DEVELOPMENT OF AN IMPROVED TECHNIQUE FOR SENTIMENT ANALYSIS USING MACHINE LEARNING

Harshdeep Singh, Kanwal Preet Singh Attwal ,Madan Lal

harshdeeps99@gmail.com,kanwalp78@yahoo.com,mlpbiuni@gmail.com

Department of Computer Science and Engineering, Punjabi University, Patiala

Abstract

Sentiment Analysis is an application that concentrates on the identification and classification of ideas indicated mainly in the form of positive, negative and neutral values. One of the most important parts of machine learning is Feature Extraction and Selection. This paper revolves around Term Frequency Inverse Document Frequency (TF-IDF) as Feature Extraction and Chi-Square (Chi) as a Feature Selection technique to generate feature vocabulary and then use it with a Hybrid Tree to bring out better results than by using machine learning algorithms. We compared the performance of Feature Selection trained using Hybrid Tree (HT) against K-Nearest Neighbor (KNN), Support Vector Machines (SVM), KNN+SVM, Sequential Minimal Optimization (SMO), Decision Tree (DT) and SMO+DT. Our proposed work with TFIDF+HT is far more accurate as compared to other ML algorithms.

Keywords – Feature Selection, TF-IDF, Chi-Square, Sentiment Analysis, Machine Learning

DOI Number: 10.48047/NQ.2022.20.20.NQ109140 **NeuroQuantology2022;20(20): 1368-1378**

1. Introduction

Sentiment Analysis is one of the most popular and used applications of machine learning. Information gathered includes opinions of the users on products, movies, music, disease, politics, tourism spots,etc [1-2]. It helps in knowing what people think and where the market is heading. Almost all industries are using Sentiment Analysis to understand their customers or what people think about their services or products [3]. Opinion Mining is mainly needed with a large amount of text data. Text Mining is used to process, organize and analyze a large amount of unstructured textual data. Classification is one important part of Text Mining with Machine and Deep Learning [4-5]. There are many algorithms used for the classification of textual data like Support Vector Machines (SVM), K-Nearest Neighbors (KNN) [6], Random Forest [7], Naïve Bayes, Logistic Regression etc. [8]. A large amount of data is gathered from applications like Twitter, Facebook, IMDB, Amazon, Flipkart etc., but mostly in unstructured format [9]. The problem with most of text classification methods is that there isa large number of features and attributes on which they have to work. This is where Feature Selection comes in place which is used to eliminate any irrelevant features and attributes. One of the most popular Feature Selection methods is Chi-Square or Chi that we have used in our proposed work. Along with Feature Selection, Feature Extraction is another method that betters the accuracy of the solution. We have used the Terms Frequency Inverse Document Frequency (TF-IDF) [10] which is fast and reliable. The purpose of using the Feature Selection and Feature Extraction method along with Classification is to improve the accuracy of the solution.

1.1. Background

1.1.1. Machine Learning

Machine Learning is a technique to program computers to optimize or better their performance based on history or experience. It comes under the branch of Artificial

eISSN 1303-5150

Intelligence (AI) and uses data and algorithms to behave in the way humans learn, automatically increasing the accuracy. Machine Learning is classified into four main categories i.e.

(a) Supervised Learning–Every observation present in the dataset is labeled with prediction is mainly derived from the input data.
(b) Unsupervised Learning – Every observation present in the dataset is unlabeled. Here algorithms learn to build the structure using input data [11].
(c) Reinforcement Learning – It is based on the interaction with the specific environment on their own.
(d) Semi-Supervised Learning - Here some part of the dataset is labeled, while most of it is unlabeled. So a collaboration of supervised and unsupervised is used.

1.1.2. Sentiment Analysis

Sentiment Analysis is the process of classifying the emotion in the text i.e. whether the text is positive, negative, or neutral [12]. There isa large number of applications related to sentiment analysis that are used in different businesses and industries like Finance, Ecommerce, Health, Tourism, Sports, Movies, Music, etc [13-16]. It is also used heavily in making predictions in Protests, Politics and Sports [17-19].It helps in increasing business productivity by understanding clients' feedback and how they are reacting to the business object which can be a movie, some product in an e-commerce application, a song etc. It not only pays attention on polarity i.e. positive, negative, or neutral [20] but also to emotions like happiness, sadness, funny, angeretc [21].

1.1.3. TF-IDF

It is an unsupervised algorithm used for feature extraction. TF-IDF works at the lexical level. It mainly deploys based onthe frequency of words in the text [22]. TF stands for Term frequency which means how many times a word has been in the document. The calculation formula is as below:

tf(w,d) = count of w in d / number of words in d (1)

Inverse Document Frequency (IDF) highlights and gives importance to words that are less occurred in the document. It is calculated as:

idf(w) = log(N/ df(w))(2)

where N is the total number of documents, and df(w) is the total number of documents with the word or term "w" in them.

TF-IDF score is calculated as:

TF-IDF(w) = TF(w) x IDF(w) (3)

1.1.4. Chi-Square (Chi)

Chi performs the measurement of relationships among variables. Variables with a value of zero mean they are independent [24]. Variables with higher $\chi 2$ are tightly coupled with each other. The

eISSN 1303-5150

sentiment Analysis application uses χ2 as the feature selection method. The relationship is measured between the feature and target variables. If the outcome is higher, the feature is then having higher importance. χ2 is among the most used and popular feature selection methods in sentiment analysis applications. The calculation formula is as below:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

where

χ2 = chi square

O_i = Observed frequency

E_i = Expected frequency

∑ = Summation

2. Related Work

Researchers [23] used Machine Learning along with optimization techniques to classify the tweets into different classes. Around 90 percent accuracy was found with Machine Learning integrated with Optimization Techniques for Sentiment Analysis. Machine Learning classification algorithms like SVM and Naïve Bayes were used for Sentiment Analysis on product reviewing datasets that outperformed the features crafted by humans. Mining is performed for product analysis by using characteristic word extraction, and word re-occurrence. Opinion Mining was performed on reviews collected from clients [24-25]. A revision of the Categorical Proportion Difference (CPD), and Categorical Probability Proportion Difference (CPPD) is presented for sentiment analysis [26]. Fisher's formula of discriminant ratio was performed for feature selection in sentiment classification [27]. Performance is improved for teacher assessment when the feature selection method is used [28].The chi-square feature selection method is used along with machine learning algorithms to better the performance of classification. A two-stage feature selection process is presented to classify sentiments with the first stage focuses on generating feature scores using train data and in the second stage, weight is assigned based on test data [29]. An upgraded Chi-Square Feature Selection technique is presented based on Chi-Square to perform Arabic Text Classification [30]. Researcher [31] analyses the outcomes of pre-processing and machine learning to classify Arabic Text. A growing ensemble learning method is presented to reduce the problem related to domain adaptation [32]. Another Memetic Feature Selection method is presented based on frequency difference for textcategorization [33].A Multivariate filter technique i.e. Multivariate Relative Discriminative Criterion (MRDC) is presented based on Relative Discriminative Criterion (RDC) along with association among features [34]. In recent years, Feature Fusion [35] is also becoming popular to classify sentiments. A true feature set is created with the fusion of word embedding and sentiment/statistical knowledge. Researchers proposed a fusion

eISSN 1303-5150

and category-type dictionary learning model which can be used for multi-view human action recognition [36].

3. Methodology

The proposed work consists of the following steps as shown in fig. 1.

1. Data Collection
2. Data Preprocessing
3. Feature Extraction and Selection
4. Classification using Machine Learning
5. Performance Evaluation

3.1. Data collection

The first stage involves a collection of data where data is collected from a Twitter source with the help of the Application Program Interface. It allows the user and source to communicate. The airline dataset is used as an input which contains multiple attributes tweet_userid, the sentiment of the tweet, retweet, and time posted etc. [37].

3.2. Data Preprocessing

Preprocessing is essential to remove noise and insufficient, inconsistent data from datasets for better accuracy. Tweets consist of many unusual special characters (%,$,*,&), hash (#Support), emojis, and URLs removed as it is irrelevant and noisy data. Table 1 and 2 shows the data set before and after preprocessing.

Table 1: Noisy Data

User_id	Sentiment	Tweet
570267956648792000	positive	@VirginAmerica what would be amazingly awesome??
570258822297579000	neutral	You will be making BOS>LAS nonstop permanently anytime soon?
569986348041547000	negative	Hi! i'm so excited about your $99 LGA->DAL-deal-but I've
569986348041547000	positive	I'm #elevategold for a reason: you rock!!

Table 2: Preprocessed data

User_id	Sentiment	Tweet
459216672264645455	positive	Awesome. I flew yall Sat morning. Any way we can correct my bill
459423647592124515	neutral	What happened to Doom
957878554946519923	negative	Cant bring up my reservation online using flight booking problems
65121651655649495	positive	You are the best whenever use any other airline I am delayed

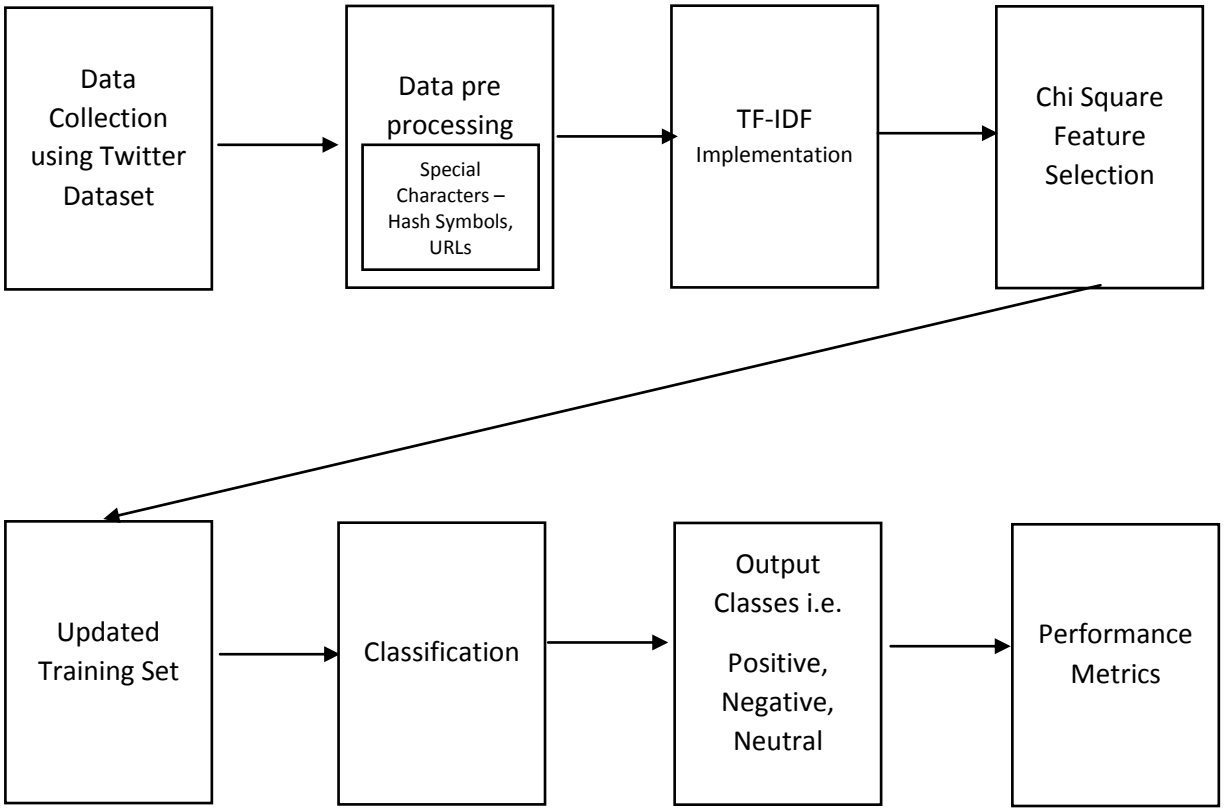
3.3. **Feature Extraction and Selection:** The third stage consists of Feature Extraction and Feature Selection where Feature Extraction is performed using TF-IDF at the lexical level and finds how times a word has been in the document and then it assigns importance to the word which occurred less in the document. For Feature Selection Process, we have used the Chi-Square method that does the measurement of the relationships among variables.

3.4. **Classification using Machine Learning:** In stage 4, we train the data using machine

eISSN 1303-5150

learning classification methods, where we used Hybrid Tree along with the data we have after TF-IDF and Chi-Square process. A hybrid Tree is a combination of two or more algorithms and we have used Random Forest and XGBoost. Random Forest is a simple algorithm than XGBoost, but both algorithms use the same model representation and inference. XGBoost permits the models to be trained in a manner that reprocesses and controls the computational efficiencies that help in training random forest models and provides better results than using individual algorithms.

3.5. **Performance Evaluation:** Performance evaluation is performed in the last stage where a comparison has been made between recent work Figure below lists the different stages and steps involved in stages:



4. Experimental Results

This section mainly revolves around the experimentation and results where we analyze the performance of the Feature Selection Technique with a Hybrid Tree. We have used the [37] dataset for our experimentation work. Feature Extraction is performed using the TF-IDF method, and Feature Selection is performed using the Chi-Square method. Features that are selected using Chi-Square are trained using a Hybrid Tree.

To assess how Feature Selection algorithms have performed along with Hybrid Tree, Accuracy, F1, Precision, and Recall are used as the metrics. Accuracy is the correctly specified samples against the total number of samples. It can be defined as below:

Accuracy Score = (TP + TN)/ (TP + FN + TN + FP) (4)

where TP = True Positives

FP = False Positives

TN = True Negatives

FN = False Negatives

Accuracy doesn't use False Positives, therefore is ineffective in some conditions or classifications, in that case, F1-Score can be used, which is defined as:

F1 Score = 2* Precision Score * Recall Score/ (Precision Score + Recall Score) (5)

Where Precision is the ratio of TP to the number of TP and FP. The recall is the ratio of TP to the number of TP and FN. Both Precision and Recall methods are given below:

Precision Score = TP / (FP + TP) (6)

Recall Score = TP / (FN + TP) (7)

Accuracy Scores of Feature Selection Technique along with Hybrid Tree on dataset is provided along with the comparison with machine learning only results [23].

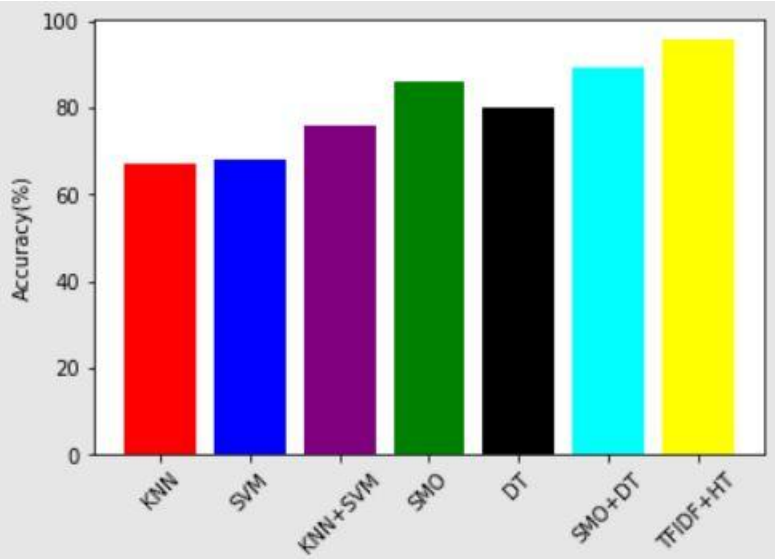


Figure – TFIDF + HT Accuracy Score against ML Algorithms

eISSN 1303-5150

As Accuracy does not uses False Positives and is ineffective sometimes, so we have also used F1-Score to compare our proposed work (TF-IDF + Chi-Square + Hybrid Tree) with ML algorithms. To get F1-Score, we need to find Precision and Recall Scores also. Figures below show Precision, Recall, and F1-Scores:

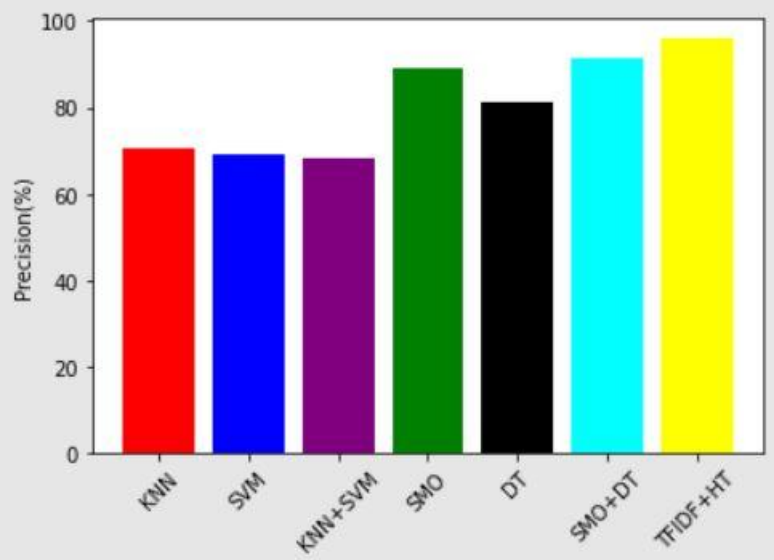


Figure – TFIDF +HT Precision Score against ML Algorithms

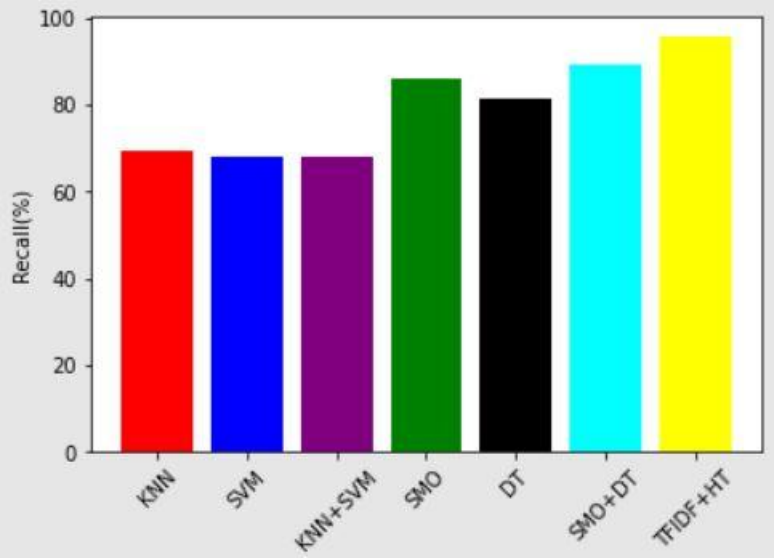


Figure - TFIDF +HT Recall Score against ML Algorithms

Now that we have our Precision and Recall Scores, we can get the F1-Score as shown below:

eISSN 1303-5150

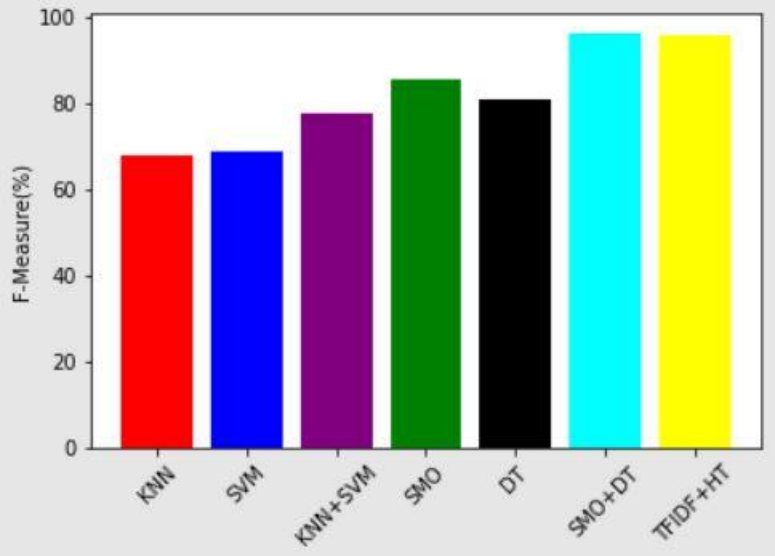


Figure - TFIDF +HT F1 Score against ML Algorithms

As results are showing, our proposed work where we have used TF-IDF for feature extraction and Chi-Square for feature selection along with Hybrid Tree has performed better than KNN, SVM, KNN+SVM, SMO, DT, SMO+DT.

5. Conclusion and Future Scope

With the results that we got from experimentation, it is clear that Feature Selection techniques help in better the results of Sentiment Analysis. TF-IDF and Chi-Square bring out superior results as compared to only machine learning algorithms. TF-IDF and Chi-Square along with Hybrid Tree showed more accuracy than KNN, SVM, KNN+SVM, SMO, DT, and SMO+DT. TFIDF+HT shows the highest accuracy with 96 percent which is much better than other algorithms used in experimentation. Also, by integrating Feature Selection with Hybrid Forest not only betters the accuracy but also it lessens the training time and parameters that automatically reduces the requirement of complex hyperparameter tuning. In the future, more Feature Selection and Machine Learning combinations should be experimented with to get a view of how other algorithm integrations behave and what impact they make on Sentiment Analysis outcomes.

References

[1] Rane A, Kumar A (2018) Sentiment classification system of twitter data for US airline service analysis. In: 2018 IEEE 42nd annual computer software and applications conference (COMPSAC). IEEE, vol 1, pp 769–773

[2] Sarirete, A. (2022). Sentiment analysis tracking of COVID-19 vaccine through tweets. *Journal of Ambient Intelligence and Humanized Computing*, 1-9.

[3] Falasari, A., & Muslim, M. A. (2022). Optimize naïve bayes classifier using chi square and term frequency inverse document frequency for amazon review sentiment analysis. *Journal of Soft Computing Exploration*, *3*(1), 31-36.

[4] Akpatsa, S. K., Li, X., & Lei, H. (2021, July). A survey and future perspectives of hybrid deep learning models for text classification.

eISSN 1303-5150

In *International Conference on Artificial Intelligence and Security* (pp. 358-369). Springer, Cham.

[5] Yaremenko, V. S., Rogoza, W. S., &Spitkovskyi, V. I. (2021). Application of neural network algorithms and naive bayes for text classification. *Journal of Theoretical and Applied Information Technology*, *99*(1), 125-134.

[6] Prasanti, A. A., Fauzi, M. A., &Furqon, M. T. (2018). Neighbor Weighted K-Nearest Neighbor for Sambat Online Classification. *Indonesian Journal of Electrical Engineering and Computer Science*, *12*(1), 150-160.

[7] Salles, T., Gonçalves, M., Rodrigues, V., & Rocha, L. (2018). Improving random forests by neighborhood projection for effective text classification. *Information Systems*, *77*, 1-21.

[8]SelvaBirunda, S., &Kanniga Devi, R. (2021). A review on word embedding techniques for text classification. *Innovative Data Communication Technologies and Application*, 267-281.

[9] Wang L, Niu J, Yu S (2019) SentiDiff: Combining textual information and sentiment diffusion patterns for twitter sentiment analysis. IEEE Trans Knowl Data Eng.

[10] Patil, R. S., &Kolhe, S. R. (2022). Supervised classifiers with TF-IDF features for sentiment analysis of Marathi tweets. *Social Network Analysis and Mining*, *12*(1), 1-16.

[11] Bibi, M., Abbasi, W. A., Aziz, W., Khalil, S., Uddin, M., Iwendi, C., &Gadekallu, T. R. (2022). A novel unsupervised ensemble framework using concept-based linguistic methods and machine learning for twitter sentiment analysis. *Pattern Recognition Letters*, *158*, 80-86.

[12] Monika, P., Kulkarni, C., Harish Kumar, N., Shruthi, S., & Vani, V. (2022). Machine learning approaches for sentiment analysis: A survey. *International Journal of Health Sciences*, *6*, 1286-1300.

[13] Sunitha, D., Patra, R. K., Babu, N. V., Suresh, A., & Gupta, S. C. (2022). Twitter sentiment analysis using ensemble based deep learning model towards COVID-19 in India and European countries. *Pattern Recognition Letters*, *158*, 164-170.

[14] Sachar, D., Goyal, H., & Sharma, V. K. (2022). Comparative Sentiment Analysis on Stock Market News Using Machine Learning. In *Intelligent Computing Techniques for Smart Energy Systems* (pp. 199-210). Springer, Singapore.

[15] Nithya, T., Atchaya, S., Dharshini, M., Harinesha, D., &Monisha, M. (2022). SENTIMENT ANALYSIS OF CUSTOMER REVIEWS USING MACHINE LEARNING TECHNIQUES. *International Journal of Advanced Engineering Science and Information Technology*, *10*(6).

[16] Wang, J., Du, J., Shao, Y., & Li, A. (2022). Sentiment Analysis of Online Travel Reviews Based on Capsule Network and Sentiment Lexicon. *arXiv preprint arXiv:2206.02160*.

[17] Mamun, M., Rahaman, M., Sharif, O., & Hoque, M. M. (2022). Classification of textual sentiment using ensemble technique. *SN Computer Science*, *3*(1), 1-13.

[18] Mee, A., Homapour, E., Chiclana, F., & Engel, O. (2021). Sentiment analysis using TF–IDF weighting of UK MPs' tweets on Brexit. *Knowledge-Based Systems*, *228*, 107238.

[19] Neogi, A. S., Garg, K. A., Mishra, R. K., & Dwivedi, Y. K. (2021). Sentiment analysis and classification of Indian farmers' protest using twitter data. *International Journal of*

Information Management Data Insights, *1*(2), 100019.

[20] Srivastava, R., Bharti, P. K., & Verma, P. (2022). A review on multipolarity in sentiment analysis. *Information and communication technology for competitive strategies (ICTCS 2020)*, 163-172.

[21] Srivastava, R., Bharti, P. K., & Verma, P. (2022). Comparative Analysis of Lexicon and Machine Learning Approach for Sentiment Analysis. *International Journal of Advanced Computer Science and Applications*, *13*(3).

[22] Antonio, V. D., Efendi, S., &Mawengkang, H. (2022). Sentiment analysis for covid-19 in Indonesia on Twitter with TF-IDF featured extraction and stochastic gradient descent. *International Journal of Nonlinear Analysis and Applications*, *13*(1), 1367-1373.

[23] Naresh, A., & Venkata Krishna, P. (2021). An efficient approach for sentiment analysis using machine learning algorithm. *Evolutionary Intelligence*, *14*(2), 725-731.

[24] Agarwal, B. and Mittal, N., 2012, December. Categorical probability proportion difference (CPPD): a feature selection method for sentiment classification. In Proceedings of the 2nd Workshop on Sentiment Analysis where AI meets Psychology (pp. 17-26).

[25] Pang, Bo and Lee, Lillian and Vaithyanathan, Shivakumar: Thumbs Up?: Sentiment Classification Using Machine Learning Techniques, In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Vol 10,EMNLP '02,pp 79-86 (2002)

[26] Morinaga, S., Yamanishi, K., Tateishi, K. and Fukushima, T., 2002, July. Mining product reputations on the web. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 341-349). ACM.

[27] Wang S., Li D.,Wei Y., Li H. (2009) A Feature Selection Method Based on Fishers Discriminant Ratio for Text Sentiment Classification. In: Liu W., Luo X., Wang F.L., Lei J. (eds) Web Information Systems and Mining. WISM 2009. Lecture Notes in Computer Science, vol 5854. Springer, Berlin, Heidelberg.

[28] C. Pong-Inwong and K. Kaewmak : Improved sentiment analysis for teaching evaluation using feature selection and voting ensemble learning integration, In: 2nd IEEE International Conference on Computer and Communications (ICCC), pp 1222-1225 (2016).

[29] X. Chi and T. P. Siew and E. Cambria: Adaptive two-stage feature selection for sentiment classification, In: IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp 1238-1243 (2017).

[30] Bahassine, S., Madani, A., Al-Sarem, M. and Kissi, M., 2018. Feature selection using an improved Chi-square for Arabic text classification. Journal of King Saud University-Computer and Information Sciences.

[31] Oussous, A., Lahcen, A.A. and Belfkih, S., 2019, March. Impact of Text Pre-processing and Ensemble Learning on Arabic Sentiment Analysis. In Proceedings of the 2nd International Conference on Networking, Information Systems and Security (p. 65). ACM.

[32] Lopez, M., Valdivia, A., Martnez-Comara, E., Luzn, M.V. and Herrera, F., 2019. E2SAM: Evolutionary ensemble of sentiment analysis methods for domain adaptation. Information Sciences, 480, pp.273-286.

[33] Lee, J., Yu, I., Park, J. and Kim, D.W., 2019. Memetic feature selection for multilabel text categorization using label frequency

difference. Information Sciences, 485, pp.263-280.

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

[35] Abdi, A., Shamsuddin, S.M., Hasan, S. and Piran, J., 2019. Deep learning-based sentiment classification of evaluative text based on Multi-feature fusion. Information Processing and Management, 56(4), pp.1245-1259.

[36] Gao, Z., Xuan, H.Z., Zhang, H.,Wan, S. and Choo, K.K.R., 2019. Adaptive fusion and category-level dictionary learning model for multi-view human action recognition. IEEE Internet of Things Journal.

[37] F. Eight, "Twitter US Airline Sentiment," 2019. [Online]. Available: https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment?resource=download.

eISSN 1303-5150