A Framework for Sentiment Extraction using Hybrid Feature Extraction Method

K ANUSHA[1] , D. VASUMATHI[2]

[1]Research Scholar, Dept. of Computer Science and Engineering, University College of Engineering, Science&Technology Hyderabad, JNTU, Hyderabad, 500085, India.
E-Mail: anusha.kollu87@gmail.com

[2]Professor & HOD, Dept. of Computer Science and Engineering, University College of Engineering, Science&Technology Hyderabad, JNTU, Hyderabad, 500085, India.
E-Mail: rochan44@gmail.com

***Abstract*— Due to the increasing importance of sentiment analysis in analyzing customer feedback, it has been suggested that a framework be developed that combines the various features of deep learning and traditional lexical features. This paper aims to provide a framework that enables companies to perform effective sentiment analysis. The framework is composed of three components: post-processing, feature extraction, and sentiment classification. The first one uses a hybrid approach that combines the traditional features of lexical analysis and contextual features extracted from deep learning models. This method provides a more complete understanding of the sentiments expressed in texts. The second component is sentiment classification, which is performed through a machine learning model that has been trained on hybrid features. The BoW, TF-IDF, Word Embedding perform well in sentiment analysis and outputs a label for each input. The post-processing component of the framework involves analyzing the output of the classification process and applying regression method to improve its accuracy. The results of the study revealed that the hybrid feature extraction technique performed better than the traditional methods when it came to sentiment extraction. The paper also highlighted the advantages of this framework in sentiment analysis by combining the three different techniques.**

Keywords—Automated Sentiment Extraction, Feature Selection, Natural Language Processing, Sentiment Analysis, Machine Learning, Efficiency

I. INTRODUCTION

Sentiment analysis has been given a boost by big data. before sentiment analysis. Entropy and evolutionary techniques fuel our sentiment analysis binary classification model. Two domains were thoroughly researched using Stanford Sentiment Treebank and WHO COVID-19 public responses. Selecting appropriate features helps with more precise dataset labeling. 70% feature size decreases processing time and space [1].

Henceforth, it is natural to realize that the demand for building a feature driven sentiment extraction model is crucial.

The rest of the work is elaborated as in Section – II, the recent research outcomes and the problems in the existing research are furnished, the proposed solution details are furnished in Section – III, the obtained results are discussed and compared with other parallel research outcomes in Section – IV and V and the last Section – VI furnished the research conclusion**.**

II. RECENT RESEARCH REVIEWS

Twitter sentiment analysis and dimensionality reduction are challenging areas of research. As a result of dimensionality reduction, classification and computation may be improved. Reactions may be measured in terms of tweet features. Sparse data points in feature matrices increase the difficulty and error rate of Twitter sentiment analysis. We compare the performance of Naive Bayes and Support Vector Machines (SVMs) on three Twitter datasets in terms of classification accuracy (area under curve) and efficiency. Two well-known feature selection (dimensionality reduction) approaches, IG and Pearson's correlation, are evaluated with CAARIA (PC). The performance of CAARIA is outstanding. The classification information obtained by CAARIA is well-prepared [2].

Stock market volatility, non-linearity, and complexity are difficult to predict. Academics are split on whether or not investors' emotions affect the stock market. Feelings-based math did not work. In this study, we use a refined method of emotion analysis to foretell the behavior of the stock market. Five predictions for the stock market based on how investors feel. Specifically, we use algorithmic feature selection and stacked regularized models to examine causality. Our strategy for predicting the stock market's direction, based on emotion, is superior to [3] by a factor of 60%.

Sentiment analysis is used in a wide variety of contexts, such as gauging customer satisfaction or determining the source of hate speech. Twitter sentiment analysis is complicated by missing data, high-dimensional feature vectors, and skewed classifications. In the end, we overcame these hurdles to quantify the impact of Twitter sentiment

analysis. Urdu Tweets were the first to use sentiment analysis to identify hate speech. Sentiment classifier performance was improved with the use of dynamic stop words filtering, VGFSS, and SMOTE by addressing sparsity, dimensionality, and class imbalance. Sentiment analysis is the most effective method for identifying hate speech [4]. Class skew and high dimensionality reduction are particularly helpful.

Dimensionality reduction and a good classification model are necessary for microblog sentiment classification because of the size of the texts involved. The TF-IDF value doesn't take into account microblog equivalents Form-semantic. This is followed by a feature selection using NewChi-TF-IDF. Generalization based on a single classifier does not work. Compared to other feature selection methods, NewChi-TF-IDF is more effective in reducing dimension, increasing generalization, and increasing average F-score. [5].

Depression has worsened as a result of modern ways of living. Despite progress, psychiatric disorders are often misdiagnosed. Sad or potentially dangerous persons are identified automatically. Analysis of language use and representation of linguistic features are required for depression diagnosis. Choosing and blending features boosts efficiency. [6].

III. Feature Selection for Automated Sentiment Extraction

Finally the predictive sentiment extraction method is furnished and is illustrated here [Fig – 1].

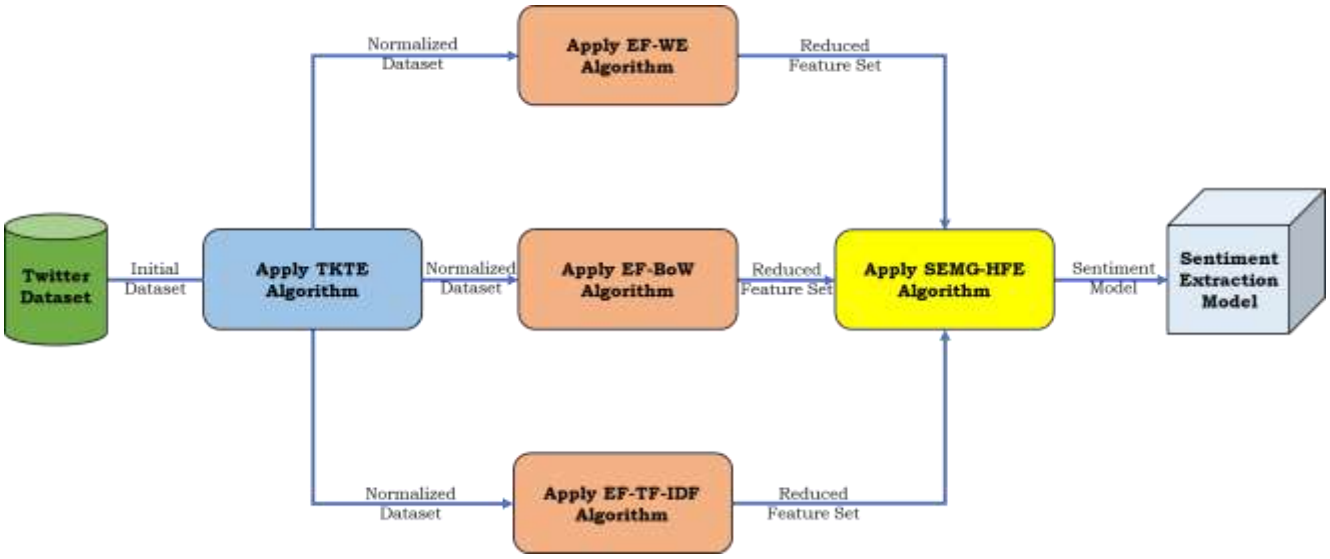

Fig. 1. Feature Selection for Automated Sentiment Extraction Model

One of the most crucial tasks in the extraction of sentiment from a text is to identify the overall opinion or sentiment expressed in it. A proposed model would involve selecting subsets of the extracted features from the vast set. The goal of this method is to improve the extraction's efficiency and accuracy and reduce the number of features involved. The goal of the proposed model is to analyze the various features of a text that are related to its sentiment. It then uses a feature extraction algorithm to find the most relevant ones, and it trains a machine learning algorithm that can classify the text into neutral, positive, or negative. Automated sentiment extraction can be performed with the help of feature selection, which can lead to faster processing times and improve the accuracy of the analysis. It can also reduce the noise in the text and provide a more accurate analysis. In addition, by reducing the number of features, the extraction process' computational resources are saved. The proposed model can improve the efficiency and accuracy of sentiment analysis by identifying the most relevant elements from the text. It can also reduce the number of features that are used in the extraction process and save time and resources.

Further, in the next section of this work, the obtained results are discussed.

IV. Results and Discussions

After the detailed discussions on the proposed method and the algorithm driven framework, in this section of the work, the obtained results are discussed.

The initial dataset analytics are visualized graphically here [Fig – 2 to 4].

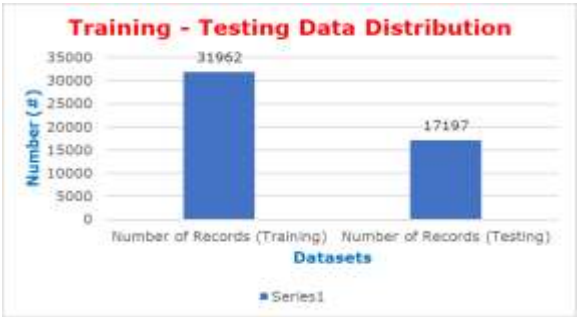

Fig. 2. Dataset Distribution

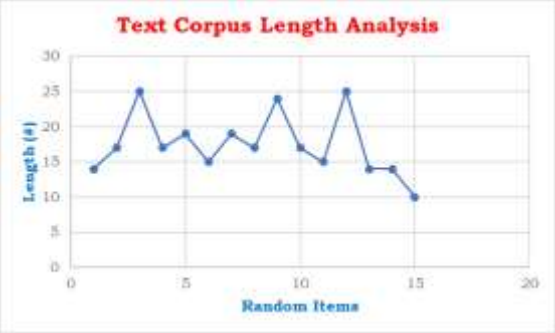

Fig. 3. Text Corpus Length Analysis

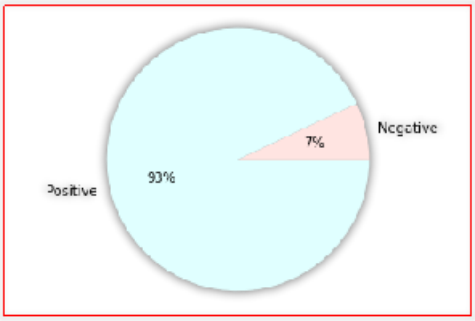

Fig. 4. Positive and Negative Distribution

Further processes are performed on the dataset more than 1200 times, however only 15 random sequences are furnished for presentation purposes.

Secondly, the three primary methods, BoW (Bag of Words), TF-IDF (Term Frequency and Inverse Document Frequency) and WE (Word Embedding) are applied on the initial dataset and the key term extraction process is identified. The outcomes are visualized graphically here [Fig – 5].

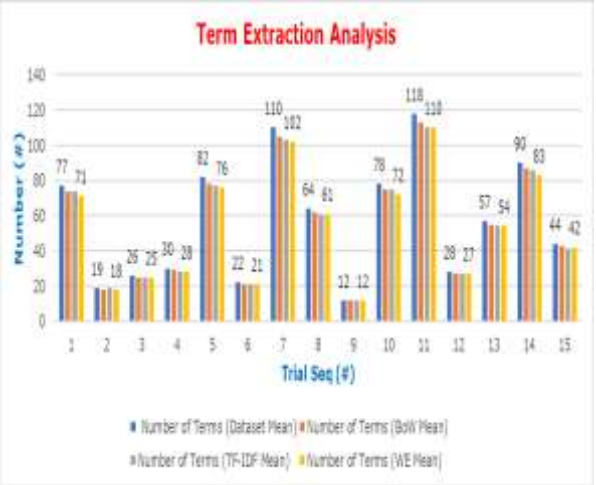

Fig. 5. Term Extraction Outcomes

It is natural to realize that there are significant overlaps during the term extraction process and few of the processes report a similar number of extracted terms in few trial sequences.

Hence, the term frequencies must be verified in order to confirm there are no miscalculations during the term counts. Hence, the term frequencies for a few sample elements must be verified.

Thirdly, the term frequency results from three methods are listed here [Table – 1].

TABLE I. TERM FREQUENCY

Number of Terms (BoW)	*Number of Terms (TF-IDF)*	*Number of Terms (WE)*
love, 1654	love, 1654	love, 1654
posit, 917	posit, 917	posit, 917
smile, 676	smile, 676	smile, 676
healthy, 573	healthy, 573	healthy, 573
thank, 534	thank, 534	thank, 534
fun, 463	fun, 463	fun, 463
life, 425	life, 425	life, 425
affirm, 423	affirm, 423	affirm, 423
summer, 390	summer, 390	summer, 390
model, 375	model, 375	model, 375

From the random analysis of the term frequencies, it is natural to realize that the term frequencies are identical and differences in terms of number of terms extracted are due to the dynamic frequencies set during the extraction process. Hence, the dynamic thresholds from each iteration must be analyzed.

Fourthly, the threshold value analysis is carried out. The threshold value is considered to reduce the number of terms as the term frequencies are expected to be higher than the threshold values.

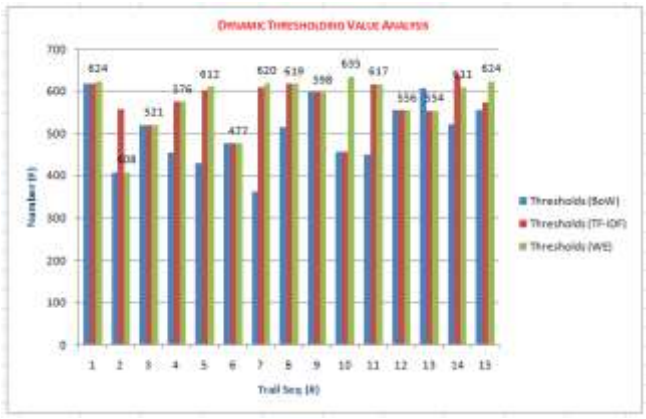

Fig. 6. Dynamic Thresholding Value Analysis

After this phase, the number of extracted terms and the threshold values are compared and it is significant to observe that similar threshold values are seen for the sequences, where the same number of terms are extracted. Hence, the previous assumption is justified.

Once the methods are validated in terms of the correctness, the introspection for hybrid method must be

validated. Hence, with the extracted terms the models must be validated for identifying the sentiments with F1 scores.

Fifthly, the F1 measures are tested for these three methods with the extracted terms.

Fig. 7. F1 Measure Analysis

The results are visualized graphically here [Fig – 8].

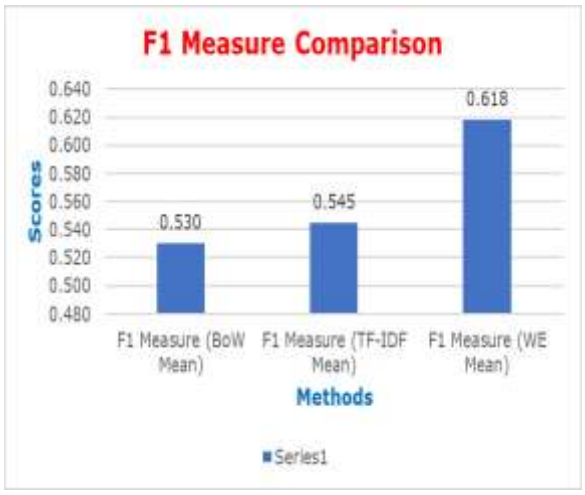

Fig. 8. F1 Measure Comparison

From the detailed and comparative results, it is natural to realize that few of the models produce better and optimal set of terms for few sequences. Hence, building the combined model for generating the final terms list as intersection of the methods is highly justified.

Sixthly, a sample of a few final terms are listed [Table – 2].

TABLE II. SAMPLE LIST OF EXTRACTED TERMS

Term	*Frequency*
posit	917
smile	676
healthi	573
thank	534
fun	463
life	425
affirm	423
posit	917

Finally, the accuracy of the three models with the hybrid model is furnished here

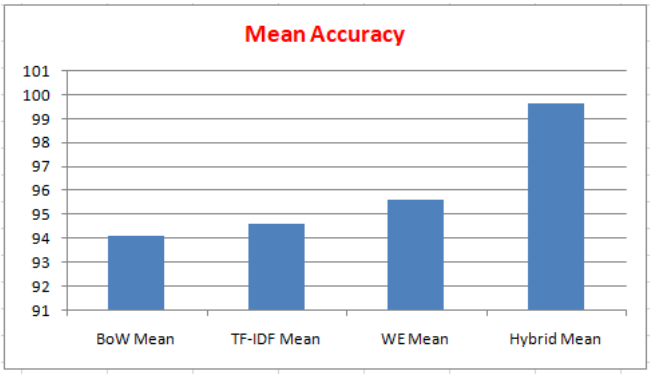

Fig. 9. Mean Accuracy

It is natural to realize that the hybrid method is performing better in terms of detection of the sentiment accurately.

Further, in the next section of this work, the work is compared with the other parallel research outcomes.

V. COMPARATIVE ANALYSIS

After analyzing the obtained results in detail, in this section of the work, the same results are compared with the other parallel research outcomes

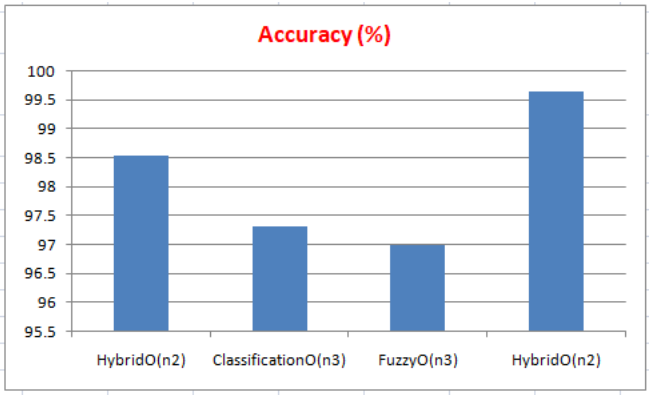

Fig. 10. Comparitive analysis

Henceforth, it is natural to realize that the proposed method outperformed the existing methods.

Finally, in the next section of this work, the research conclusion is furnished.

VI. CONCLUSION

Due to the increasing need for sentiment analysis in order to analyze customer feedback, a framework that can combine the traditional lexical features and deep learning has been

proposed. This paper will provide a framework that can be used for this type of analysis. The framework consists of three components: sentiment classification, post-processing, and feature extraction. The first one combines the features of both lexical analysis and deep learning models to provide a deeper understanding of texts' sentiments. The second component of this framework is sentiment classification, a process that is performed through a deep learning model that has hybrid features. The various components of this framework, such as Word embedding, BoW, and TF-IDF perform well in this process and produce labels for each input. Post-processing involves applying a regression method to improve the accuracy of the classification. The paper analyzed the advantages of the framework and its hybrid feature extraction method in terms of performance when it comes to sentiment analysis.

REFERENCES

[1] A. Deniz, M. Angin and P. Angin, "Evolutionary Multiobjective Feature Selection for Sentiment Analysis," in IEEE Access, vol. 9, pp. 142982-142996, 2021.

[2] M. Bibi et al., "Class Association and Attribute Relevancy Based Imputation Algorithm to Reduce Twitter Data for Optimal Sentiment Analysis," in IEEE Access, vol. 7, pp. 136535-136544, 2019.

[3] S. Bouktif, A. Fiaz and M. Awad, "Augmented Textual Features-Based Stock Market Prediction," in IEEE Access, vol. 8, pp. 40269-40282, 2020.

[4] M. Z. Ali, Ehsan-Ul-Haq, S. Rauf, K. Javed and S. Hussain, "Improving Hate Speech Detection of Urdu Tweets Using Sentiment Analysis," in IEEE Access, vol. 9, pp. 84296-84305, 2021.

[5] H. Li, Z. Ma, H. Zhu, Y. Ma and Z. Chang, "An Ensemble Classification Algorithm of Micro-Blog Sentiment Based on Feature Selection and Differential Evolution," in IEEE Access, vol. 10, pp. 70467-70475, 2022.

[6] L. Ansari, S. Ji, Q. Chen and E. Cambria, "Ensemble Hybrid Learning Methods for Automated Depression Detection," in IEEE Transactions on Computational Social Systems, vol. 10, no. 1, pp. 211-219, Feb. 2023.