|| Volume 6 || Issue 3 || March 2021 || ISSN (Online) 2456-0774

INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH AND ENGINEERING TRENDS

AN APPROACH FOR ASPECT BASED SENTIMENT CLASSIFICATION USING MACHINE LEARNING ALGORITHMS

Mr.Bipin Ghorpade[1], Prof. Vandana Navale[2]

ghorpadebipin23@gmail.com[1], vandananavale@dpcoepune.edu.in[2]

Computer Engineering Dhole Patil College of Engineering Pune, India.[1]

Computer Engineering Dhole Patil College of Engineering Pune, India.[2]

-- ***---

Abstract: - Aspect-based sentiment analysis is divided into two tasks aspect extraction and related sentiment identification. To carry out this task, features play an important role to determine the accuracy of the model. Feature extraction and feature selection techniques contribute to increase classification accuracy. Feature selection strategies reduce computation time, improve prediction performance, and provide a higher understanding of the information in machine learning or pattern recognition applications. This work focuses on aspect extraction from restaurant review dataset. In this system, we proposed a hybrid approach of feature selection which works on lemma features. Initially, the extracted features undergo pre-processing and then the term frequency matrix is generated which contains the occurrence count of features with respect to aspect category. In the next phase, different feature selection strategies are applied which includes selecting features based on correlation, weighted term frequency and weighted term frequency with the correlation coefficient. The performance of weighted term frequency with correlation coefficient approach is compared with the existing system and shows improvement classification accuracy of system.

Keywords: *Aspect-Based Sentiment Analysis (ABSA), Natural Language Processing (NLP), Term Frequency-Inverse Document Frequency (TF-IDF), feature extraction, feature selection, correlation coefficient.*

---***---

I INTRODUCTION

As the Web expands, the horizons are expanding. Social media and micro blogging sites such as Facebook, Twitter, Tumbler dominate fast-paced dissemination of encapsulated news and trending topics around the globe. A subject becomes a phenomenon when more and more people share their views and observations, making it a reliable source of understanding online. Twitter is an online tweet-driven networking site that is limited to 140 character tweets. The character limit therefore enforces the use of hash tags to classify text. Roughly 6,500 tweets are currently being published per second, resulting in around 561.6 million tweets a day. Such tweet streams are typically noisy representing various subjects, shifting details about attitudes in unfiltered and unstructured format. Analysis of feelings on Twitter means using natural language processing to isolate, classify and describe the meaning of feelings. Sentiment Analysis is often conducted at two stages (1) coarse, and (2) fine. At the gross level, the analysis of whole documents is carried out while at the fine level, the analysis of attributes is carried out. In terms of tonality, polarity, lexicon and tweet grammar, there are many challenges involved. We tend to be highly unstructured and unstructured. This project uses the rapid processing capabilities of Apache Spark to analyze feelings from such high-speed tweets in real-time. A topic becomes trending if more and more users are contributing their opinion and judgments, thereby making it a valuable source of online perception. These topics generally intended to spread awareness or to promote public figures, political campaigns during elections, product endorsements and entertainment like movies, award shows. Large organizations and firms take advantage of people's feedback to improve their products and services which further help in enhancing marketing strategies. One such example can be leaking the pictures of upcoming iPhone to create a hype to extract people's emotions and market the product before its release. Thus, there is a huge potential of discovering and analyzing interesting patterns from the infinite social media data for business-driven applications. Sentiment analysis is the prediction of emotions in a word, sentence or corpus of documents.

II LITERATURE SURVEY

According to [11] sentiment analysis is a vastly used term to classify user's opinion using Natural Language Processing (NLP) and Machine Learning (ML) Approaches. Various researchers have used different methods for aspect based classification and polarity based classification [1], [3], [8], etc. Product review based sentiment analysis is similar to the proposed sentiment analysis approach. Figure 1 summaries the basic model of the sentiment classification task.

Mahdieh Labani et. al. [12] proposed a multivariate filter method for feature selection which is used for various text classification approach. This method focuses on the reduction of redundant features using minimal-redundancy and maximal-relevancy concepts. The proposed method takes into account document frequencies for each term, while estimating their usefulness. It not only selects the features with maximum relevancy but also the redundancy between them is taken into account using a correlation metric. Results obtained using this approach are better than state-of-the-art filter methods.

Asriyanti Indah Pratiwi and Adiwijaya[13] proposed feature selection and classification based on information gain for document sentiment analysis. Information Gain Classifier (IGC) is used to extract the various features from movie review dataset. Authors proposed IG-DF-FS based hybrid method called a combination of Information Gain + Document Frequency Feature Selection etc.

Haoyue Liu et. Al. [14] proposed a system of feature selection for imbalanced data. If the dataset is imbalanced, it has a problem of bias-to-majority. This issue is solved using Weighted Gini Index (WGI) approach. The WGI approach calculates an impurity reduction score for each feature and features with a high score are considered as important.

Asha S Manek et. al. [15] proposed aspect term extraction for sentiment analysis of movie review dataset. The work is carried out using the Gini index approach for feature selection after NLP processing. It uses SVM classifier to classify test data. This study illustrates a statistical method for weight calculation by Gini Index method for feature selection in sentiment analysis. This framework for sentiment analysis using SVM classifier is compared with other feature selection methods on movie reviews and results have shown that classification by using this efficient method has improved accuracy.

Muhammad ZubairAsghar et. al. [16] proposed a system aspect-based opinion mining framework using heuristic patterns. The work proposed an integrated framework comprising of an extended set of heuristic patterns generated using POS tags for aspect extraction, a hybrid sentiment classification module with the additional support of intensifiers and negations, and a summary generator. The system obtained classification results with improved precision (0.85) when compared to the alternative methods available. This method is quite generalized and it can classify aspect-based opinions on multiple domains.

Kim Schouten et. al. [17] proposed a system aspect category detection for Sentiment Analysis for supervised as well as unsupervised learning. In this work, the first method presented is an unsupervised method that applies association rule mining on co-occurrence frequency data obtained from a corpus to find aspect categories. The second, supervised, method uses a rather straightforward co-occurrence method where the co-occurrence frequency between annotated aspect categories and both lemmas and dependencies is used to calculate conditional probabilities. If the maximum conditional probability is higher than the associated, trained, threshold, the corresponding aspect category is assigned to that sentence. The accuracy of the system is around 83% for a supervised method.

Laith Mohammad Abualigah et. al. [18] proposed a feature selection method which is a hybrid of Genetic operators (GA) and particle swarm optimization algorithm for text clustering. The hybrid approach improved the accuracy of text clustering. The GA is used to solve the unsupervised feature selection problem, called Feature Selection based Genetic Algorithm for Text Classification (FSGATC). This method is used to create a new subset of informative features in order to obtain more accurate clusters on different review text datasets. This method also overcomes the other comparative methods in improving text clustering results based on different common benchmark datasets used in the domain of text mining.

Basant Agarwal et. al. [19] proposed a system concept-level sentiment analysis with Dependency-Based Semantic Parsing. This system illustrates a fundamental issue of the sentiment analysis task and uses concepts as features. It presents a concept extraction algorithm based on a novel concept parser scheme to extract semantic features that exploit semantic relationships between words in natural language text. The system also extracts the actual concept using ConceptNet ontology like RDF framework. Concepts extracted from the text are sent as queries to ConceptNet to extract their semantics. It selects important concepts and eliminates redundant concepts using the Minimum Redundancy and Maximum Relevance feature selection technique. All selected concepts are then used to build a machine learning model that classifies a given document as positive or negative.

V. K. Singh et. al. [20] proposed a method for sentiment analysis of movie review dataset. This work illustrates a domain-specific feature-based heuristic for aspect-level sentiment analysis of movie reviews. The system devised an aspect-oriented system that analyses the textual reviews of a movie and assigns it a sentiment label on each aspect. Authors also have done document-level classification involving use of different linguistic features (ranging from Adverb+ Adjective combination to Adverb + Adjective + Verb combination). This System used SentiWordNet dictionary to compute the sentiment class label for document-level sentiment analysis as well as for aspect-based sentiment analysis.

III PROPOSED SYSTEM DESIGN

In the system, various feature selection approaches are analysed and hybrid approach for feature selection is proposed. The feature selection strategies analysed are:

In this approach, feature selection is done on the basis of term frequency count. Term frequency of each feature with respect to each aspect category is calculated. A threshold is set for feature selection The result of this is, for each aspect category a respective term frequency matrix is generated. Further, a compound matrix is generated containing terms with their occurrence count in all aspect categories. From this matrix, a binary train matrix is generated where '1' is considered for non-zero term frequency. The supervised approach for aspect category extraction which selects relevant features and avoids redundancy by calculating correlation among features. Acquired results show that weighted term frequency with correlation approach has comparatively more F-score. In this work, it is observed that features selected using weighted term frequency are more relevant but also redundant. Redundancy among features in an aspect category is avoided by calculating the correlation.

Algorithm Design

1 : Stop word Removal Approach

Input: Stop words list L[], String Data D for remove the stop words.

Output: Verified data D with removal all stop words.

```
Step 1: Initialize the data string S[].
Step 2: initialize a=0,k=0
Step 3:  for each(read a to L)
             If(a.equals(L[i]))
Then Remove S[k]
End for
Step 4: add S to D.
Step 5: End Procedure
```

2 Stemming Algorithm.

Input : Word w

Output : w with removing past participles as well.

```
Step 1: Initialize w
Step 2:  Intialize all steps of Porter stemmer
Step 3: for each (Char ch from w)
        If(ch.count==w.length()) && (ch.equals(e))
          Remove ch from(w)
Step 4: if(ch.endswith(ed))
   Remove 'ed' from(w)
Step 5: k=w.length()
          If(k (char) to k-3 .equals(tion))
  Replace w with te.
Step 6: end procedure
```

3 TF-IDF

Input: Each word from vector as Term T, All vectors V[i…n]

Output: TF-IDF weight for each T

```
Step 1: Vector = {c1, c2, c3….cn}
Step 2: Aspects available in each comment
Step 3: D = {cmt1, cmt2, cmt3, cmtn}
        and comments available in each document
        Calculate the Tf score as
Step 4: tf (t,d) = (t,d)
         t=specific term
         d= specific document in a term is to be found.
Step 5: idf = t → sum(d)
Step 6: Return tf *idf
```

IV RESULTS AND DISCUSSIONS

The proposed system is implemented in java with windows environment; some inbuilt functions are used during the feature selection as well as extraction. Experimental analysis is explained below.

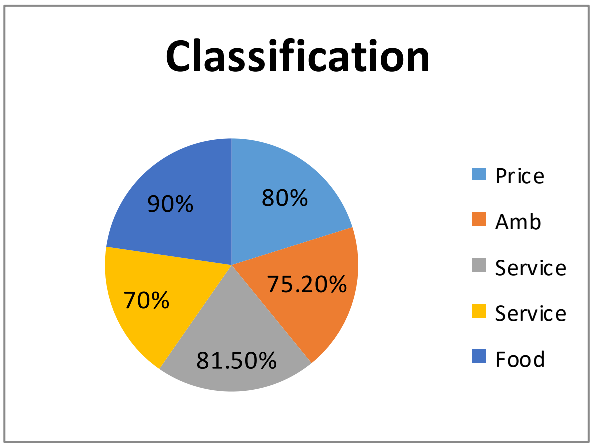

Figure 2 : Aspect category classification accuracy

Figure shows the percentage distribution of the number of sentences in each aspect category in test data. For this experimentation, restaurant review dataset is used which contains 700 training instances and 300 test instances.

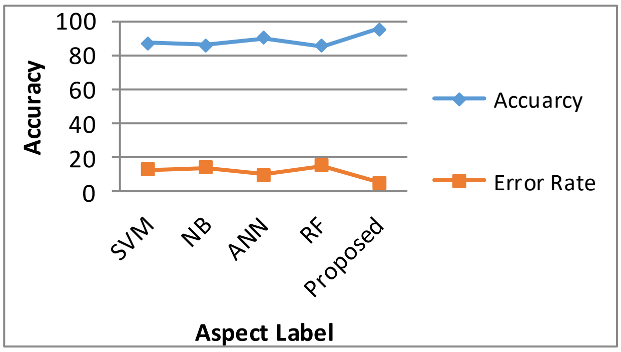

Figure 3 : Overall classification accuracy

V CONCLUSION

The proposed system describes feature extraction and feature selection approach using various techniques, basically the system proposed NLP approach for data pre-processing as well as data normalization. Select important features from entire data set all document it is much important for accurate classification. The system works with basic NLP features like tokenization, stopword removal, lemmatization, POS tagging and dependency parser respectively. Once the pre-processing has done system deals with feature extraction, in this phase we extract Bi-tagged features as well as dependency rule base features including lemmas features. To select specific features from extracted vector according to aspect category, five aspect categories has considered during the feature selection. In this phase we also add some synonyms for respective tokens to achieve accuracy for build the train model. After completion of whole process you will apply prospective classifier to generate the rules and system training has completed.

REFERENCES

[1] Randa Benkhelifa, Nasria Bouhyaoui and Fatima Zohra Laallam, “A Real-Time Aspect-Based Sentiment Analysis System of YouTube Cooking Recipes”,Springer Nature Switzerland AG, Machine Learning Paradigms: Theory and Application, Studies in Computational Intelligence 801, https://doi.org/10.1007/978-3-030-02357-7_11, 2019.

[2] Nadeem Akhtar, NashezZubair, Abhishek Kumar, Tameem Ahmad, “Aspect based Sentiment Oriented Summarization of Hotel Reviews”, Procedia Computer Science 115, 5 63–571, 2017.

[3] Satuluri Vanaja, Meena Belwal, “Aspect-Level Sentiment Analysis on E-Commerce Data”, Proceedings of the International Conference on Inventive Research in Computing Applications (ICIRCA 2018).

[4] Deepa Ananda, Deepan Naorema, “Semi-supervised Aspect Based Sentiment Analysis for Movies using Review Filtering”, 7th International conference on Intelligent Human Computer Interaction, IHCI 2015.

[5] Muhammad Afzaal, Muhammad Usman, Alvis Fong, “Tourism Mobile App with Aspect-Based Sentiment Classification Framework for Tourist Reviews”, IEEE Transactions on Consumer Electronics Vol: 65 , May 2019.

[6] Hajar E Hannach, Mohammed Benkhalifa, “WordNet based Implicit Aspect Sentiment Analysis for Crime Identification from Twitter”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 9, No. 12, 2018.

[7] Asha S Manek, P DeepaShenoy, M Chandra Mohan, Venugopal K R, “Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier”, Springer Science+Business Media New York 2016.

[8] MahdiehLabani, Parham Moradi, FardinAhmadizar, Mahdi Jalili, “A novel multivariate filter method for feature selection in text classification problems”, Engineering Applications of Artificial Intelligence 70, Elsevier, 2018.

[9] AlperKursatUysal, “An improved global feature selection scheme for text classification”, Expert Systems with Applications, 2015.

[10] SoujanyaPoria, Erik Cambria, Lun-Wei Ku, Chen Gui, Alexander Gelbukh, “A Rule-Based Approach to Aspect Extraction from Product Reviews”, Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP), pages 28–37, Dublin, Ireland, August 24 2014.

[11] Data extraction performance with pos tag pattern of dependency relation in aspect-based sentiment analysis. In2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP) 2018 Mar 26 (pp. 1-6). IEEE.

[12]Labani M, Moradi P, Ahmadizar F, Jalili M. A novel multivariate filter method for feature selection in text classification problems. Engineering Applications of Artificial Intelligence. 2018 Apr 1;70:25-37.

[13] Pratiwi AI. On the feature selection and classification based on information gain for document sentiment analysis. Applied Computational Intelligence and Soft Computing. 2018;2018.

[14] Liu H, Zhou M, Lu XS, Yao C. Weighted Gini index feature selection method for imbalanced data. In2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC) 2018 Mar 27 (pp. 1-6). IEEE.

[15] Manek AS, Shenoy PD, Mohan MC, Venugopal KR. Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier. World wide web. 2017 Mar 1;20(2):135-54.

[16] Asghar MZ, Khan A, Zahra SR, Ahmad S, Kundi FM. Aspect-based opinion mining framework using heuristic patterns. Cluster Computing. 2017:1-9.

[17] Schouten K, Van Der Weijde O, Frasincar F, Dekker R. Supervised and unsupervised aspect category detection for sentiment analysis with co-occurrence data. IEEE transactions on cybernetics. 2017 Apr 14;48(4):1263-75.

[18] Abualigah LM, KhaderAT, Al-Betar MA. Unsupervised feature selection technique based on genetic algorithm for improving the text clustering. In2016 7th international conference on computer science and information technology (CSIT) 2016 Jul 13 (pp. 1-6). IEEE.

[19] Agarwal B, Poria S, Mittal N, Gelbukh A, Hussain A. Concept-level sentiment analysis with dependency-based semantic parsing: a novel approach. Cognitive Computation. 2015 Aug 1;7(4):487-99.

[20] Singh VK, Piryani R, Uddin A, Waila P. Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification. In2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s) 2013 Mar 22 (pp. 712-717). IEEE.