Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Applying EMONO Variants to Multi-Class Sentiment Analysis for Short-Distance Inter-Class Frequency of Term

[1] Cristopher C. Abalorio, [2] Ariel M. Sison, [3] Ruji P. Medina, [4] Gleen A. Dalaorao

[1][3] Graduate Programs, Technological Institute of the Philippines – Quezon City, [2] School of Computer Studies, Emilio Aguinaldo College, [4] College of Computing and Information Sciences, Caraga State University

[1] cristopher.abalorio@aclcbutuan.edu.ph, [2] ariel.sison@eac.edu.ph, [3] ruji.medina@tip.edu.ph, [4] gadalaorao@carsu.edu.ph

Article Info
Page Number: 1938-1947
Publication Issue:
Vol. 71 No. 4 (2022)

Article History
Article Received: 25 March 2022
Revised: 30 April 2022
Accepted: 15 June 2022
Publication: 19 August 2022

Abstract— Opinion mining has become one of the most sought-after interests by researchers because of the advent of the internet and relevant technologies. Analyzing people's opinions and emotions or sentiment analysis is a subdomain of text classification under NLP. Feature vectorizer is a technique in sentiment analysis frequently used in machine learning approaches to improve classification performance. However, working with multiple categories of sentiments becomes challenging in the machine learning approach using TF-IDF vectorizer as a word tends to be spread out in various classes. In this paper, EMONO, a supervised feature vectorizer with variants TF and SRTF, was implemented to answer the problems in the appropriate representation of terms in multiple sentiments due to a term's short-distance frequency. Results showed for Sentiment Analysis that an EMO value of 3 obtained 74% for KNN and 82% for SVM using EMONO variants compared with 69% of KNN and 81% for SVM in TF-IDF, respectively, on the Commodity News (Gold) dataset. It is evident that EMO that sets extensions of inter-classes coverage in max-occurrence values in EMONO vectorizer improves the classification performance in sentiment analysis with multi-classes.

Index Terms- EMONO Variants, Sentiment Analysis, Short-distance inter-class, TF-IDF

I. INTRODUCTION

Sentiment analysis, often called opinion mining is one of the subdomains of Text Classification (TC). A widely studied area where the polarity of the sentiments within the natural language is investigated through different approaches to improve customer service, discover trends in the market, find problems, understand the target audience, etc. Internet and developing technology promote the dramatically increased volume of data, primarily through social media [1].

Feature vectors play a significant role in sentiment analysis as it affects the classification results. The famous feature vectorizer comes with TF-IDF [2]–[5]. TF-IDF is an unsupervised method of weighting terms to extract, select, and assign weight to terms in sentiment annotated text in binary and multiclass classification. However, in literature, supervised term weighting

schemes outperformed the unsupervised approach in text classification and are favored in TC [6].

Sentiment analysis with multiclass is more challenging than binary classification with only positive and negative sentiments [7]. The text is spread out in different emotions in a multiclass, and identifying the category becomes challenging using an unsupervised feature vectorizer. In a supervised feature vectorizer, the utilization of the inter-class information improves the classification [8],[9]. However, in multiclass sentiment analysis where terms are statistically counted, the words or features tend to occur in many sentiments, e.g., the term 'gold, price, and increase' appears multiple times in positive, negative, neutral, and irrelevant classes where inter-class distance information is required [10].

This paper aims to address the short-distance frequency of a term in inter classes for sentiment analysis using supervised feature vectorization techniques TF-EMONO and SRTF-EMONO [11]. Short-distance frequency of a term occurs when the total document frequency value is more than 50 percent in the succeeding sorted category.

The succeeding sections of this paper are ordered as follows. In Section 2, the literature review contains two subsections discussing the related works in sentiment analysis and multi-class text classification. In Section 2, the methods used to conduct this study. Section 4 discusses the results of the text classification performance with its success measures, and finally, in Section 6, the conclusion and recommendation is presented.

II. LITERATURE REVIEW

This section presents the relevant works of literature for multiclass sentiment analysis. It is divided into sub-sections. This first section is about sentiment analysis and the second section is about multiclass text classification

A. Sentiment Analysis

Investigating people's emotions through analyzing text using computational linguistics is a data mining technique called sentiment analysis. Machine learning and lexicon-based approach are the two main strategies used to obtain sentiments from tweets and reviews with their given polarity positive, negative, and neutral [12]. The lexicon-based approach [13]–[15] has been successfully used in many applications, such as developing a cross-lingual Malay language using Twitter data [16]; sentiment analysis research focuses commonly on the English language lexicon. However, the machine learning approach [17]–[19] dominates the applications of sentiment analysis through the use of learning algorithms such as SVM, KNN, Random Forest, and DT [20]–[22].

The opinion mining techniques fall under text classification, commonly evaluated by Precision, Recall, Accuracy, and F-measure. Extraction of terms of features in natural language text is one of the core components of text analysis. In order to proceed, terms should be modeled as feature vectors and will be assigned weights according to their relevance within the entire document. A feature vectorizer often used in sentiment analysis is the TF-IDF. Since then, TF-IDF has been a machine learning tandem to vectorize the text documents, especially with

annotated sentiments to classify the documents and automatically predict the new text's feelings or opinions.

B. Multi-class Text Classification

Binary text classification has been commonly used in sentiment analysis with positive and negative polarities. However, nowadays, people's feelings are explicitly expressed in product reviews and tweets. It has now added neutral and other sentiments, adding multiple categories for sentiment analysis as subdomains in text classification.

In Multi-class text classification using a feature vectorizer, there are two primarily used methods supervised and unsupervised extraction, selection, and assigning weights to terms. In literature, the unsupervised way used TF-IDF because it does not consider the information of the sentiments or class in setting weights to terms. Meanwhile, supervised methods include sentiment information in adding weights to texts [8],[9]. The class information, such as the inter-class values, refers to the amount of occurrence of a term in every sentiment. Since the multi-class dataset is often imbalanced, the number of inter-class values fluctuates in every term. Reviews and literature enforce that the supervised way is the suggested approach as a feature vectorizer in using the machine learning approach. It outperforms the unsupervised vectorizer in classifying performance [6],[23],[11].

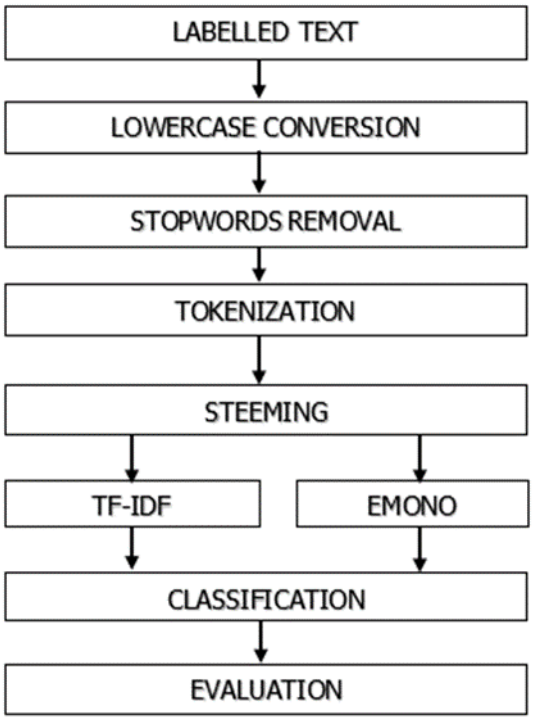

Fig. 1. Framework of the Study

III. METHODOLOGY

This section presents the methods and techniques of how the study is conducted. The framework is shown in Fig.1 as the overall workflow in addressing the short-distance inter-class frequency of a term in multi-class sentiment analysis.

A. Dataset

The benchmark dataset used in this paper is the Commodity News (Gold), a news dataset for the commodity market that manually annotated 11,412 news headlines across multiple dimensions into various sentiments. It is one of the most used imbalanced datasets favored for sentiment analysis. The data is split into 70% for training and 30% for test data. There are a total of 14,412 news documents with four classes as sentiments: positive (4759), negative (4138), neutral (448), and irrelevant (2067).

B. Pre-processing

The following are the pre-processing steps to prepare the dataset in this study:

1. Each text in the corpus was transformed into lowercase characters. The method was used to uniformly treat all text equally because the term "Gold" was not treated with "gold."

2. The stop-words such as "the," "an," "or," etc., were removed from the text documents. These words were the common words that exist in any language. The primary purpose of omitting stop words is to obtain more relevant and significant terms in the text corpus. The number of essential terms represents the text category.

3. In the feature, vectorization text is tokenized to get the single terms to assign individual weight values. There are two tokenization performed within data pre-processing. The first is eliminating special characters within the terms through a regular expression. Next is to obtain single terms.

4. Stemming was implemented with the tokenized texts to treat terms with similar meanings uniformly. The Porter stemming is preferred to use in this paper. This was done to reduce every term and transform it into its root word.

5. The terms that seldom occur were omitted. If the words appear more than once, they are retained as features.

C. TF-IDF

The Term Frequency Inverse Document Frequency is a method used to compute the score of every term signifying the term's relevance within the documents and the corpus. It is an unsupervised feature vectorizer because it does not consider class information in setting weights to terms. Class information refers to the information about the categories belongingness to every document, such as the number of classes pertaining to certain documents and the like. To generate the features using TF-IDF, Term Frequency in (1) is multiplied by Inverse Document Frequency in (2), as shown in (3).

$$TF(t) = \frac{(No.of\ times\ term\ t\ appears\ in\ document)}{(Total\ no.of\ terms\ in\ the\ document)} \quad (1)$$

$$IDF(t) = log\left(\frac{Total\ no.of\ documents}{No.of\ documents\ where\ term\ t\ appears}\right) \quad (2)$$

$$TF - IDF = TF(t) * IDF(t) \quad (3)$$

D. TF-EMONO and SRTF-EMONO Feature Vectorizer

The Extended MONO with Term Frequency and Squared-root of Term Frequency variants is a supervised feature vectorizer used in text classification. EMO parameter holds value for classes covered in max-occurrence computation. The j classes in (4) are the sorted document frequencies of a term. Let us assume that the EMO parameter holds the value of 2 where $1 < EMO < d_{ij-1}$.The max-occurrence is computed in (5) while normalized non-occurrence in (6). The EMONO local and global weights for feature vectors are calculated in (7). Finally, in (8), the EMONO variants with term frequency and squared root of term frequency are computed.

$$sorted\_df_{t_i} = \left\{ \overbrace{d_{i3}, d_{i1}}^{MO} |, \overbrace{\overline{d_{i4}}, \overline{d_{i5}}, \dots, \overline{d_{ij}}, \overline{d_{ij-1}}}^{NO} \right\} \quad (4)$$

$$EMO_{t_i} = \frac{MO_{t_{i1}} + MO_{t_{i2}}}{D_{total}\left(MO_{t_{i1}} + MO_{t_{i2}}\right)} \quad (5)$$

$$Normalized\_NO_{t_i} = \frac{NO_{t_{i1}} + NO_{t_{i2}}}{200} \quad (6)$$

$$EMONO_{Local}(t_i) = EMO_{t_i} * Normalized\_NO_{t_i} \qquad EMONO_{Global}(t_i) = [1 + \alpha * EMONO_{Local}(t_i)] \quad (7)$$

$$TF - EMONO = TF(t_i, d_k) * [EMONO_{Global}(t_i)] \qquad SRTF - EMONO = SRTF(t_i, d_k) * [EMONO_{Global}(t_i)] \quad (8)$$

E. Machine Learning Algorithm

The Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) [22]–[25] are utilized to classify documents from the benchmark Commodity News (Gold). These two learning algorithms top the list of classifiers frequently used in binary or multiclass sentiment analysis classification tasks.

F. Success Measures

The metrics to evaluate this study is Accuracy and F1 scores (10). F1 score is primarily used in an imbalanced dataset derived from Precision and Recall (9).

$$\text{Precision}_{c_k} = \frac{\text{TP}_{c_k}}{\text{TP}_{c_k} + \text{FP}_{c_k}} \qquad \text{Recall}_{c_k} = \frac{\text{TP}_{c_k}}{\text{TP}_{c_k} + \text{FN}_{c_k}} \tag{9}$$

$$\text{Accuracy} = \frac{2 * \sum_k^c \text{TP}_{c_k}}{2 * \sum_k^c \text{TP}_{c_k} + \sum_k^c \text{FP}_{c_k} + \sum_k^c \text{FN}_{c_k}} \qquad \text{F1}_{c_k} = \frac{2 * \text{Precision}_{c_k} * \text{Recall}_{c_k}}{\text{Precision}_{c_k} + \text{Recall}_{c_k}} \tag{10}$$

IV. RESULTS AND DISCUSSION

The classification results of the multi-class sentiment analysis were examined in this section, along with the utilization of the learning algorithms SVM and KNN to address inter-class distance in document frequency of a term.

Fig. 2 shows the unique features extracted from Commodity News (Gold) sentiment analysis dataset. The features were filtered and selected after data pre-processing was performed. Of the total 14,412 news documents, only 3,203 features were generated and used for this study. The generated features displayed in Word Cloud are taken from the maximum p-values using chi-squared.

Table I shows the top ten inter-class document frequency values that statistically counted from the total number of documents in the term. The values will determine the number of classes or extensions for the EMO parameter of the feature vectorization using TF-EMONO and SRTF EMONO. If EMO = 2, then two categories or classes of sentiments covered the extended max-occurrence and normalized non-occurrence.

In Fig. 3, the supervised feature vectorizer outperformed the unsupervised technique. EMO value sets the extension for the EMONO variants with TF and SRTF addressing short-distance inter-classes. EMO value of 3 has the promising results of accuracy and F1 scores for both TF and SRTF variants of EMONO.

In Table II the comparative results of the classification performance for unsupervised (TF-IDF) and supervised (TF-EMONO and SRTF-EMONO) feature vectorizer techniques, along with the accuracy and F1 scores of the metric, are shown. It can be seen that the feature vectorizer technique for KNN has the lowest accuracy value of 67% for TF-IDF and the highest value of

74% for TF-EMONO with EMO = 2 and SRTF-EMONO with EMO = 3. However, SVM has the higher classification performance in terms of accuracy, with its lowest value of 81% for TF-IDF and 82% for EMONO variants in all EMO values.

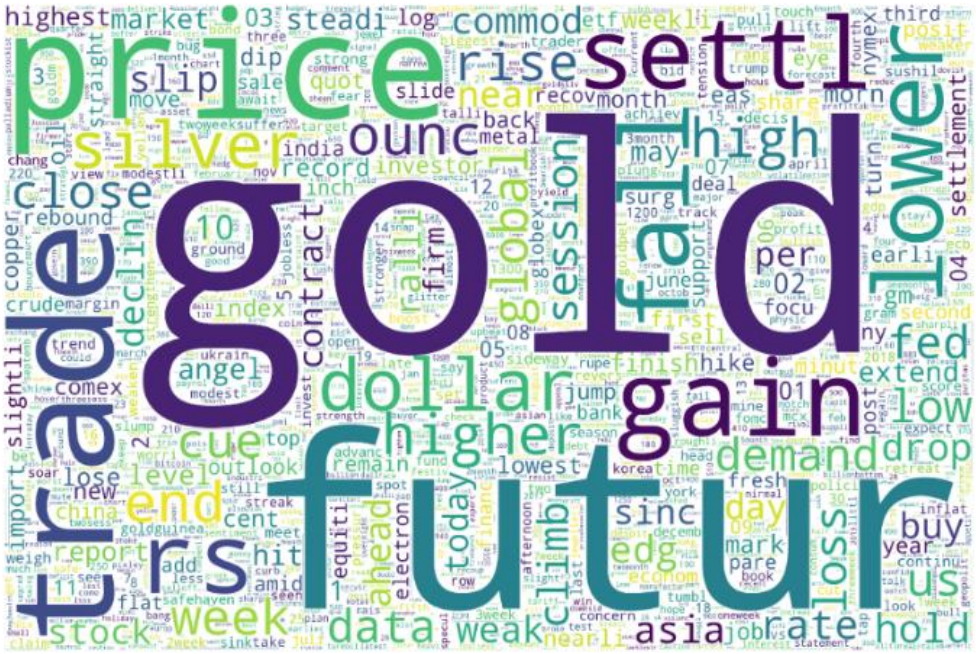

Fig. 2. Word Cloud for Extracted Features

TABLE I. TOP TEN FEATURES WITH INTER CLASS DOCUMENT FREQUENCY VALUES

Features	**Inter-class document frequency (Sentiments)**			
	Positive	**Negative**	**Neutral**	**Irrelevant**
gold	4759	4137	448	2067
futur	952	837	39	89
trade	406	376	135	125
price	653	577	127	231
gain	1074	171	23	41
settl	524	489	4	18
dollar	449	4	74	54
silver	358	333	29	159
lower	60	776	2	10
fall	76	748	4	29

TABLE II. COMPARATIVE RESULTS OF DIFFERENT FEATURE VECTORIZATION

Vectorizer	**Category**	**KNN**		**SVM**	
		Accuracy	**F1-weighted**	**Accuracy**	**F1-Weighted**
TF-IDF	Unsupervised	67%	67%	81%	81%
TF-EMONO EMO=2	Supervised	74%	73%	82%	82%
TF-EMONO EMO=3	Supervised	74%	74%	82%	82%
SRTF-EMONO EMO=2	Supervised	73%	73%	82%	82%
SRTF-EMONO EMO=3	Supervised	73%	73%	82%	82%

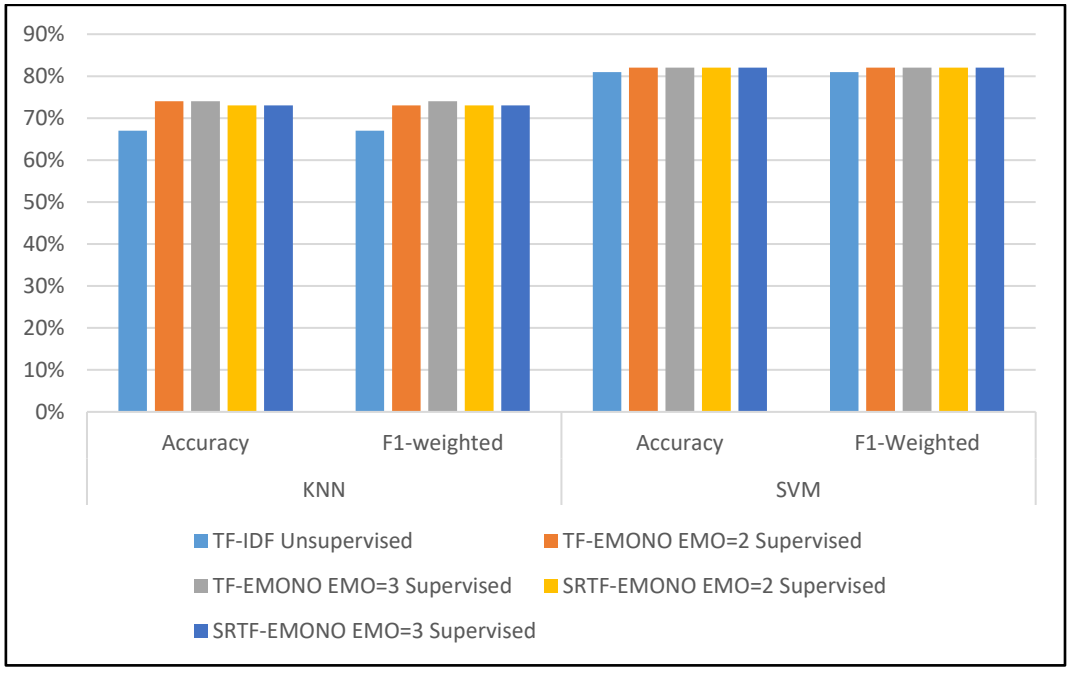

Fig. 3. Comparative result with machine learning classifier

V. CONCLUSION AND RECOMMENDATION

The inter-class information for the supervised feature vectorizer that is implemented in multi-class sentiment analysis was examined in this study. The short-distances inter-class that refers to the separation between sentiments are the values of inter-class document frequency of a term determined by the value of EMO using TF-EMONO and SRT-EMONO vectorizer. The results of the conducted experiment show that the EMO value of 3 has the highest classification accuracy value of 74% in KNN and 82% in SVM classifiers compared with 67% in KNN and 81% in SVM in TF-IDF. Moreover, in a multi-class sentiment analysis with four classes EMO value of 3 is the preferred number of extensions used for EMONO variants, which improves classification performance. It is evident that supervised (EMONO) outperformed the unsupervised (TF-IDF) feature vectorization techniques in text classification of multi-class sentiment analysis. It is recommended that the EMONO variant be applied with different feature selection strategies.

REFERENCES

[1] R. Alroobaea, "Sentiment Analysis on Amazon Product Reviews using the Recurrent Neural Network (RNN)," Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 4, pp. 314–318, 2022, doi: 10.14569/IJACSA.2022.0130437.

[2] S. K. Jones, "A statistical interpretation of term specificity and its application in retrieval," vol. 60, pp. 493–502, 2004.

[3] G. A. Dalaorao, A. M. Sison, and R. P. Medina, "Applying Modified TF-IDF with Collocation in Classifying Disaster-Related Tweets," vol. 9, no. 1, pp. 28–33, 2020.

[4] G. A. Dalaorao, A. M. Sison, and R. P. Medina, "Integrating Collocation as TF-IDF Enhancement to Improve Classification Accuracy," IEEE Access, vol. 4, pp. 282–285, 2019, doi: 10.1109/TSSA48701.2019.8985458.

[5] B. Trstenjak, S. Mikac, and D. Donko, "KNN with TF-IDF Based Framework for Text Categorization," Procedia Eng., vol. 69, pp. 1356–1364, 2014, doi: 10.1016/j.proeng.2014.03.129.

[6] A. Alsaeedi, "A survey of term weighting schemes for text classification Abdullah Alsaeedi," vol. 12, no. 2, pp. 237–254, 2020.

[7] G. Mutanov, V. Karyukin, and Z. Mamykova, "Multi-Class Sentiment Analysis of Social Media Data with Machine," 2021, doi: 10.32604/cmc.2021.017827.

[8] K. Chen, Z. Zhang, J. Long, and H. Zhang, "Turning from TF-IDF to TF-IGM for term weighting in text classification," Expert Syst. Appl., vol. 66, pp. 245–260, 2016, doi: 10.1016/j.eswa.2016.09.009.

[9] T. Dogan and A. K. Uysal, "Improved inverse gravity moment term weighting for text classification," Expert Syst. Appl., vol. 130, pp. 45–59, 2019, doi: 10.1016/j.eswa.2019.04.015.

[10] U. I. Akpana and A. Starkey, "Review of classification algorithms with changing inter-class distances," vol. 4, no. November 2020, 2021, doi: 10.1016/j.mlwa.2021.100031.

[11] C. C. Abalorio, A. M. Sison, R. P. Medina, and G. A. Dalaorao, "Extended Max-Occurrence with Normalized Non-Occurrence as MONO Term Weighting Modification to Improve Text Classification," Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 4, pp. 91–

97, 2022, doi: 10.14569/IJACSA.2022.0130411.

[12] R. Srivastava, "Comparative Analysis of Lexicon and Machine Learning Approach for Sentiment Analysis," vol. 13, no. 3, pp. 71–77, 2022.

[13] C. Chang, S. Hwang, and M. Wu, "Learning bilingual sentiment lexicon for online reviews," Electron. Commer. Res. Appl., vol. 47, no. November 2020, p. 101037, 2021, doi: 10.1016/j.elerap.2021.101037.

[14] M. E. Mowlaei, M. S. Abadeh, and H. Keshavarz, "Aspect-Based Sentiment Analysis using Adaptive Aspect-Based Lexicons," Expert Syst. Appl., p. 113234, 2020, doi: 10.1016/j.eswa.2020.113234.

[15] S. S. Sharma and G. Dutta, "SentiDraw : Using star ratings of reviews to develop domain specific sentiment lexicon for polarity determination," Inf. Process. Manag., vol. 58, no. 1, p. 102412, 2021, doi: 10.1016/j.ipm.2020.102412.

[16] N. I. Zabha, Z. Ayop, S. Anawar, E. Hamid, and Z. Z. Abidin, "Developing Cross-lingual Sentiment Analysis of Malay Twitter Data Using Lexicon-based Approach," vol. 10, no. 1, pp. 346–351, 2019.

[17] V. Govindan and V. Balakrishnan, "A machine learning approach in analysing the effect of hyperboles using negative sentiment tweets for sarcasm detection," J. King Saud Univ. - Comput. Inf. Sci., no. xxxx, 2022, doi: 10.1016/j.jksuci.2022.01.008.

[18] H. Zhao, Z. Liu, X. Yao, and Q. Yang, "A machine learning-based sentiment analysis of online product reviews with a novel term weighting and feature selection approach," Inf. Process. Manag., vol. 58, no. 5, p. 102656, 2021, doi: 10.1016/j.ipm.2021.102656.

[19] S. Mohan, A. Kumar, H. Kumar, and A. Singh, "Predicting the impact of the third wave of COVID-19 in India using hybrid statistical machine learning models : A time series forecasting and sentiment analysis approach," Comput. Biol. Med., vol. 144, no. February, p. 105354, 2022, doi: 10.1016/j.compbiomed.2022.105354.

[20] B. Andrian, T. Simanungkalit, I. Budi, and A. F. Wicaksono, "Sentiment Analysis on Customer Satisfaction of Digital Banking in Indonesia," vol. 13, no. 3, pp. 466–473, 2022.

[21] M. Al-ayyoub and A. Nuseir, "Hierarchical Classifiers for Multi-Way Sentiment Analysis of Arabic Reviews," vol. 7, no. 2, pp. 531–539, 2016.

[22] M. Ahmad, S. Aftab, M. S. Bashir, N. Hameed, I. Ali, and Z. Nawaz, "SVM Optimization for Sentiment Analysis," vol. 9, no. 4, 2018.

[23] T. Dogan and A. K. Uysal, "A novel term weighting scheme for text classification: TF-MONO," J. Informetr., vol. 14, no. 4, p. 101076, 2020, doi: 10.1016/j.knosys.2012.06.005.

[24] H. T. Sueno, B. D. Gerardo, and R. P. Medina, "Multi-class Document Classification using Support Vector Machine (SVM) Based Multi-class Document Classification using Support Vector Machine (SVM) Based on Improved Naïve Bayes Vectorization Technique," no. June, 2020, doi: 10.30534/ijatcse/2020/216932020.

[25] A. J. P. Delima, "An Enhanced K-Nearest Neighbor Predictive Model through Metaheuristic An Enhanced K - Nearest Neighbor Predictive Model through Metaheuristic Optimization," no. October, 2020, doi: 10.46604/ijeti.2020.4646.