International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 05 | May -2017 www.irjet.net p-ISSN: 2395-0072

SENTIMENT ANALYSIS BY USING UNSUPERVISED COMMENT SUMMARIZATION

Anjali Rai, Kuldeep Jaiswal

M.Tech Student,Dept. of Computer Science and Engineering, BBD University, Lucknow, India

Assistant Professor,Dept. of Computer Science and Engineering, BBD University, Lucknow, India

---

Abstract : *Sentiment analysis is used to determine the attitude of a writer or a speaker with respect to some topic. In this paper we focus on the classification of features and their performance. We studies movie reviews using different sentiment analysis methods. We compared Decision Tree, Support Vector Machine, Naive Bayes (Kernel) and Stacking for the sentiment classification of movie reviews. And we evaluate the performance vector like accuracy, root mean squared error and absolute error for all the methods. Result analysis shows that Stacking gives better accuracy than all the other methods.*

Keyword: **Sentiment analysis, Decision tree, Support vector machine, Navie bayes kernel, Stacking.**

1.INTRODUCTION

Opinions are key influencers of human behaviours. We want to know others' opinions to make any decision. In the real world, business organizations consider consumers' opinion about their product and services which helps them to improve the quality of the product. It is also beneficiary for consumers as they can consider others opinions and experiences while taking decision about products and services. E-commerce, social media, forums, blogs etc help people to acquire these opinions as the amount of user generated data on internet increases day by day in the form of reviews, opinions and comments. However, finding and tracking opinion sites on the internet and filtering the data contained in them remains a intimidating task owing to the expansion of numerous sites. Each web site generally contains a huge volume of opinion text that's not always simply discerned in long blogs and forum postings. The average human reader will have issues distinguishing relevant sites and extracting and summarizing the opinions in them. Therefore automated sentiment analysis systems are required. In recent years, we have witnessed that opinionated postings in social media helped in reshaping businesses, and affect public sentiments and emotions. It has thus become a necessity to gather and study opinions on the online. Of course, opinionated documents not merely exist on the internet (called external data), many organizations additionally have their internal data, e.g., customer feedback collected from emails and call centres or results from surveys conducted by the organizations. Due to these applications, industrial activities increased in recent years. Sentiment analysis applications have spread to nearly every potential domain, from consumer merchandise, services, healthcare, and financial services to social events and political elections.

Sentiment analysis is a natural language processing technique that automatically extracts opinions, views, sentiments, emotions etc and classify them into different categories like positive, negative and neutral. In general, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or a document. Sentiment analysis is different from text classification because in text classification we have many classes corresponding to different topics where as in sentiment analysis we have only three classes i.e. positive, negative and neutral. The objective of sentiment analysis is to analyze the data and extract the opinionated phrases as features and use it for sentiment analysis, after this we perform classification.

Classification of data is done at three different levels like aspect level, sentence level and document level. In the case of document level sentiment analysis, we classify the whole document as positive or negative. The sentence level sentiment analysis is related to subjectivity analysis. At this level each sentence is analyzed and its opinion is determined as positive, negative or neutral. The aspect level sentiment analysis aims at identifying the target of the opinion.

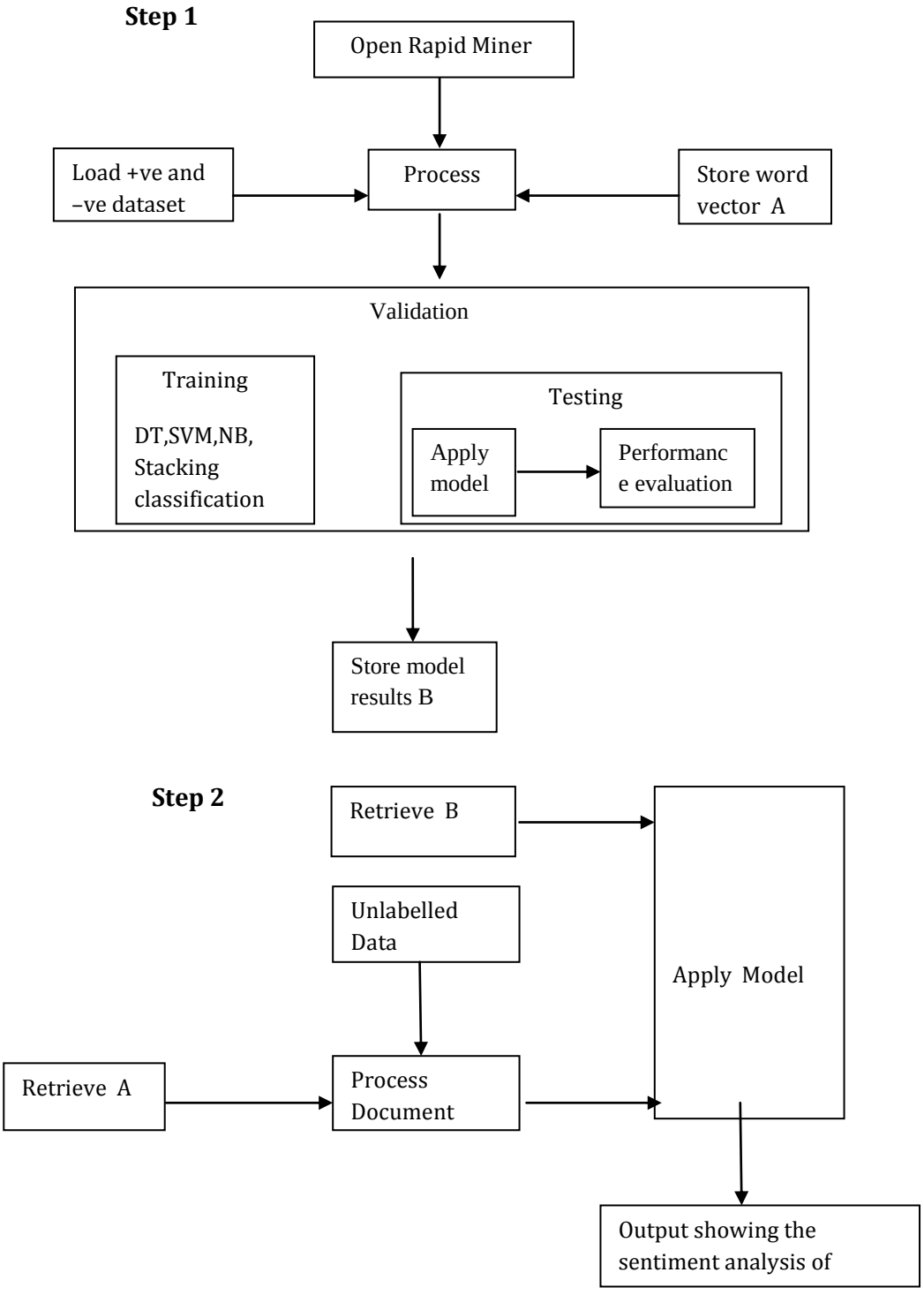
In the represented work an attempt has been made to develop various machine learning classification approaches with different algorithms to obtain sentiment analysis models for the movie review dataset and to compare them.

Handling of this user generated data on web is difficult because of its huge size and we have to process this data on daily basis. We face many challenges while processing this data like implicit sentiments and sarcasm, domain dependency, thwarted expectations, pragmatics, world knowledge, subjectivity detection, entity identification and negation. Complexity of these

challenges varies from high to low. Some of these problems are easily solvable like world knowledge and some are difficult like sarcasm and negation.

2.PROPOSED TECHNIQUE

This section presents the proposed technique to analyze sentiments in a movie domain. The proposed approach uses combinations of natural language processing techniques and supervised learning algorithms. The model for the proposed technique is depicted in figure



3.DATA USED

The proposed work is evaluated by running experiments with the polarity dataset V2.0, available at http://www.cs.cornell.edu/people/pabo/movie-review-data. Sentiment model has been built using supervised learning technique. For this, a set of 200 movie review data available from Pang and Lee at Cornell university has been used. It has 200 positive reviews, 200 negative reviews and 200 unlabelled reviews for testing of the model.

4.MODELS AND METHODOLOGY

We used three models for sentiment classification of the movie review data set. These models are Decision Tree (DT), Support Vector Machine (SVM), Naive Bayes kernel (NB) and Stacking model using DT, SVM and NB. Using Rapid Miner tool for model development Stacking Meta learning algorithms have been used along with other algorithms for enhancing the performance of the model as well as using different combinations of the model.

4.1 Decision Tree

Decision tree classifier used in decision support system and machine learning processes. A decision tree is a predictive modelling technique that used in classification, clustering and predictive task. Decision tree classifier provides a hierarchical decomposition of the training data space in which a condition on the attribute value is used to divide the data into collection of simple decisions, thus providing a solution which is easy to interpret. This classifier able to handle a variety of input data like nominal, numeric and textual. It can also process datasets that have errors and missing values and gives high predictive performance for relatively small computational effort. We construct decision tree in three different phases, construction, pruning and processing. In the first phase we build the tree by using entire training data set. It requires recursively partitioning of the training set into two or more, sub-partition using a splitting criterion, until a stopping criteria is met. In pruning phase, the constructed tree may not result in the best possible set of rules due to over fitting. The pruning phase removes some lower branches and nodes to improve its performance. And at the end we process the pruned tree to improve the understanding.

4.2 Support Vector Machine

Support vector machine is a method for classification of data generally organized into linear separable categories. Non linear mapping is used to transform the training data set into a higher dimension. Support vector machine (SVM) classifier is used for classification and regression tasks. It minimizes the classification error and maximizes the geometric margin. So it is also called maximum margin classifiers. As it maps input vector into a higher dimensional space, there a maximal separating hyper plane is constructed. Two parallel hyper planes are constructed on each side of the hyper plane that separates the data. This separating hyper plane maximizes the distance between the two parallel hyper planes. Large distance between these two parallel hyper planes indicates the better the generalization error of the classifier.

4.3 Naive Bayes

Naive Bayes text classification is used as a probabilistic learning method. Naive Bayes (NB) classifier is the most common method for classifying text documents. It assumes that the probabilities being combined are independent of each other. The probability of one word in the document being in a specific category is unrelated to the probability of the other words being in that category. Calculating the entire document's probability is a matter of multiplying all the probabilities of the individual words in that document. Bayesian classifiers are often used for document classification because they require less computing power than other methods. The NB classifier is often used as a good baseline method, with results that are sufficiently good for practical use. Spam filtering is one of the common uses of Naive Bayes text classification.

4.4 Stacking

Stacking sometimes called stacked generalization involves the training of a learning algorithm to combine the predictions of several other learning algorithms. Firstly, all the other algorithms are trained using the available data and then a combiner algorithm is used to make a final prediction using all the predictions of all the other algorithms as additional inputs. This operator is used for combining many models of different types, thus introducing the meta learner concept. We first split the training data set into two disjoint subsets. Stacking improves the performance by combining all the models thus resulting in a better performance than any single trained models. It has been successfully used on both supervised learning tasks (regression, classification and distance learning) and unsupervised learning (density estimation). The stacking operator is a nested operator, having two sub processes, the base learner and the stacking model learner.

5.EXPERIMENTS

In the first step we perform three tasks, data input, text processing and validation of data. In data input, we load positive and negative labelled data of the movie reviews. Text processing consists several tasks like transform cases, tokenize, filter tokens, filter stop words, replace tokens, stems, generation of n grams and store operator. Validation of data consists two sub

processes, training and testing. The training sub process, we train all models. The trained model is then applied in the testing sub process. During the testing phase, the performance of the model is also measured. For training we use decision tree, support vector machine and naive bayes models. But in case of stacking, some base learners and stacking model learner are used for training. Testing is performed with the help of apply model and performance evaluation. Apply model applies the trained model on the data set and performance evaluation measures the different performance criteria values.

In the second step, apply model operator takes a model from the retrieve operator and unlabelled data as input and provides an output.

6.RESULT ANALYSIS

Performance of all the models is calculated on three performance criteria values, accuracy, root mean squared error (RMSE) and absolute error. Here the dataset consists of 200 reviews equally divided into 100 positive and 100 negative. The performance of the stacking model has a better accuracy (86.75%) than the rest of the models. Also the model performance has RMSE value of 0.361 and absolute error of 0.133, better than the rest of the models, which again clearly demonstrates the good predictive capability of the model.

Table 1 : Performance Vector of all the models

Algorithms Used	Accuracy	RMSE	Absolute Error
Decision Tree	72.50 %	0.470	0.313
Support Vector Machine	80.50 %	0.478	0.477
Naive Bayes (kernel)	78.00 %	0.440	0.429
Stacking	86.75 %	0.361	0.133

7.CONCLUSION

In the present work machine learning techniques are used to detect the sentiments of movie reviews. We consider movie review data of 200 reviews which is divided into 100, positive and 100, negative reviews. On the basis of these reviews we have to categorised the movies into good, bad or average. Multiple experiments were carried out using different feature sets and parameters to obtain maximum accuracy. The proposed work is evaluated by running experiments with the polarity dataset V2.0, available at http://www.cs.cornell.edu/people/pabo/movie-review-data. Out of the four models developed using Support Vector Machine (SVM), Decision Tree (DT), Naive Bayes kernel (NB) and Stacking Model, Stacking Meta Algorithm using SVM, DT and NB proved to be the best model to learn sentiments from a review, achieving considerable accuracy of 86.75% with very basic feature set. Overall, gaining further improvements using the supervised model of machine learning for the purpose of sentiment analysis is studied in Natural Language Processing and requires sufficient knowledge of Linguistics. More sentence structure analysis and understanding of words in different part of speech and sarcasm is required. Extracting features based on semantic structure of the text will improve the accuracy of these classifiers.

REFRENCES

[1] M.N. Moreno, et al., Web mining based framework for solving usual problems in recommender systems. A case study for movies' recommendation, Neurocomputing (2015), http://dx.doi.org/10.1016/j.neucom.2014.10.097i

[2] P. Turney, “Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews”, Proceedings of the Association for Computational Linguistics (ACL), 2002, pp. 417–424.

[3] B. Pang and L. Lee, “Opinion mining and sentiment analysis,” Foundations and Trends in Information Retrieval 2(1-2), 2008, pp. 1–135

[4] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews," Proceedings of WWW, 2003, pp. 519–528.

[5] A. Harb, M. Planti, G. Dray, M. Roche, Fran, o. Trousset and P. Poncelet, "Web opinion mining: how to extract opinions from blogs?", presented at the Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology, Cergy-Pontoise, France, 2008.

[6] W. Zhang, H. Xu, W. Wan, "Weakness Finder: Find product weakness from Chinese reviews by using aspects based sentiment analysis," Expert Systems with Applications, Elsevier, vol. 39, 2012,pp.10283-10291.

[7] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu, "Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis", Technical report, HP Laboratories, 2011

[8] A. Mudinas, D. Zhang, M. Levene, "Combining lexicon and learning based approaches for concept level sentiment analysis", Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining, ACM, New York, NY, USA, Article 5, pp. 1-8, 2012.

[9] Ji Fang and Bi Chen, "Incorporating Lexicon Knowledge into SVM Learning to Improve Sentiment Classification", In Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), pages 94–100, 2011

[10] Lina L. Dhande and Dr. Prof. Girish K. Patnaik, "Analyzing Sentiment of Movie Review Data using Naive Bayes Neural Classifier" Indian Journal of Computer Science and Engineering (IJCSE), Volume 3, Issue 4 July-August 2014

[11] P.Kalaivani, "Sentiment classification of movie reviews by supervised machine learning approaches" P.Kalaivani et.al / Indian Journal of Computer Science and Engineering (IJCSE) Vol. 4 No.4 Aug-Sep 2013

[12] Vidisha M. Pradhan, Jay Vala and Prem Balani, "A Survey on Sentiment Analysis Algorithms for Opinion Mining" International Journal of Computer Applications (0975 – 8887), Volume 133 – No.9, January 2016

[13] A. Suresh and C.R. Bharathi, "Sentiment Classifi cation using Decision Tree Based Feature Selection" International Science Press IJCTA, 9(36), 2016, pp. 419-425

[14] Bhumika M. Jadav and Vimalkumar B. Vaghela, "Sentiment analysis using support vector machine based on feature selection and semantic analysis" International Journal of Computer Applications (0975 – 8887), Volume 146 – No.13, July 2016

[15] Aamera Z. H. Khan, Dr. Mohammad Atique, Dr. V. M. Thakare, "Sentiment Analysis Using Support Vector Machine", International Journal of Advanced Research in Computer Science and Software Engineering , Volume 5, Issue 4, April 2015

[16] Preety , Sunny Dahiya, "Sentiment analysis using SVM and NAÏVE BAYES algorithm" , International Journal of Computer Science and Mobile Computing, Vol.4 Issue.9, September- 2015

[17] Jayashri Khairnar1 , Mayura Kinikar, "Sentiment analysis based mining and summarizing using SVM-MapReduce", International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014

BIOGRAPHY

Anjali Rai : She is pursuing her M.Tech from BBD University, Lucknow, Uttar Pradesh on Computer Science and Engineering. She has completed her B.Tech from BBDNIIT(UPTU), Lucknow, Uttar Pradesh on Information and Technology. Her area of interest is web mining.