based approach. TSA plays an important role in sensing, computing and supporting the decision making in ITS. Rule-based approach is used for dealing with real problems, presented an architectural design, constructed related bases, demonstrated the process. It is the first attempt to apply sentiment analysis on the area of traffic. The functions of TSA system are Investigation, Evaluation and Prediction.

Index Terms— Traffic sentiment analysis, machine learning techniques, Rule base

1 Introduction

Sentiment analysis is a machine learning approach in which machines analyze and classify the human's sentiments, emotions, opinions, etc. about topics which are expressed in the form of either text or speech. Modern intelligent transportation systems (ITSs) failed to concern about the public opinions. For the completeness of ITS space, it is necessary to collect and analyze the public wisdom and opinions. TSA treats the traffic problems in a new angle, and it supplements the capabilities of current ITS systems.

2 Literature Review

The growth of the internet is exponentially and faster so that web-based information has drawn much significance. Enormous amounts of information are available in online documents. Sentiment analysis is used to find sentiments of different context in the text document.

Pang, Lee and Vaithyanathan [1] examine the effectiveness of applying machine learning techniques to the sentiment classification problem.

Naive Bayes:

Assign a given document d to the class

$$\text{class } c^* = \arg\max_c P(c \mid d)$$

Naive Bayes (NB) classifier is derived by first observing that by Bayes rule,

$$P(c \mid d) = (P(c)\, P(d \mid c)) / P(d)$$

To estimate the term P (d | c), Naive Bayes decomposes it by assuming the f_is are conditionally independent given ds class. Naive Bayes is best for certain problem classes with highly dependent features.

Maximum Entropy:

Maximum entropy classification (MaxEnt or ME) is an alternative technique which outperforms Naive Bayes at standard text classification. MaxEnt make no assumptions about the relationships between features and when the conditional independence assumptions are not met, it might perform better. The underlying philosophy is that we should choose the model making the fewest assumptions about the data which is consistent with it, by intuitive sense.

Estimation of P(c | d) takes the following exponential form:

$$P_{\mathrm{ME}}(c \mid d) := \frac{1}{Z(d)} \exp\left(\sum_i \lambda_{i,c} F_{i,c}(d,c)\right)$$

Where Z (d) - normalization function.

$F_{i,c}$ is a feature/class function for feature f_i and class c, defined as follows:

Support Vector Machine (SVM):

Support Vector Machine is non-probabilistic classifiers and large-margin unlike, Naive Bayes and Maximum Entropy. The basic principle behind the training procedure is to find a hyper plane, represented by vector w that separates the document vectors in from one class with those in the other. Let Cj ∈ {1,-1} be the correct class of document d_i

$$\vec{w} := \sum_j \alpha_j c_j \vec{d}_j, \quad \alpha_j \geq 0$$

Where α_is are obtained by solving a dual optimization prob-

- *Anju Murali J is currently pursuing Master of Technology degree program in Computer Science & Engineering in M.G. University, Rajagiri School of Engineering & Technology, Kerala, India. E-mail: anjumurali.j@gmail.com*
- *Varghese S Chooralil is currently working as Assistant Professor, Department of Computer Science and Engineering, M.G. University, Rajagiri School of Engineering & Technology, Kerala, India. E-mail: varghesesc@rajagiritech.ac.in*

lem and $\vec{d}_i$ such that α_i is greater than zero are called support vectors, since they are the only document vectors contributing to w. Classification of test instances consists simply of determining which side of w's hyperplane they fall on.

The machine learning techniques produced good results when compared to human-generated ones. The Naive Bayes tends to do the worst and SVMs tend to do the best, in terms of relative performance. The topics are mainly identifiable by keywords; so that sentiment can be expressed in a more subtle manner is a challenging task.

Turney [2] presents an unsupervised learning algorithm for classifying a review as recommended or not recommended. Pointwise Mutual Information and Information Retrieval (PMI-IR) algorithm is used to measure the similarity of pairs of words or phrases and to estimate the semantic orientation of a phrase.
Steps:
Input: written review
Output: classification

1. POS tagger to identify phrases in the input text
2. Estimating the semantic orientation of each extracted phrase
3. Assigning the given review to a class, recommended or not recommended, based on the average semantic orientation of the phrases extracted from the review.

The Pointwise Mutual Information (PMI) between two words, word1 and word2, is defined as follows

PMI (word1, word2) = log2 [p(word1 & word2) / p(word1) p(word2)]

Applications of Thumbs up or Thumbs down (Turney, 2002) are
(i) Providing summary statistics for search engines. Given the query "Akumal travel review", a search engine could report, "There are 5,000 hits, of which 80% are thumbs up and 20% are thumbs down."
(ii)Filtering flames for newsgroups

The advantages are simple and movie reviews are difficult to classify, because the whole is not necessarily the sum of the parts. It is more preferred for banks and automobiles over movie domain because the whole is the sum of the parts. The drawback includes the time required for queries and level of accuracy.

Lun-Wei Ku [3] proposes algorithms for opinion extraction, summarization, and tracking. Opinion extraction, opinion summarization and opinion tracking are three important techniques for understanding opinions. Watching specific information sources and summarizing the newly discovered opinions are important for governments to improve their services and for companies to improve their productsAn opinion tracking system provides not only text-based and graph-based opinion summaries, but also the trend of opinions from many information sources. Opinion summaries show the reasons for different stands people take on public issues.

A.-M. Popescu [6] introduces OPINE, an unsupervised informationextraction system which mines reviews in order to build a model of important product features, their evaluation by reviewers, and their relative quality across products. OPINE's novel use of relaxation labeling for finding the semantic orientation of words in context leads to strong performance on the tasks of finding opinion phrases and their polarity.

Cao and Zeng [4] propose a traffic sentiment analysis (TSA) as a new tool to provides a new prospective for modern intelligent transportation systems (ITSs). In the advancement of ITS, need of public opinions is very important. TSA is a subfield of SA, which gives focus to the issues of the traffic and can be used for processing traffic information from websites. TSA plays an important role in sensing, computing and supporting the decision making in ITS and it is the first attempt to apply SA in the area of traffic. The functions of TSA system are as follows.
1. Investigation: It is the most effective method to collect public opinion than public poll.
2. Evaluation: It is used to evaluate the performance of traffic services and policies.
3. Prediction: TSA system can be used to predict the trends of some social events. For example, in 2009, the volcano ash from Iceland caused the malfunction of traditional social sensors such as cameras. In such emergency situations, TSA is independent of current systems and can detect it in a new humanized perspective.These can be shown in Fig.1

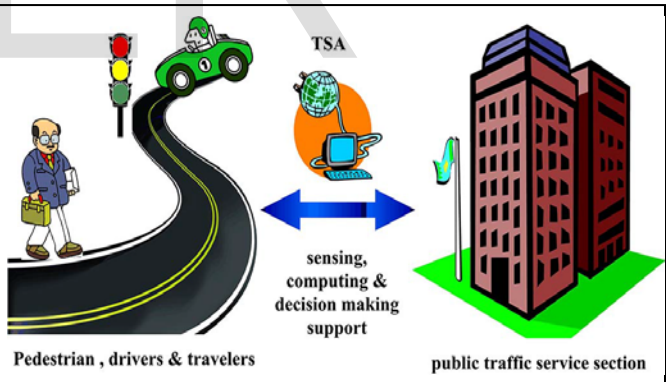

Fig. 1. TSA system plays the role of sensing, computing, and supporting the decision making in ITS.

The primary issue of TSA is the selection of sentiment analysis approaches for Web-based data. The Properties of web data can be described as follows.
(i) Lengths of the texts vary: Some texts contain thousands of words, whereas others consist of only one sentence that can be as short as a one word.
(ii) Stylistic features of texts are diverse: Different users express in different ways, hence the features of expressions on the Web vary.
(iii) New Internet expressions frequently emerge: Same sen-

timent may be expressed in different ways. In extreme cases, the same word may carry a different sentiment polarity in each context.

The Rule-based approach is applied over learning-based approaches.

Learning-based Approach:

The advantage of the learning-based approach is that it does not need expert knowledge to build the related bases; instead classifier is simply trained without considering the context. The texts should first be categorized based upon their sizes. The document and sentence-level clauses should then be trained separately. The disadvantage of learning-based approach is that large training data sets with positive and negative examples are required, which is expensive and time consuming. Moreover, classification standard of different levels of clauses must be carefully learned and investigated.

Rule-based approach:

The disadvantage of the approach is that if the context of the texts is not considered, the sentiment polarity results cannot be as precise as expected. The advantage of rule-based approach is that the precision of the rule-based approach is independent of the sizes of the clauses. Despite the differences in the stylistic features of various users, the syntax rule of a certain language is static and basic. The word choice and thought process basically remain unchanged. Therefore, the rules of the rule-based approach are relatively static. Finally, it can be easily extended by simply updating the sentiment lexicon, although new sentimental words rapidly emerge and several words' sentiments may be changed with words.
The architecture as shown in fig.2 is based on the tackling process and its main components, including

1) Web data collection
2) Pre-processing
3) Extraction of subjects and objects
4) Extraction of sentiment properties
5) Sentiment calculation and classification
6) Evaluation or applications
7) Feedback.

Extraction of properties: The extraction of properties is based on the sentiment, modifier, and rule bases.
Evaluation: The efficiency and precision of the algorithms are tested in this step. If the test performance is lower than expectation, proper words will be identified and updated to improve the related bases.
The basic step in the construction of sentiment word base is the calculation of sentiment polarity of words. The text processing includes the polarity calculation of sentence level text and document level text. Fig.3 shows the overall process involved in the proposed approach.

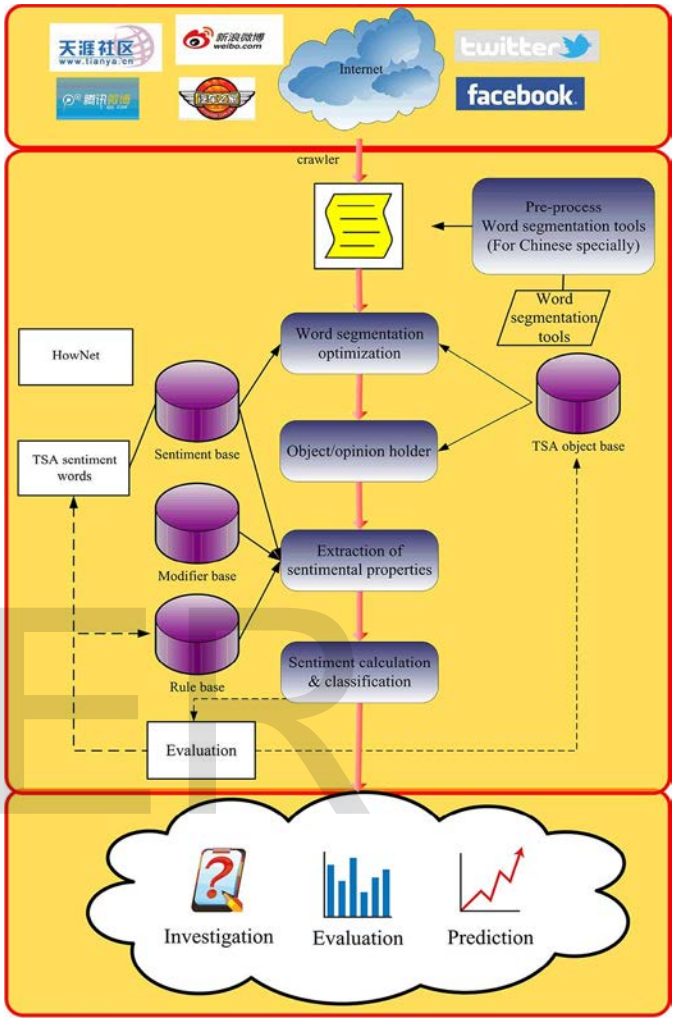

Fig. 2. Rule-based TSA Architecture

First, decompose a document into constituting sentences and determine the sentiment polarity of each sentence. The polarity scores of all the sentences are then synthesized to compute for the overall polarity of the entire document. The importance of a sentence to a document can be represented by the weight in the overall polarity computation. We formalized the problem as follows. Given t be a text which containing sentences $s_1,\ldots,s_n$ as inputs, the system must calculate p_s, the polarity score for each sentence s_i and determine the sentiment polarity, where w_i is the weight of sentence s_i. The document shows a positive sentiment if $P_t > 0$, otherwise, it shows a negative sentiment.

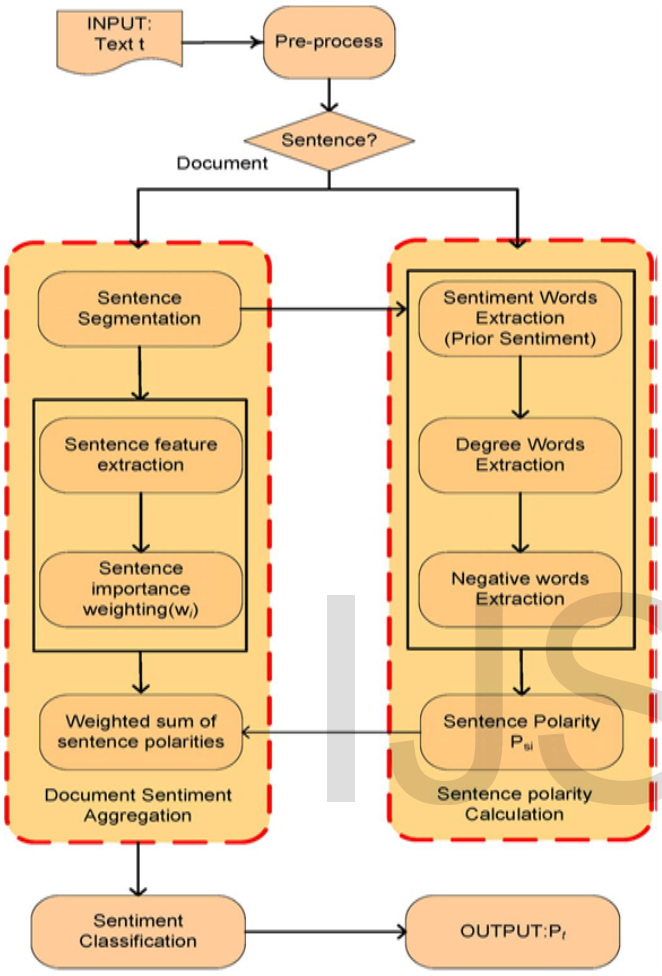

Fig. 3. Illustration of Rule-based Sentiment Analysis

Algorithm Rule-based Sentiment Analysis

1. Input: a set of texts, obtained a text t
2. do pre-process on t
3. If t is on sentence level
4. do SND extraction on t
5. calculate P_t of t
6. End
7. Else if t is a sentence
8. segment t into sentences $S_1,.....,S_n$
9. for each sentence S_i
10. do 4-6 get P_{s_i}
11. do sentence feature extraction
12. get sentence S_i weighting W_i , $\sum_{i=1}^{n} w_i = 1$
13. end for
14. get P_t by $P_t = \sum_{i=1}^{n} P_{s_i} w_i$
15. End if
16. do sentiment classification on P_t
17. Output : sentiment polarity of text t

Fig. 4. Algorithm for rule-based sentiment analysis

The Traffic regarding information classified into three:

1) News, Expert commentaries, announcements
2) Post from transport sector in forum
3) Real time traffic information on social media

Fig.4 shows the algorithm for rule-based sentiment analysis which takes text as input and gives sentiment polarity value as output. Fig.5 shows the evaluation result for various machine Learning Techniques accuracy rate in percentage. Here we can consider only the Naive Bayes (NB), Maximum Entropy (MAXENT), Support Vector Machine (SVM), KU, Rule-Based Sentiment Analysis (R-BSA) protocols for the evaluation.

S.-M. Kim [5] present a system that, given a topic, it finds the people who hold opinions about that topic and the sentiment of each opinion. The system consists of a module for determining word sentiment and another for combining sentiments within each sentence by using various models of classifying and combining sentiment at word and sentence levels such as Word Sentiment Classifier Sentence Sentiment Classifier. Model 0, 1 and 2 are the reffered models for classification and combination of sentiments.

Sl.no	Protocol	Goals	Gains	Limitations
1.	NB SVM MAXENT	Find the three cross fold validation	Best performance in accuracies	Unreliable for low frequency feature
2.	ULT PMI-IR	Determine the semantic orientation of adjectives	Domain specific	For large dataset, distant search engine for rare words
3.	LCT	Opinion mining	Automatically identify the product feature	Strength of opinions and accuracy rate
4.	KU'S R-BSA	Addressed the key problems of TSA, including the design of architecture and the construction of related bases.	R-BSA provided the higher accuracy rate.	Difficult to classify the low importance sentences

Fig. 5. Performance Comparison

4 Conclusion

Web-based TSA is proposed to analyze the traffic problems in a most humanizer way. It is the first attempt to apply sentiment analysis on the area of traffic. The study of TSA will provide us a new perspective when facing with traffic problems. The main contributions of this paper designing the application architecture of TSA, Constructing the related bases for the TSA system, Comparing the advantages and disadvantages of both rule- and learning-based approaches based on the characters of web data, proposing an algorithm for the sentiment polarity calculation based on the rule-based approach and consideration of the modifying relationships of sentence patterns and locations in the sentiment polarity calculations.

The future work will focus on the prediction by incorporating web-based TSA. The task to implement the TSA system into existing ITSs is also critically important. To support the decision making of managers, take the policy evaluation part and view the evaluation results related to specific location as sensor information. More techniques will be developed for the joint performance of ITS with the TSA system in the future.

References

[1] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classi_cation using machine learning techniques," in Proc. ACL Conf. Empirical Methods Natural Lang. Process., vol. 10 , pp. 79 86, 2002.

[2] P. D. Turney,"Thumbsup or thumbsdown?: Semantic orientation applied to unsupervised classication of reviews," in Proc. 40th Annu. Meet. Assoc. Comput. Linguist., pp. 417 424 , 2002.

[3] L.-W. Ku, Y.-T. Liang, and H.-H. Chen, "Opinion extraction, summarization and tracking in news and blog corpora," in Proc. AAAI Spring Symp., Comput. Approaches Anal. Weblogs, pp. 100 107, 2006.

[4] J. Cao, K. Zeng, H. Wang, F. Qiao, D. Wen, Y . Gao and J. Cheng, "Web-based trac sentiment analysis: methods and applications," IEEE Intell. Trans.Syst., vol. 15, no. 2, Apr. 2014.

[5] S.-M. Kim and E. Hovy, "Determining the sentiment of opinions," in Proc. 20th Int. Conf. Comput. Linguist., 2004, pp. 1367–1373.

[6] A.-M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," in Natural Language Processing and Text Mining. New York, NY, USA: Springer-Verlag, 2007, pp. 9–28.