A Supervised Opinion Mining for IPL 2018 Using Feature Subset Selection

S.Fouzia Sayeedunnisa[1] , Dr.Nagaratna P Hegde[2], Dr.Khaleel Ur Rahman Khan[3]
[1]*Dept. Of IT ,M.J. College of Engineering and Technology,*
[2]*Dept. Of CSE, Vasavi College of Engineering,*
[3]*Dept. Of CSE, ACE Engineering College,*
[1, 2, 3]*Hyderabad, Telanagana State, India*

ABSTRACT: With the growing popularity of Opinion rich resources such as online review sites, micro blogs and social networking sites which actively use information technology to seek out and understand the opinions of others. Opinion Mining of these opinion rich resources is "contextual mining" of text which identifies whether piece of text is positive, negative or neutral. Most of the existing sentiment analysis models envisaged the complexities, which is due to high volume of features . In this manuscript, the proposed solution is about identifying different features and applying a feature reduction technique for selecting optimal features from tweets of Twitter. Tweets can reveal mass opinion when taken in large amount i.e. events like IPL 2018.We analyze the performance of Naive Bayes classifier under the reduced feature subset selected using Information gain with respect to accuracy and the time taken for processing o tweets. The basis of our paper is to effectively perform sentiment Classification using reduced subset of features for the event IPL 2018. Results from the experimental study depict that the proposed solution can support in attaining effective classification accuracy levels of 85.9%, upon using less than 40% of the features .

KEYWORDS—*Social network, Opinion Mining, Machine Learning, Twitter.*

1 INTRODUCTION

The main aim of Sentiment Analysis is to extract and identify perceptions, sentiments from the huge data generated by the various social networking sites. With the immense growth of user generated content expressing opinions have become an integral source of information for the stakeholders of the business **.** This opinionated text gives more insights to the decision makers over the market trends and the consumer perceptions over a brand, product, and the services[1], [2].

With the recent advances in deep learning, the ability of algorithms to analyse text has improved considerably. Creative use of advanced artificial intelligence techniques can be an effective tool for doing in-depth research. Inferring sentiment is one of the significant challenges envisaged in the process of sentiment analysis [3]. The sentiment analysis process can be defined in two steps, i.e. the first step which is extracting the features from opinionated text and second, sentiment classification of this opinionated features. In the Sentiment analysis process of micro blogs like Twitter the key challenge is the huge size of text, irrelevant and overlapping data.[4].Feature Selection plays a wide role in the process of sentiment classification, and selecting an optimal feature subset based on Information Gain as the criteria, is the crux of process [5].

Feature selection process using the machine learning [6] has significant results earlier. Elimination of redundant and irrelevant features provides It is imperative from [7] that if right kind of feature selection methods were proposed. Such solutions can lead to elimination of irrelevant features from the vector, thus leading to reduced size of feature vector and increased accuracy of sentiment classification

Three different categories are defined of Feature Extraction and Selection techniques for Sentiment Classification. The first techniques deals with the problem of over-fitting and improving the performance of Sentiment Classification. Second dimension is about emphasis on cost-effective and less time-complexity oriented solutions. The third dimension is regarding the process of handling the enormous data generated from social media sources.The three main Feature Selection[8] methods are filtering, wrapping, and embedded approaches. In the filter model, it offers choice of selecting optimal subset of features using scaling and elimination of low-scoring features. The various filter methods are Information Gain, Chi square, Pearsons Correlation, LDA and ANOVA. In filter methods, features are selected on the basis of their scores in various statistical tests for their correlation with the outcome variable.

Twitter is a free micro-blogging service which allows users to publish their opinions about events and activities. Twitter posts called tweets are short messages generated continuously which is a huge source of streaming data for sentiment polarity detection**.**Key aim is about identifying and extracting opinions and sentiments from Twitter using most informative features obtained using IG and then measuring the accuracy of the classifier using he reduced subset of features.

2 RELATED WORK

Sentiment analysis is classified into three levels i.e. the document level, the sentence level and aspect-level [11], [12]. The document level analysis classifies a complete document as positive or negative sentiment [11], [13].

Sentence level classification analyses a complete sentence as positive , negative or neutral[114][15]. Five filter methods used for feature selections are DF (Document Frequency), IG (Information Gain), CHI (Chi-Square), MI (Mutual Information), and TF (Term Frequency) by [17] to eliminate

97% of the low-scoring features significant improved the accuracy.

A key role played by Feature selection and extraction methods in the selection and extraction of highly relevant features, which outputs optimal subset of features used for training the classifier with various machine learning algorithms[10].

In the other work by [18] classification accuracy has significantly improved by eliminating low scoring features, using IG and CHI methods than the other three methods [9].Different Supervised Machine Learning Techniques such as SVM, Naïve Bayes and CART , BackPropagation networks are used by several authors. Sentiment Classification using algorithms such as MaxEnt (Maximum Entropy), SVM and Naïve Bayes[18] are used to classify the text as positive or negative. Zhang et al . [19],examines the effectiveness of various machine learning techniques on providing a positive or negative sentiment on a twitter corpus.

Integration of two strategies towards feature selection is more resourceful from the review of literature In the proposed study, the focus is on integration of varied feature vectors and feature subsets by using IG and Naive Bayes as the classifier.

3 METHODOLOGY FOR IMPLEMENTATION.

Figure 1 describes the steps taken for building the sentiment analysis model using Twitter as a mode of streaming data. The data is extracted, collected and pre-processed .Feature extraction is done to extract six different features. IG is applied on these features to get a reduced set of most informative features which applied to Naive Bayes classifier gives significant accuracy in less time for analysing the opinions of the event IPL 2018.

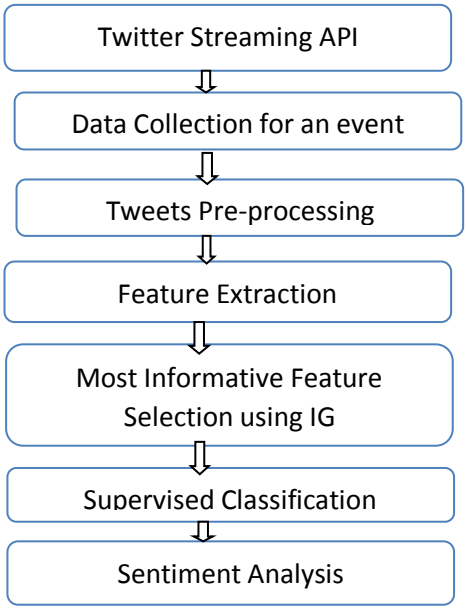

Fig 1:Block Diagram of Sentiment Classification

A. Data Collection

Twitter's streaming API was used for collecting the tweets containing the term "IPL 2018".Tweets were filtered and parsed containing the term "IPL 2018" and were stored in a Comma Separated Values (CSV) file format. A database of 3240 tweets was collected

for analysis. It contained 1927 positive sentiments and 1313 showed negative sentiment towards "IPL 2018".The csv data file is processed using Python. Python nowadays is used for data analytics and encloses several packages for text data analysis.

B. Twitter Tweet Pre-processing

Pre-processing plays a major role in the Sentiment Analysis task. It will clean the dataset by reducing its complexity in order to prepare the data for the classification task. Pre processing is a four step process in our model i.e.

a) Firstly, the data is changed to lower case.

b)Second the data is tokenized to split up the words into terms of tokens.

c)Third the stop words are removed to give the set of opinionated words

d)Fourth stemming which will reduce the tokens into a single type, normally a root word; for example, the word “pictures” will be reduce to “picture”. As such, the stemming process reduces redundant words in a document.

C. Feature Extraction

We incorporated six different features in our model. Opting a useful list of words as features of a text and eliminating a large number of words that do not contribute to the text’s opinion is termed as feature extraction.

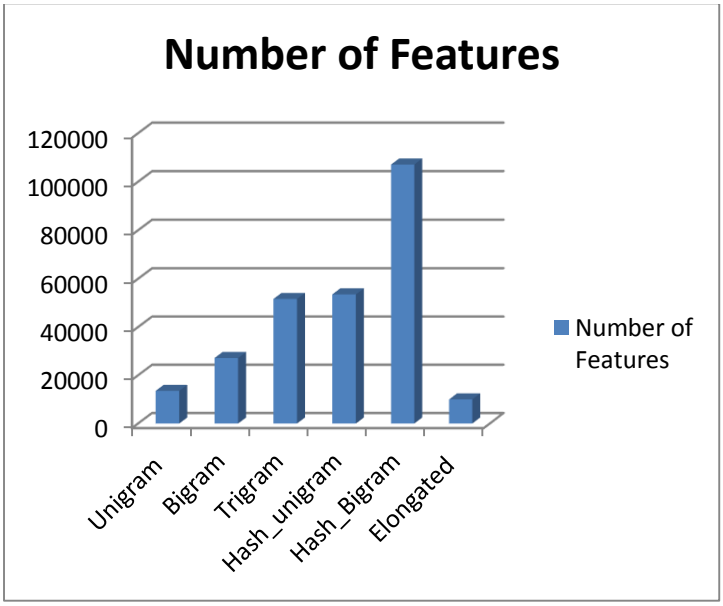

Fig 2 :A bar graphshowing "Frequency of Features Extracted"

The devised feature selection approach is extracting uni-gram, bi-gram, tri-gram, elongated words and hash tagged unigram, bigram features. In the initial level of the approach, the significant sentiment lexicons presented are compared with the tweets and the n- grams are separated. Hash tagged features play a major role in micro blogging site in carrying the opinions. Hash tagged unigrams and bigrams words are collected and compared with the sentiment lexicon set to separate as positive and negative hash tagged words. Another important feature which is very often used in users of micro blogging sites are the elongated words which carry sentiments. Elongated words are collected and separated using sentiment lexicon set.

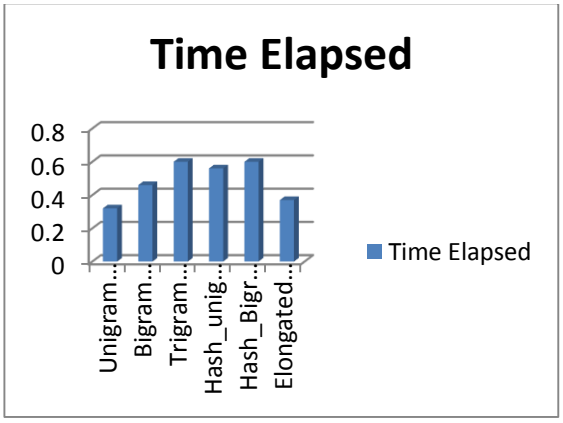

Fig3Time Elapsed Positive Feature Extraction

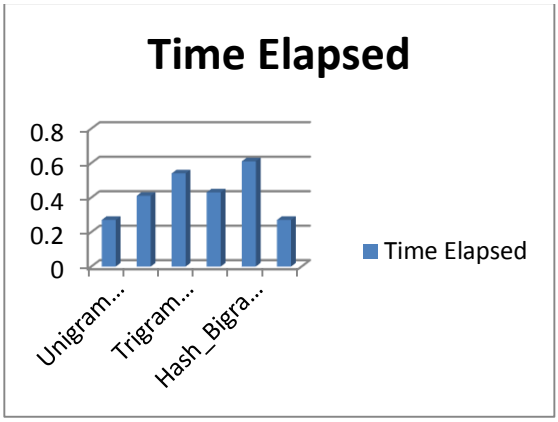

Fig4 Time Elapsed Negative Feature Extraction

4. MOST INFORMATIVE FEATURES SELECTION USING I.G.

For information gain (IG) we used the Shannon entropy measure [Shannon, 1948] in which:

IG(C,A) = H(C) - H(C|A)

where: IG(C, A) information gain for feature A;

H(C)=- ∑ p(C= i)log p (C= i) entropy across sentiment classes C;

H(C|A)=-∑p(C= i/ A)log p(C= i/ A) specific feature conditional entropy;

If the number of positive and negative sentiment messages is equal, H(C) is 1. Furthermore, the information gain for each attribute A will vary along the range 0-1 with higher values indicating greater information gain. All features with an information gain greater than 0.0025 are selected.

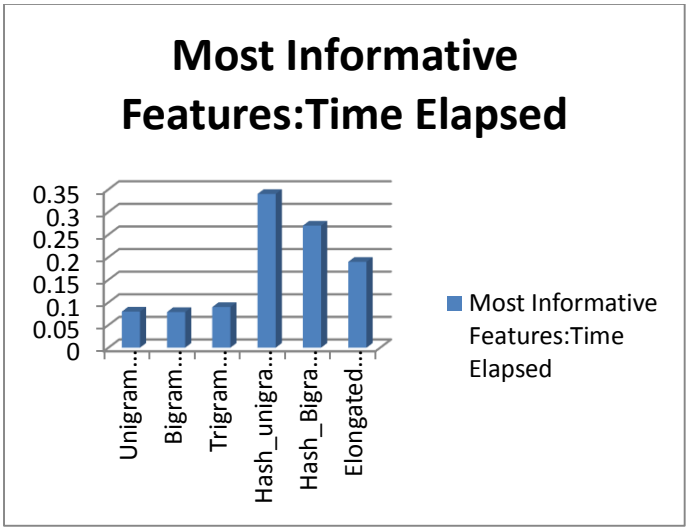

Fig5:Time Elapsed Most Informative Positive Feature Extraction

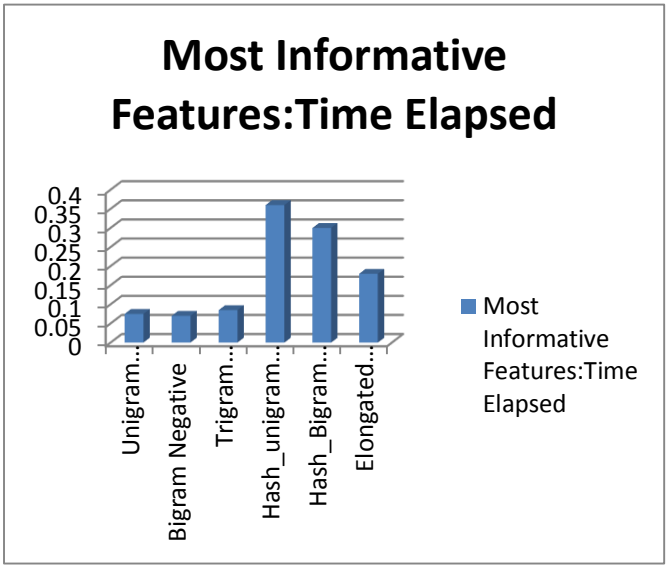

Fig6: Time Elapsed Most Informative Negative Feature Extraction

5.PERFORMING SUPERVISED LEARNING USING NAIVE BAYES CLASSIFIER

Naive Bayes Classifier and SVM are one of the most eminent and frequently used classifying algorithms for text analysis and categorization .A Naive Bayes classifier is used to train a model based on Twitter data. Three fourth of the twitter data which is compared with the lexicon set and

labelled as positive and negative are used for training the model. One fourth of the unlabelled data is used for testing. Naive Bayes classifier, Decision trees such as Random Forest are one of the most widely used machine learning algorithms. Much of their popularity is due to the fact that they can be adapted to almost any type of data. They are a supervised machine learning algorithm that divides the training data into smaller and smaller parts in order to recognize patterns which can be used for classification. The information is then presented in the form of logical structure similar to a flow chart that can be simply understood without any statistical knowledge.

6. RESULT EVALUATION AND SENTIMENT CLASSIFICATION

The models which were built using IG as the feature reduction technique and Naive Bayes

algorithm were applied to the corresponding testing set, to predict the opinion of the tweets. This outcome showed that with less number of features selected using IG the accuracy ofthe classifier was never compromised. For analysis, all the results were combined. A table showing the accuracy of the multiple features and the time elapsed for classification is shown below . The performance of the proposed "IG using Naive Bayes" approach was effective in attaining classification accuracy levels of 85.9%, upon using less than 40% of the features .

Table 1:Accuracy and Time Elapsed for Most Informative Features

Most Informative Features	Unigram	Bigram	Trigram	Hash_Unigram	Hash_Bigram	Elongated words
Time Elapsed	0.14	0.23	0.24	0.71	0.75	0.11
Accuracy	78.3	78.9	85.4	81.2	85.9	75.23

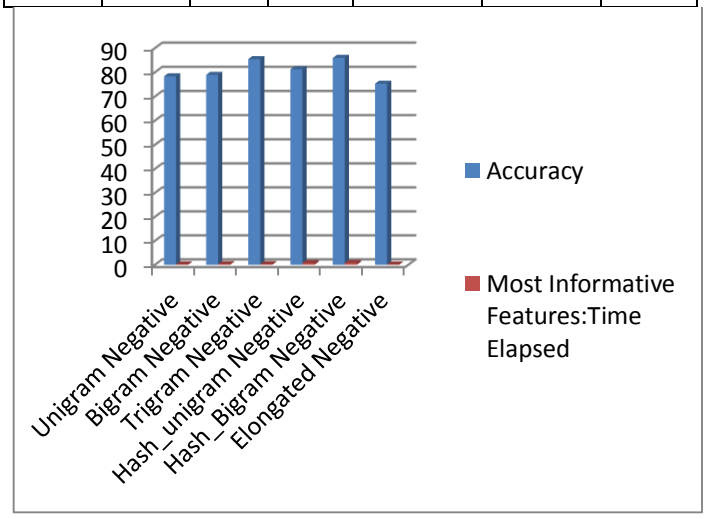

Fig7:Accuracy and Time Elapsed for Most Informative Features

7 CONCLUSION

Sentiment analysis has become an integral need for the organizations to understand the consumer expectations, buying behaviour and towards analysing the key factors that could influence the decision-making.Twitter tweets Sentiment analysis gives a detail insights of opinions of thousands of micro blogging users for sporting event IPL 2018.The results demonstrate the positive as well as negative insights of people. Such an analysis could provide valuable feedback to the organisations intended to do business and help them to gain users perception. If users perception are negative then early determining of negative trends can allow them to make well defined business decisions on how to target specific aspects of their services and products in order to increase business. This paper illustrates the approach of 'Most Informative Feature Selection using IG' and classification using Naive Bayes Classifier which has major effect on overall accuracy of the analysis. In future work, author(s) will try to include sentiment from emoticons, make the classification multiclass i.e. showing strongly positive, positive, neutral, negative or strongly negative opinions.

REFERENCES

[1.] Hu, Minqing, and Bing Liu. "Mining opinion features in customer reviews." AAAI.Vol. 4.No. 4. 2004.

[2.] Liu, Yang, et al. "ARSA: a sentiment-aware model for predicting sales performance using blogs." Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval.ACM, 2007.

[3.] Gangemi, Aldo, Valentina Presutti, and Diego ReforgiatoRecupero. "Frame-based detection of opinion holders and topics: a model and a tool." IEEE Computational Intelligence Magazine 9.1 (2014): 20-30.

[4.] Elawady, Rasheed M., SherifBarakat, and Nora M.Elrashidy. "Different feature selection for sentiment classification." International Journal of Information Science and Intelligent System 3.1 (2014): 137-150.

[5.] Liu, Huan, and Lei Yu. "Toward integrating feature selection algorithms for classification and clustering." IEEE Transactions on knowledge and data engineering 17.4 (2005): 491-502.

[6.] Koncz, Peter, and Jan Paralic. "An approach to feature selection for sentiment analysis." Intelligent Engineering Systems (INES), 2011 15th IEEE International Conference on IEEE, 2011.

[7.] Agarwal, Basant, and Namita Mittal. "Sentiment classification using rough set based hybrid feature selection." Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis (WASSA'13), NAACL-HLT. Atlanta. 2013.

[8.] Saeys, Yvan, IñakiInza, and Pedro Larrañaga. "A review of feature selection techniques in bioinformatics." bioinformatics 23.19 (2007): 2507-2517.

[9.] Yousefpour, Alireza, et al. "A comparative study on sentiment analysis." Advances in Environmental Biology (2014): 53-69.

[10.] Bharti, KusumKumari, and Pramod Kumar Singh. "Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering." Expert Systems with Applications 42.6 (2015): 3105-3114.
[11.] Liu, Bing. "Sentiment analysis and opinion mining." Synthesis lectures on human language technologies 5.1 (2012): 1-167.
[12.] Yousefpour, Alireza, Roliana Ibrahim, and HazaNuzlyAbdullHamed. "A novel feature reduction method in sentiment analysis." International Journal of Innovative Computing 4.1 (2014): 34-40.
[13.] Taboada, Maite, et al. "Lexicon-based methods for sentiment analysis." Computational linguistics 37.2 (2011): 267-307.
[14.] McDonald, Ryan, et al. "Structured models for fine-to-coarse sentiment analysis." Annual Meeting-Association for Computational Linguistics.Vol. 45.No. 1. 2007.
[15.] Nakagawa, Tetsuji, Kentaro Inui, and SadaoKurohashi. "Dependency tree-based sentiment classification using CRFs with hidden variables." Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.Association for Computational Linguistics, 2010.
[16.] Dessì, Nicoletta, and Barbara Pes. "Similarity of feature selection methods: An empirical study across data intensive classification tasks." Expert Systems with Applications 42.10 (2015): 4632-4642.
[17.] Rogati, Monica, and Yiming Yang. "High-performing feature selection for text classification." Proceedings of the eleventh international conference on Information and knowledge management.ACM, 2002.
[18.] Selmer, O., & Brevik, M.: Classification and Visualisation of Twitter Sentiment Data. Master's Thesis, NTNU- Trondheim. (2013).
[19.] Zhang, Linhao: Sentiment analysis on Twitter with stock price andsignificant keyword correlation. Diss. 2013. The University of Texas at Austin. (2013.
[20.] Liao, T. Warren. "Feature extraction and selection from acoustic emission signals with an application in grinding wheel condition monitoring." Engineering Applications of Artificial Intelligence 23.1 (2010): 74-84.
[21.] Tripathy, Abinash, Ankit Agrawal, and Santanu Kumar Rath. "Classification of sentiment reviews using n-gram machine learning approach." Expert Systems with Applications 57 (2016): 117-126.
[22.] Murphy, Kevin P. "Naive bayes classifiers." University of British Columbia (2006).
[23.] Speriosu, Michael, et al. "Twitter polarity classification with label propagation over lexical links and the follower graph." Proceedings of the First workshop on Unsupervised Learning in NLP.Association for Computational Linguistics, 2011.
[24.] Suykens, Johan AK, and JoosVandewalle. "Least squares support vector machine classifiers." Neural processing letters 9.3 (1999): 293-300.
[25.] Guyon, Isabelle, et al. "Gene selection for cancer classification using support vector machines." Machine learning 46.1-3 (2002): 389-422.
[26.] An, Tae-Ki, and Moon-Hyun Kim. "A new diverse AdaBoost classifier." Artificial Intelligence and Computational Intelligence (AICI), 2010 International Conference on.Vol. 1. IEEE, 2010
[27.] Rey, Denise, and Markus Neuhäuser. "Wilcoxon-signed-rank test." International encyclopedia of statistical science.Springer Berlin Heidelberg, 2011. 1658-1659.
[28.] B. Pang, L.Lee, and S.Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In EMNLP '02: Proc. of the ACL-02 conf. on Empirical methods in natural language processing, pages 79–86.ACL, 2002.
[29.] http://web.stanford.edu/class/cs424p/materials/ling287-handout-09-21-lexicons.pdf
[30.] Sahoo, PK-Riedel, and T. Mean Value Theorems. Functional Equations.World Scientific, 1998.
[31.] Ihaka R, Gentleman R. R: a language for data analysis and graphics.Journal of computational and graphical statistics. 1996 Sep 1;5(3):299-314
[32.] http://thinknook.com/wp-content/uploads/2012/09/Sentiment-Analysis-Dataset.zip
[33.] http://www.cs.cornell.edu/people/pabo/movie-review-data/
[34.] http://jmcauley.ucsd.edu/data/amazon/