International Journal of Scientific Research in Computer Science and Engineering **Research Paper**
Vol.6, Issue.3, pp. 8-12 , June (2018) **E-ISSN:** 2320-7639

Sentiment Analysis of movie reviews: A new feature-based sentiment classification

Ketan Sarvakar[1*], Urvashi K Kuchara[2]

[1]Ganpat University- U V Patel college of Engineering, Kherva, Mehsana, Gujarat, India
[2]Computer Engineering Department Ganpat University, Mehsana, Gujarat, India

Available online at: www.isroset.org

Received: 03/Jun/2018, Revised: 11/Jun/2018, Accepted: 21/Jun/2018, Online: 30/Jun/ 2018

***Abstract*—** Sentiment Analysis also known as opinion mining is the task of detecting, extracting and classifying opinions or sentiments related to different topics. One such area of interest is sentiment classification or polarity determination of movie reviews in user specific choice which are dependent on either mood or emotion of user perspectives. This plays an important role in today's world where the promotion of nay product or movies. Polarity determination is an important task for both the user and producer. They can take appropriate decision based on these results of classification. Thus, considering the needs and developing interests in social data mining and increasing dependency of users on customer reviews here we proposed a method to classify the data more accurately by altering the pre-processing tasks mainly filtering. Proposed methodology will be used for the available classification techniques by using the available dataset more consistently by working on the dictionary built by the filters.

***Keywords*—** Sentiment Analysis; Dictionary; tokenizer; Accuracy; StringToWordVector, CustomStringToWordVector filter

I. INTRODUCTION

Sentiment Analysis (SA) [1] is the task of analyzing the user opinions on different topics. SA is the computational study of opinions, sentiments, emotions and attitude expressed in texts towards an entity. Nowadays sentiments and reviews are becoming much evident due to growing interest in the e-commerce. Online social media and social networking sites are the main platform to express and evaluate various entities. Customers Viewers, producers and service providers mainly rely on the user reviews to make their further decision which can improve their business.

Reviews are generally classified based on their polarity which may be positive or negative. To make the classification more accurate we worked with the available dataset and tried to use it more sufficiently. We worked on the number of words in the dictionary built by the filter. Using the enhanced dictionary with the appropriate tokenizer to classify the data yields more accuracy. Thus, in this paper we discuss the importance of pre-processing tasks, how the proposed filter will yield more accuracy and also compared the results of previous and new filter based on the parameters like accuracy and precision.

Kumar Ravi, Vadlamani Ravi [1] surveyed on work done in different areas, tasks and sub-tasks of sentiment analysis. They compared the work of different authors on different tools. They compared the classifiers such as Naïve Bayes (NB), Support Vector Machine (SVM) and other classifiers. Out of these in most of the Works the SVM gives the highest accuracy. Asmita Dhokrat, Sunil Khillare, C Namrata Mahender [10] performed a survey on the different tools for the opinion mining. Peter F Brown, Peter V. deSouza, Robert L. Mercer [9] worked on predicting a word from next word they discussed the n-gram models based on the frequency of their co-occurrence with other words.

In this paper we show the comparison of tools on the same dataset also the comparison of different classifiers in different tools based on their accuracy we also compared the performance of the modified filter with the existing one. Thus, the paper is organized as follows, Section Ⅰ contains the introduction and work done by various authors, section Ⅱ describes the previous and related work in this area and our motivation for this work, Section Ⅲ contains the brief explanation about the preliminary steps of Data mining, Section Ⅳ shows the existing problem with the available techniques and some of the preliminary steps, Section Ⅴ shows the flowchart of the proposed work and the solution to the problem mentioned in section Ⅳ, Section Ⅵ shows how the proposed method works along with measures used lastly, Section Ⅶ shows the comparison and results of the previous and the proposed method. Also, we have compared the above-mentioned methods to show the performance of different tools.

II. Related Work

Sentiment Analysis has been mainly approached as classification. In which many classification techniques are used for example Naïve Bayes, SVM, etc. The classification with accuracy is a major task in which the accuracy matters a lot like many of the authors like Kumar Ravi, Vadlamani Ravi [1] and Asmita Dhokrat, Sunil Khillare; C Namrata Mahender [10] has survey on tools and techniques of sentiment analysis. They have performed and explained each phase of sentiment analysis. Also they have concluded that the SVM classifier gives the highest accuracy at all levels of classification.

The Survey till now has focused on tools and techniques for comparing the accuracy of the classifier, but a little work has been done on the preprocessing tasks or in enhancing the preprocessing techniques. Dr. S Vijayarani and Ms. R. Jinani [3] discussed about one of the preprocessing tasks i.e. tokenization. They discussed the importance of tokenization and various tools for tokenization. Similarly, Ms. Anjali Jivani [4] has survey on various stemming algorithms and stemming methods. These works can be more enhanced with the classification techniques to increase the performance of classifier.

Thus, by getting inspirations from this previous works we proposed this method which works on the preprocessing tasks. Our method shows how the preprocessing tasks play an important role in sentiment classfication. Also the classifier when used with properly preprocessed dataset is must for classification. Thus, we worked on these tasks to retain the important words in the dataset which can be used to increase the accuracy of the classifier. We discussed and compared the same in this paper.

III. Preliminary steps of datamining

Sentiment Analysis is a complicated yet important task in the field of data mining; it needs various steps to be performed for satisfactory opinion mining from the texts. The main steps are Data Acquisition and Pre-processing.

A. Data Acquisition

Data Acquisition means gathering proper dataset for classification. It can be obtained from online resources like social networking sites, micro-blogging sites and various other media. The datasets are freely available on sites like imdb.com, rotten tomatoes, etc.

B. Pre-processing

After data acquisition the other major task is to preprocess the dataset. The data available on the online resources contains noise. It can be inconsistent; it may be containing various noises like missing data and may lack in various behavior or trends and is likely to contain many errors. Thus, pre-processing involves techniques to transform the raw data into an understandable format. Some steps involved in pre-processing are Stemming, Stop word removal, tokenization, parts of speech (POS) tagging, etc. [1].

- **Stemming**- Stemming is the process to reducing word forms or various grammatical forms of words into its root form. The various grammatical forms may be noun, adjective, verbs, adverbs, etc. Stemming is usually done by removing the suffixes and prefixes from an index terms before an actual assignment of the term to the index.
- **Tokenization**- Tokenization is the process in which a stream of data or textual input is broken into words, terms, symbols, or some other meaningful elements called tokens. It is used to identify the meaningful keywords [3]. It is the main step as it is used by the classifier to make comparison if the words.
- Some words need to be removed or dropped as they do not contribute much to the analysis such as **Stop words**. Stop words are some common words which are of a little value. To recognize different parts of speech present in the text POS tagging is performed.

IV. Problem with the existing preprocessinf filters

The pre-processing task contributes much in classifying the dataset properly. The Pre-processing tasks are available in the StringToWordVector (STWV) Filter. The Filter uses this all features and creates the dictionary of the words. The dictionary is used to classify the dataset further.

Here the 1st problem is that the generated dictionary size is much smaller than the actual words or instances present in the dataset because of which there are not enough words for the classifier to classify the review or sentiment properly.

Suppose if there are 10000 instances in the dataset containing 5000 positive and 5000 negative review and the dictionary size or the words to keep in the dictionary is just 1000 then the dictionary will not have sufficient words to classify the sentiments properly. Thus, the size of the dictionary is a big problem.

The 2nd problem is that the tokenizer used must be properly chosen to make the classifier more powerful. The words in the dictionary express different meaning and describe the review or sentiment as positive or negative. As by a single word or token, classifier is not able to classify that token properly as positive or negative.

For example: The sentiment “This is the most beautiful palace in India” is divided into tokens in the dictionary like

Table 1 Example of Tokenization

This	is	the	Most	Beautiful	Palace

Here the token ‘most’ does not classify any meaning. Similarly, the token ‘Palace’ also does not signify any meaning properly. But instead if we use group of three words then it conveys some useful meaning like ‘Most beautiful Palace’ this specify it as the positive sentence. For this reason, it becomes necessary to supply proper group of words in the dictionary to the classifier.

Thus, we tried to solve this problem by proposing the method discussed in the next section.

V. Proposed Work

Looking to the problem we worked on increasing the size of the dictionary by proposing a new filter. The whole proposed work is divided in the number of steps. These steps are as shown below in the Figure 1.

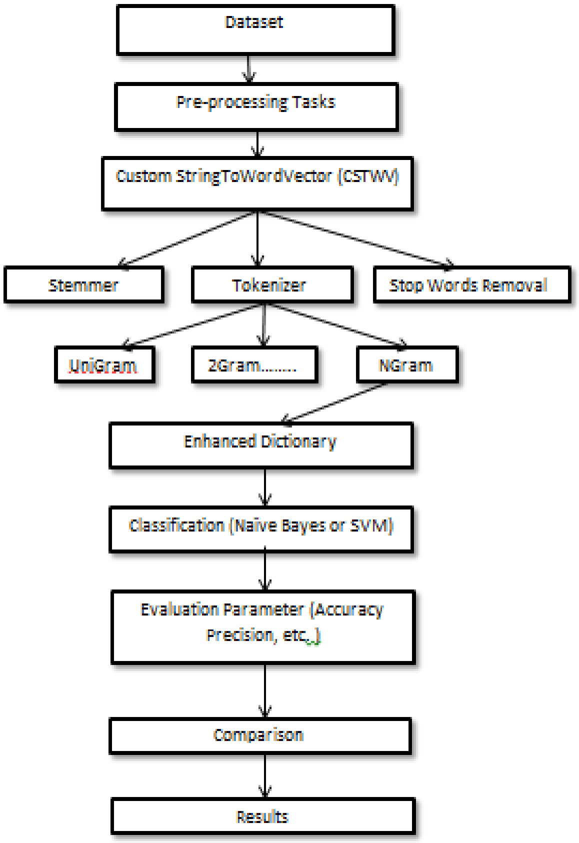

Figure 1 Flow chart of the proposed work

VI. Implementation of the Proposed Work

Dataset –Here we collected the movie review from the imdb.com website. We worked on the 10000(5000 positive and 5000 negative) movie reviews collected from the website.

Tool used- We used Weka-3.8.2 for classifying the dataset.

We Pre-processed the dataset by using the modified filter (STWV) and using this filter with NGramTokenizer.

To enhance the accuracy, we first increased the size of the dictionary by proposing a new CustomStringToWordVector filter by modifying the StringToWordVector filter. For the modification and enhancing the capability in this filter we used the variable ‘wordstokeep’ for setting the appropriate words in the dictionary. The filter then chooses the size of the dictionary by counting the instances present in the dataset. We count the number of instances present in the dataset and assign this value to the wordstokeep parameter present in the filter to increase the no. of words in the dictionary. Thus, if the size of the dictionary is same as the number of instances present in the dataset then there will be less loss of words and there will be more words in the wordlist of the dictionary.

Further the length of the tokens is specified by the NGramTokenizer by which proper group of words can be created, that can be used by the classifier for classification.

It allows us to choose the length of the words like 1Gram or Unigram, 2Gram upto NGrams as shown in Figure 2.

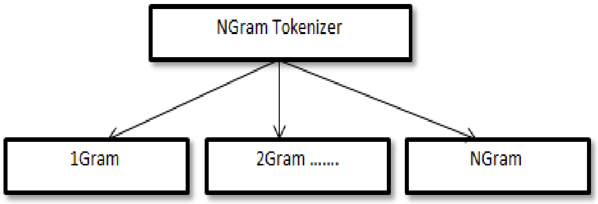

Figure 2 NGramTokenizer

We used the NGramTokenizer with minimum and maximum values of the token size as 1 and 5 respectively.

We modified the source code of the filter in Weka by extracting Weka in Net Beans for which we used Net Beans 8.0.2. We calculated the size of the instances coming out from the method for calling the dataset and assigned that size in the filter for choosing the wordstoKeep parameter. After applying these specifications of the filter performed classification on the dataset by using the Naïve Bayes classifier and SVM classifier. The results of previous and the modified filter are compared in the next section.

VII. Result and Discussion

After performing the whole classification by using both the filters we got better results from the

CustomStringToWordVector filter. The Classification is performed on the 10000 movie reviews. The graph in the figure shows the comparison between the previous and the modified filter with the enhanced size of the dictionary. We measured the accuracy and precision of the classifier and compared it. The results of comparison are shown in the Figure below.

The performance measures are calculated as:

Accuracy = (Sum of correctly classified reviews/ Total reviews),

Precision = TP/ (TP+FP),

Recall = (TP/ number of positively oriented reviews in the dataset)

Where TP, FP are the True positive and False positive respectively obtained during the classification.

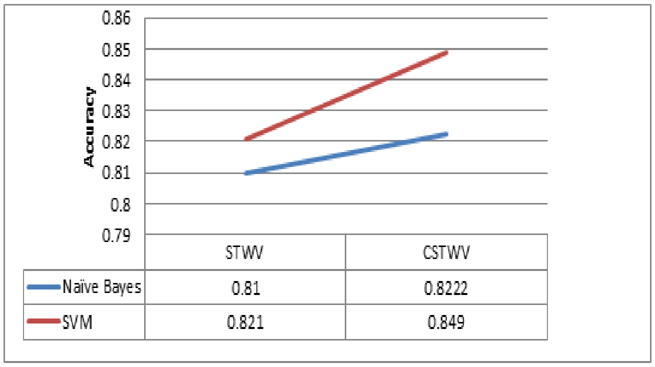

Figure 3 Comparison of accuracy

The graph in the Figure 3 shows the comparison of accuracy. Here we have compared the accuracy of Naïve Bayes (NB) and SVM classifier. After performing preprocessing tasks using STWV and CSTWV we got higher accuracy by using CSTWV filter for both NB and SVM. The values are in terms of percentage.

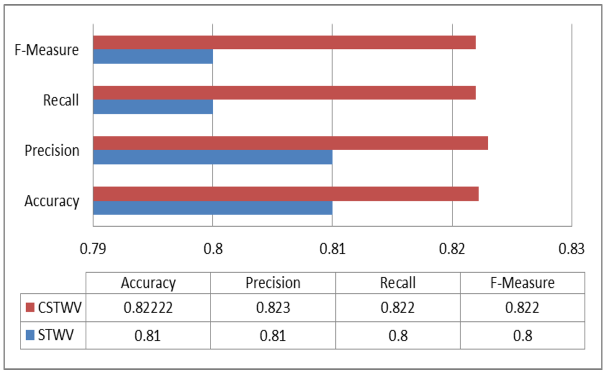

Figure 4 Comparison of performance parameters of NB

The chart in the Figure 4 shows comparison of various measurement parameters of Naïve Bayes Classifier. They are compared using the two filters (STWV and CSTWV). In this comparison we have used parameters like accuracy, precision, recall and F-Measure. In each parameter the result is improved using CSTWV filter. The value of each parameter is calculated in percentage.

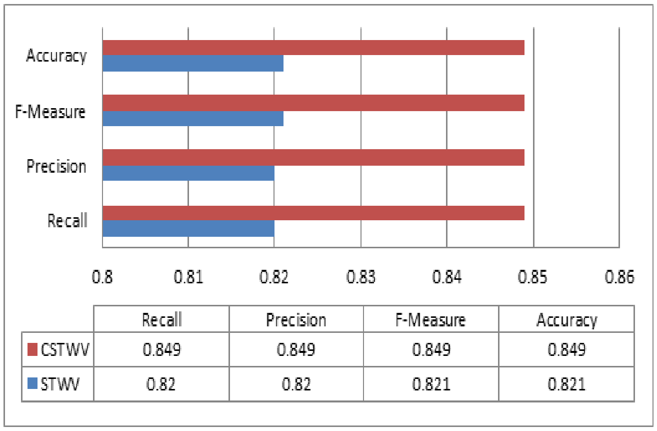

Figure 5 Comparison of performance parameters of SVM

The chart in the Figure 5 shows various measure of SVM classifier compared using STWV and CSTWV filters. Again, it shows higher values of the parameters for CSTWV filter compared to STWV filter.

This comparison also marks that the SVM classifier is more powerful than Naïve Bayes. As it yields more accuracy by using each filter.

VIII. CONCLUSION

According to our proposed work we worked with the pre-processing tasks of the String to word vector filter. We enhanced the pre-processing features and increased the word list in the dictionary. The enriched dictionary with sufficient words along with the NGram tokenizer is used to get higher accuracy of classification. By the experiment and the modified filter, we got high accuracy and precision for both Naïve Bayes and SVM Classifiers. With the String to word vector filter we get Accuracy as 81% and 82.1% for Naïve Bayes and SVM respectively whereas by using Custom string to word vector filter we get 82.2% and 84.9% for NB and SVM respectively. Thus, we can conclude from this that the word list made by the filter and the pre-processing tasks plays an important role in enhancing the performance of the classifier. Also, the more the no. of words in the dictionary it is easy for the classifier to make comparisons and classify the dataset properly. It is concluded from this experiment that appropriate number of words along with proper tokenization method yields higher accuracy of classification.

We have used weka for this work, In future this work can be implemented in other tools also. Here we used String to word vector filter for pre-processing tasks and modified this filter in future this can be done for other filters also by using different tokenization methods and stemming algorithms.

ACKNOWLEDGMENT

I thank my grateful faculty of my college at Ganpat University for this research, without which the present research could not been completed I am especially indebted to my guide faculty Prof. Ketan Sarvakar, who have been Supportive and worked actively to provide me knowledge. I am thankful to all my friends for their support during the research. I would like to thank my parents, whose love and guidance with me in whatever I pursue.

REFERENCES

[1] K. Ravi, V. Ravi , "*A survey on opinion mining and sentiment analysis: Tasks, approaches and applications,*" Knowledge-Based Syst., Vol. 89, pp. 14–46, 2015.

[2] S. V. B. Pang, L. Lee, "*Thumbs up? Sentiment Classification using machine learning techniques,*" ACL-02 Conf. Empir. Methods Nat. Lang. Process., Vol. 10, pp. 79–86, 2002.

[3] Dr. S. Vijayarani, Ms. R. Janani, " *TEXT MINING: OPEN SOURCE TOKENIZATION TOOLS – AN ANALYSIS,*" Advanced Computational Intelligence: An International Journal (ACII), Vol. 3, No. 1, 2016.

[4] Ms. Anjali Ganesh Jivani, "*A Comparative Study of Stemming Algorithms,*" Anjali Ganesh Jivani et al, Int. J. Comp. Tech. Appl., Vol. 2 (6), pp. 1930-1938.

[5] Nan Hu, Noi Sian Koh, Srinivas K. Reddy, "*Ratings lead you to the product, reviews help you clinch it? The mediating,* " Decision Support Systems, Vol. 57, pp. 42–53, 2014.

[6] Diana Maynard, Ian Roberts, Mark A. Greenwood, Dominic Rout, Kalina Bontcheva, "*A framework for real-time semantic social media analysis,*" Web Semantics: Science, Services and Agents on the World Wide Web, 2017.

[7] Manajit Chakraborty, Sukomal Pal, Rahul Pramanik, C. Ravindranath, "*Recent developments in social spam detection and combating techniques: A survey,*" Information Processing and Management, Vol. 52, pp. 1053–1073, 2016.

[8] Donato Hernández Fusilier, Manuel Montes-y-Gómez b, Paolo Rosso c, "*Detecting positive and negative deceptive opinions using,"* Information Processing and Management, Vol. 51, pp. 433–443, 2015.

[9] Peter F. Brown, Peter V. deSouza*, Robert L. Mercer,V. Pietra, J. C. Lai, *"Class-Based n-gram Models of Natural,"* IBM T. J. Watson Research Center .

[10] Asmita Dhokrat, Sunil Khillare, C. Namrata Mahender, *"International Journal of Computer Applications Technology and Research - Review on Techniques and Tools used for Opini on Mining,"* International Journal of Computer Applications Technology and Research, Vol. 4, Issue 6, pp. 419 - 424.

[11] Jesse Read, Bernhard Pfahringer, Geoff Holmes, Eibe Frank, *"Classifier Chains for Multi-label Classificatio,"*@cs.waikato.ac.nz in machine learning, 2011.

[12] Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu, Wayne Niblack., *"Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques,"* Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03) , Vol. 0-7695-1978-4/03.

[13] J. H. Zaragoza, L. E. Sucar, E. F. Morales, C. Bielza, and P. Larranga, *"Bayesian chain classifiers for multidimensional classification,"* In the proceedings of 2011 IJCAI Int. Jt. Conf. Artif. Intell., pp. 2192–2197, 2011.

[14] S. ChandraKala and C. Sindhu, "*Opinion Mining and Sentiment Classification: A Survey,"* ICTACT JOURNAL ON SOFT COMPUTING, Vol. 03, Issue. 01, 2012.

Authors Profile

Prof. Ketan Sarvakar pursed Bachelor of Technology, in 2006 and Master of Technology from Nirma University in year 2008. He is currently pursuing Ph.D. and currently working as Assistant Professor in Department of Information Technology, Ganpat University, India since 2008. He has published more than 20 research papers in reputed international journals including conferences including IEEE and it's also available online. His main research work focuses on Cryptography Algorithms, Network Security, Cloud Security and Privacy, Big Data Analytics, Data Mining, IoT and Computational Intelligence based education. He has 10 years of Teaching and Research Experience.

Urvashi K Kuchara pursed Bachelor of Technology and Master of Technology from Ganpat University- U V Patel College of Engineering,Gujarat- INDIA in year 2018. Her main research work focuses on Sentiment Analysis, Big Data Analytics, Data Mining, IoT and Computational Intelligence based education. She has 2 years of Research Experience.

.