Optimal Feature Selection by Distribution Diversity for Sentiment Analysis

A.V.R.Mayuri, K.Asha Rani

Abstract: ***In this article, the feature selection method and the classification of opinion methods are compared and discussed. The performance of feature selection methods utilized the z-scores & t-scores statistical measure. The classifier SVM is utilized for comparing and classifying with the Adaboost and NB. The main aim of this article is to discuss and assess the range of the statistical measures to detect the features which are optimal and its importance for the opinion classification utilizing the diverse classifiers. The analysis of performance are conducted on diverse datasets with varied range such as reviews of product, reviews of movies, and tests & tweets are conducted utilizing "Wilcoxon Signed Rank based Z-score and T-score. And from the outcomes of simulation studies, it is obvious that amid 3 classifiers which are tested for the accuracy of classification, the method Adaboost has surpassed the other 2 methods of NB and SVM.***

Keywords: Adaboost, SVM, NB, Optimal Feature, Decision trees, K-nearest-Neighbors.

I. INTRODUCTION

E-commerce has become an integral part of the consumer buying trends. In the present scenario, billions of purchases are taking place over the ecommerce platforms, and with thousands of online transactions taking place every hour. According to global research stats, many consumers take a detailed review of the product, pricing and other feature exploration over the ecommerce platform, even if they are considering the offline purchase of the product. Millions of consumers provide feedback, reviews, product usage experience and their expectations over the ecommerce platforms and also in varied social networking platforms. Such reviews are effective sources of understanding the expectations of the consumers, their preferences, and the product usage experiences. Also, such a platform is becoming a very effective channel for promoting the range of new products or services to the consumers, even with customized intimations. Numerous sites like the amazon.com, epionions.com and many other such sites have emerged in the recent past that enables the consumers to exchange views on products. In the instance of a user considering buying a product, just by visiting a product review site and reading the reviews, considerable insights over the product, features, performance and other attributes could be attained. However, reviewing all the reviews and gathering insights could be time consuming process, and there is need for comprehensive system for analyzing the reviews and finding information about the product quality and other such features. In the recent past, machine learning classifiers has been effectively put to use for solving the aforesaid issue, by using the sentiment classification solutions.

In this paper, emphasis is on the sentiment mining based solutions aimed at discovering the consumer preferences, mindset of the consumers over the products, and the consumer selection reflects on the product as a negative or positive impression. In the proposed review, use of measures of statistical for theselection of feature which are optimal and its impact over the divergent classifiers in the data mining, for sentiment mining is reviewed. Also the feature selection methods effectivenesson datasets which are diverseis also evaluated.

Sentiment Analysis

Analysis based on direction is the fundamental factor considered in sentiment mining, opinion mining, polarity mining or the sentimental analysis process that are carried out by varied organizations. Pre-processing of data is very significant in terms of data gathered from the social media networks as the data could be unstructured, prone to grammatical errors or spelling mistakes or such syntactical issues. Certain pre-processing elements like the spell check, stop words removal are carried out as an integral part of natural language pre-processing solutions [1] [2].

The text containing the inputs of perceptions, emotions and opinions of consumer are deliberated for the analysis of sentiment and the procedure contains several tasks. The work [3] presents that 4 important tasks which are integral for the analysis of sentiment where many of the researchers concentrate on the data which is pre-processed, class labelling, target detection and annotation granularity.

In further class labeling procedure, some of the researchers concentrate on the text classification as objective/subjective. Mainly, the work [4] presents that, in instance of analysis of sentiment, the class labeling task is conducted former to the classification of polarity, as outcomes are imperative that such execution enhances latter. For example, when text is detected to be subjective then the classification of polarity for defining such kind of subjective text conducts either negative or positive sentiment. And several researchers have underscored on the automation of procedure of the class-labeling utilizing different supervision utilizing the noisy tags.

For instance, certain special characters & emoticons used for labeling tweets [5] are used for indicating positive or negative impression towards a specific mindset. But, such noisy label could hamper the performance of classifiers that focus on the sentiments. In [6], the method of exploiting the

Revised Manuscript Received on December 22, 2018.

A.V.R.Mayuri, Associate Professor, Department of CSE, G.Pulla Reddy Engineering College (Autonomous), Kurnool, AP, India (Email: drmayuri.cse@gprec.ac.in@gmail.com)

K.Asha Rani, Associate Professor, Department of CSE, G.Pulla Reddy Engineering College (Autonomous), Kurnool, AP, India (Email: asha.ashreddy@gmail.com)

Twitter follower graph has been proposed for improving the classification of sentiment and build a graph which makes utilize of word unigram, tweets, hashtags, emotions and word bigrams as nodes werelinked on the basis of existence of link amid them.

Further, the label propagation model with the tags of sentiment is utilized the small nodes set containing definite initial label data in the entire graph. Having subjective test which is pre-processed comprise of sentiment classification and class labels that will be performed at document [7], phrase levels [8] and sentence [9], which are deliberated usually as classification granularity. The work [10] presents that several researchers are in the opinion that analysis of sentiment knows the destination of sentiment and source is noticed as most important challenge in the analysis of sentiment.

Related Work

The literature review explore regarding some of studies which have concentrated on the assessment of the execution from the diverse algorithms of classification adapted for the mining of sentiment. The work [11] presents that comparative study is conducted on 4 classifiers such as SVM, Decision trees, K-nearest-Neighbors, NB method for assessing the sentiment mining execution for the reviews of online product. The vivid sampling models such as random sampling, bootstrap sampling and linear sampling are adapted for creating examples of training from the data set of product reviews. Outcomes from study indicate that SVM through the bootstrap sampling model executes better than other sampling models and classifiers for the rate of misclassification. The unigrams are utilized for the space of feature and occurrence of terms to settle the input classification. Nevertheless, the study did not offer any data regarding the impact of the input format relating to the outcomes of classification.

In [12], authors have assessed the performance of three classifiers. Utilizing the reviews of product like 100k reviews which are gathered online concentrate on effect of order-n-grams (N>3) that are higher, are used for evaluating PA (Passive-Aggressive) Algorithm Based Classifier, Language Modelling (LM) based classifier and Winnow Classifier. It is concluded in the study that PA classifier when integrated with maximum order n-grams as the features has achieved effective or comparable execution than the other stated models. Though the authors have reviewed 6-grams feature length, still any calculations pertaining to data representation over the performance of classifiers were not reflected.

In a study conducted by Vinodhini's and Hang's, study has focused on reviews of product comprising around 800 characters. But in the case of a Tweet, the limitation of 140 characters is the other key challenge envisaged in sentiment mining of tweets. Usually the datasets that are generate from twitter source encounter large sparsely. Higher order n-grams might not be suitable for usage. Considering superior peculiarities of text and length of a twitter, a twitter-specific comparative-study conducted to assess the execution of certain famous classification algorithms in the domain of mining of sentiment, by adapting varied formats of inputs that actually have significant effect on classification accuracy.

Methods and Materials

This segment details regarding classifiers utilized aimed at the classification of sentiment and measurements of statistical which are adapted for the selection of optimal feature.

Machine Learning Classifiers

In machine learning procedure and modeling of statistics, detecting the set classification of towards related categories and correct type of the mapping for novel notice is the main procedure. The entire procedure depends on the data training group effectiveness containing related annotations and its classification of membership is defined. And in methods of "machine learning", "supervised learning" determined to be classification whereas unsupervised learning is deliberated to be clustering. And the clustering procedure is generally clustering the data into several categories on the basis of definite measures of the innate features.

Support Vector Machine

The "SVM (Support Vector Machine)" *[13]*is an organized algorithm of "supervised machine learning" which is adapted for the regression and classification challenges. Yet, it is generally utilized in the issues of classification. In algorithm, each item in the data is graphed to be point in dimensional space-n through value of the each marked values which is coordinate for every feature. And now the classification procedure is executed by noticing hyper-plane that can distinguish 2 classes in the productive manner. SVM are easy coordinates of the single observation and the frontier is the SVM which productively partition the 2 classes.

1.1.1 Naive Bayes

The method NB*[14]* is intensely relied on the "Bayes' theorem" which have independence assumption amid predictors. The classifiers NB assume that each feature particular exist in class which is not relevant to any of the other feature. And for case, Apple is deliberated to be fruit, when it is reed color, round in shape and with definite diameter measures. Even though such kind of features will depend on the every other and with the presence of the other features, definitely all such kind of properties will be contributing independently to the possibility that apple is a fruit. And such kind of procedure is called "Naïve Bayes method". Method NB is very simple to construct and is useful for maintain voluminous datasets volume. Further to easy procedures, the NB method is called for its performance and productivity when compared with several high end models of classifications. Utilizing the former and predictive probability of class, predicting the future probability is possible through "Bayes theorem".

1.1.1 AdaBoost

The Adaboost*[15]* method is useful for the execution boosting of the decision trees on the issues of "binary-

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

classification". And it is also productive in performance enhancing of algorithms related to machine-level, and intense ininstance of the feeble learners. This kind of methods obtains accuracy through random opportunity of the issue of classification. The one of the level of decision trees are appropriate and general algorithm utilized with the Adaboost. Because trees were short and one of the decisions is comprised for the classification, they were called to be decision-stumps.

Feature Selection Strategies

The words bag is achieved from specified training set that are latter sieved to achieve the words which are optimal for the utilization as the features. And the words which are often related to the representative lexicons of sentiment is filtered initially to be the probable features. Moreover, the features which are optima is chosen on the basis of covariance among frequency of co-occurrence with sentiment lexicons of negative and positive for every feature. And the features which have important covariance among co-occurrence frequency through corresponding negative & positive lexicons of sentiment is deliberated to be optimal-features.

In esteem to predict variance of every feature frequency of co-occurrence by corresponding negative & positive representative lexicons of sentiment, here suggested a method deliberated "Wilcoxon Signed Rank based Z-score and t-score" metrics from the analysis statistically. T-score will be adapted for choosing the features which are optimal respective to every negative and positive words of representative lexicon of sentiment. Unlike to this the z-score will be deliberated to choose the features which are optimal from the negative & positive sets of training without deliberating their relationship through the representative lexicons of sentiment.

t-Score

The different values in 2 diverse vectors is depicted by t-score that is predicted in the following way:

$$t-score=\frac{(M_{v1}-M_{v2})}{\sqrt{\frac{\sum_{i=1}^{|v1|}(x_i-M_{v1})^2}{|v1|-1}+\frac{\sum_{j=1}^{|v2|}(x_i-M_{v2})^2}{|v2|-1}}}$$

In the aforesaid equation

- M_{v1},M_{v2} depicts the mean of the values noticed in corresponding vectors $v1$, $v2$
- The representations x_i,x_j depicts every element of corresponding vectors $v1$, $v2$ of respective $|v1|$,$|v2|$ sizes

The t-score will be the ratio among mean differences of corresponding vectors and "square root of aggregate of mean square distances" of the corresponding vectors.

Then evaluate "degree of probability (p-value) [16]" in the t-table [17] aimed at t-score achieved. P-value which is lower than probability of threshold signifies both the vectors are discrete; therefore, the feature depicting corresponding vectors which is an optimal feature.

Wilcoxon Signed Rank based Z-score

Evaluating the t-score which utilized for computing feature co-occurrence variance through the representative lexicons of sentiment, and here utilized the "Wilcoxon Signed Rank based Z-score" for the selection of features which are optimal.

The Z- score that denotes the given two vectors are distinct or not is assessed as follows:

- The representation *a* will be the Existence of *f* feature in the positive-set
- The representation *b* will be theExistence of *f* feature in the negative-set
- The representation *c* will be theExistence of entire other features (other than *f* feature) in the positive-set
- The representation *d* will be theExistence of entire other features (other than *f* feature) in the negative-set

Then the "Wilcoxon Signed Rank based Z-score" the "Z of feature *f*" is evaluated in the following way:

$$Z(f)=\frac{a-(a+c)*\frac{(a+b)}{a+b+c+d}}{\sqrt{(a+c)*(a+b)*(1-(a+b))}}$$

The work [17] presents that moreover, finding the "degree of probability (p-value) of the Z-score *Z*(*f*)" from the z-table. When probability of degree will be lower than specified "probability of degree of threshold" then *f* feature will be optimal.

Experimental Study

In the analysis of performance of machine learning method on the dataset which is processed before& converted into group of the records so that, every record contains group of words. Further, "Wilcoxon Signed Rank based Z-score and t-score" are implemented for evaluating the features which are optimal from the dataset which is processed.

Every technique of selection of feature will assist through unique lessened features set. And every such kind of diverse set of features are tested for the accuracy utilizing the 3 classifiers which are diverse and are called Adaboost, SVM and NB. Same type of cooperation is conducted on the normal attributes set generated from the selection of feature across entire datasets. And in last step, the analysis and validation of feature sets which are minimized are executed relying on the "anatomic-relevance".

Data Statistics & Results

The reviews dataset of movies [18], reviews dataset of product [19] and analysis of sentiment [20] are the 4 datasets which are deliberated for the analysis. Entire datasets are preprocessed and no. of cases are available for every dataset which is represented in the table-1. When the reviews of datasets of product and movies possess the uniform dissemination, the dataset of twitter is skewed.

Table-1: statistics of data for every set of data

Data Set	Total # of records	No # sentiment representative lexicons
Twitter dataset	975	31
Movie reviews	442	27
Product Reviews	836	12

Selection of feature and statistics of performance

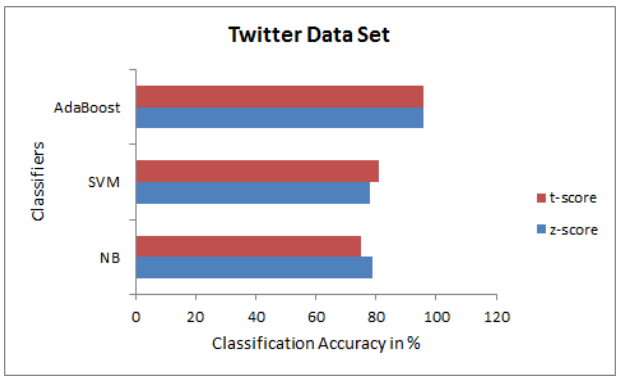
The methods of feature selection are "Wilcoxon Signed Rank based Z-score and t-score" which are selected and itemized the lessened attributes set which are represented in the table-2. Further, lessened features are evaluated utilizing the 3-classifiers and the outcomes are represented in Figure 1.

Table 2: statistics of outcomes achieved from the scheme of choosing the feature

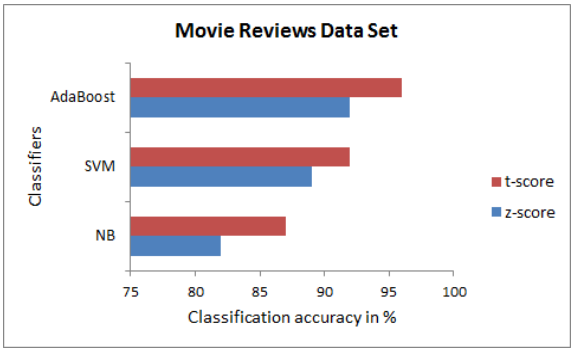
Dataset	No of Features	No # optimal Features by t-score	No # optimal Features by z-score
Twitter dataset	862	75	23
Movie reviews	240	91	53
Product Reviews	330	72	28

Validation is conducted depending on the correctly classified cases. From application of "t-score and Wilcoxon Signed Rank based Z-score feature selection" methods, it will be obvious that chosen features through t-score will be almost features selected subset through "Wilcoxon Signed Rank based Z-score". Execution dip will be noticed for the NB method & for the definite set of data when the method SVM will be applied in the t-score (in Figure 1). The outcomes represents that even though the selection of features methods are implemented, still require the developments in the entire execution of classifiers which became the issue and hence there is a requirement for concentrating on the features which are residual and their significance for the productive making of decision.

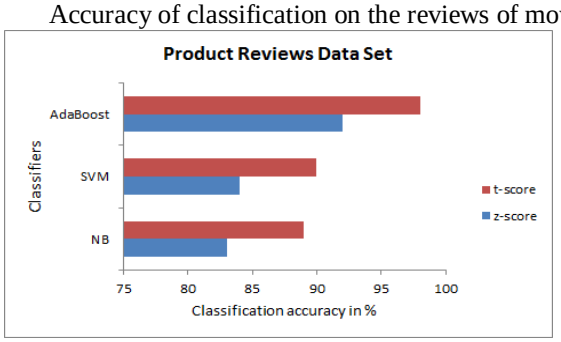
The classifier Adaboost executed better for the 3 datasets and the accuracy resulted to be 98%. And rest of the classifiers also showed the important enhancement in the execution, through NB classifier resulting 86% when compared with other classifiers execution of 90% at least. In spite of dataset of twitter and the review dataset of movie leveraging the highest execution 93% & 84% correspondingly, it is imperious that actual significance is of the chosen features by the t-score, hence resulting to the features study chosen by the "Wilcoxon Signed Rank based Z-score" across entire datasets.



a) Accuracy of classification on the dataset of Twitter



b) Accuracy of classification on the reviews of movies



c) Accuracy of classification on the reviews of product

Figure 1: performance analysis of classifier

II. CONCLUSION

The analysis of sentiment became an important segment for the administrations to know the expectations of consumer, conducts of buying and towards the analyzing, the main aspects which could affect decision-making. Upon emerging the trends of BI, the analysis of sentiment of mining the data has achieved the momentum and several researches are conducted in this field. And in this article, the importance is on utilizing the 3 classifiers Adaboost, SVM and NB for the selection of feature methods which could make important effect on the result. From the simulation studies which are conducted on the vivid range of the datasets and testing the execution utilizing the z-score and t-score, it is obvious that the classifier Adaboost has surpassed with the 98% of the accuracy levels across entire 3 data sets. Even though the features which are optimal discovered under the t-score is lower than the features which are optimal discovered utilizing the "Wilcoxon Signed Rank based Z-score", nevertheless, the classification accuracy is much greater in the instance of the analysis of t-score. So that the result indicates the effect of classifier Adaboost for the selection of feature and to productively execute the analysis of sentiment.

REFERENCES

1. Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1(12).
2. Bollen, J., Mao, H., & Pepe, A. (2011). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. Icwsm, 11, 450-453.

3. Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. ACM Transactions on Information Systems (TOIS), 26(3), 12.
4. Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.
5. Jiang, L., Yu, M., Zhou, M., Liu, X., & Zhao, T. (2011, June). Target-dependent twitter sentiment classification. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (pp. 151-160). Association for Computational Linguistics.
6. Speriosu, M., Sudan, N., Upadhyay, S., & Baldridge, J. (2011). Twitter Polarity Classification with Label Propagation over Lexical Links and the Follower Graph. EMNLP 2011, 53.
7. Salvetti, F., Lewis, S., & Reichenbach, C. (2004). Automatic opinion polarity classification of movie. Colorado research in linguistics, 17(1), 2.
8. Wilson, T., Wiebe, J., & Hoffmann, P. (2005, October). Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of the conference on human language technology and empirical methods in natural language processing (pp. 347-354). Association for Computational Linguistics.
9. Elarnaoty, M., AbdelRahman, S., & Fahmy, A. (2012). A machine learning approach for opinion holder extraction in Arabic language. arXiv preprint arXiv:1206.1011.
10. Perez-Tellez, F., Pinto, D., Cardiff, J., & Rosso, P. (2010, October). On the difficulty of clustering company tweets. In Proceedings of the 2nd international workshop on Search and mining user-generated contents (pp. 95-102). ACM.
11. Whitehead, M., & Yaeger, L. (2010). Sentiment mining using ensemble classification models. In Innovations and advances in computer sciences and engineering (pp. 509-514). Springer, Dordrecht.
12. Cui, H., Mittal, V., &Datar, M. (2006, July). Comparative experiments on sentiment classification for online product reviews. In AAAI (Vol. 6, pp. 1265-1270).
13. Suykens, J. A., &Vandewalle, J. (1999). Least squares support vector machine classifiers. Neural processing letters, 9(3), 293-300.
14. Murphy, K. P. (2006). Naive bayes classifiers. University of British Columbia, 18.
15. An, T. K., & Kim, M. H. (2010, October). A new diverse AdaBoost classifier. In Artificial Intelligence and Computational Intelligence (AICI), 2010 International Conference on (Vol. 1, pp. 359-363). IEEE.
16. Sahoo, P., & Riedel, T. (1998). Mean value theorems and functional equations. World Scientific.
17. http://www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf
18. http://www.cs.cornell.edu/people/pabo/movie-review-data/
19. http://www.dcs.bbk.ac.uk/~andrius/psenti/
20. Rosenthal, S., McKeown, K., & Agarwal, A. (2014). Columbia nlp: Sentiment detection of sentences and subjective phrases in social media. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014) (pp. 198-202).