Exploring Diverse Conventional and Deep Linguistic Features for Sentiment Analysis of Online Content

Yajun Tang*
Anhui Business and Technology College, Hefei City, Anhui Province, 231131, China

***Abstract*—Social media has changed the world by providing the facility to common person to share their views and generate their own content, known as Users Generated Content (UGC). Due to huge volume of UGC data being created at great velocity, so to analysis this big data, latest AI (Artificial Intelligence) and its sub-domain NLP (Natural Language Processing) are being used. Sentiment analysis of online content is an active research area due to its vast applications in business for review analysis, social and political issues. In this research study, we aim to carry out sentiment analysis of online content by exploring conventional features like Term Frequency – Inverse Document Frequency (TF-IDF), Count-Vectorization, and state of the art word embeddings based word2vec. Extensive exploratory data analysis has been carried out using the latest data visualization approaches. The main novelty lies in the application of unique and diverse machine learning algorithms on social media datasets and the results evaluation using standard performance evaluation measures reveal that the word2vec using Quadratic Discriminant analysis-based classifier show optimal results.**

***Keywords*—Artificial intelligence; sentiment analysis; machine learning; word embeddings; natural language programming**

I. Introduction

Opinion mining or sentiment analysis on the other hand is a highly important subfield of NLP and is used as the umbrella term for studying sentiments, opinions, and emotions in text. Due to the drastic increase in use of social networking sites and the internet, the analysis of public opinion has gained importance for various commerce, policies and academia. Text analytics include creating categories depending on whether the text is positive, negative or neutral which can be important in understanding uptake among consumers or any specific segment or the public. For sentiment classification, Extended SVM, Naïve Bayes and Logistic regression were used traditionally; yet they are highly dependent on feature engineering and could not capture the depth of human language effectively [1].

In the last few years, deep learning techniques have brought dramatic improvements in SA, because they apply neural network structures that learn multi-level representations of text data from scratch [2]. RNNs, LSTM and CNN have been found to provide better resolution in extracting sequential correlations and contextual information within textual data [3]. In addition, the rise of pre-trained language models such as BERT and GPT has enhanced the performance of the sentiment analysis systems since transfer learning reduces the rate of overfitting as well as making models better at generalizing between datasets with limited labeled data based on [1].

Sentiment analysis is not limited to the monitoring of social media such as Facebook, Instagram, and twitter [4] but can be practiced in areas such as product review analysis, the customers feedback evaluation and even in the healthcare field. Since sentiment analysis is rapidly becoming a standard method for assessing public opinion and improving customer interaction for various organizations, the issue of effective and accurate detection becomes critical. Future areas are still hard and require better solutions to decode sarcasm, different meanings in different contexts, and domain specific language which are also quite important to cover in deep learning techniques [2]. Thus, continuous study is required to overcome these challenges and equally to extend the effectiveness of sentiment analysis to help explain the multifaceted human emotions captured in the content that is generated online.

In this research study, our aim is to carry out sentiment analysis from online content of social media by exploring the role of various textual features such Count Vectorizer, and Term Frequency – Inverse Document Frequency (TF-IDF). The features are used as input to machine learning classifiers such as CalibratedClassifierCV (CACV), PassiveAggressiveClassifier (PAC), and Quadratic Discriminant Analysis (QDA). We also try to explore various deep features like word2vec which focuses on considering context for given words in local and global perspective respectively. Here local means within a sentence or few words before or after the word while global means within the whole document. The results evaluation is carried out using standard performance metrics of accuracy, precision, recall and f-measures. This paper contributes to the field of AI by carrying out machine learning analysis of human feelings and emotions derived from textual data, while contributing toward the general understanding of the relationship between textual encoding and the analysis of human behavior. The main contributions of this research study include:

- Application of an advanced feature engineering approach such as count vectorization, TF-IDF and Word2Vec embeddings proved helpful in improving model sentiment analysis performance.
- Also, other machine learning models like Calibrated Classifier CV, Passive Aggressive Classifier, and Quadratic Discriminant Analysis were used to classify the sentiment labels.
- The best results were achieved employing Word2Vec embeddings with CACV proving that the use of embedding enhances the performance of the system.

The paper is organized as follows: Section II presents an analysis of some of the prior work done regarding sentiment classification and feature extraction methods. Section III then describes the data processing methodology of this study, feature engineering techniques including TF-IDF and Word2Vec and the classification models used in this study. Section IV summarizes the findings of this work and discusses the effectiveness of the embedding techniques. Section V provides the overall conclusion of the paper.

II. Background

Therefore, over the recent past, the use of sentiment analysis has grown to be more important as the world has advanced in issues such as social media and content creation. The field has graduated from decision tree type of solutions to highly enhanced machine learning and deep learning type methods, which can now learn the tone of the text to whether it is happy, sad, angry or otherwise. The latest publications point to a revolutionary effect of generative AI for enhancing the efficacy and flexibility of SA for identifying consumer sentiment [5]. The use of sophisticated NLP ensures that there is a means of processing huge volumes of data generated by customers over social media as well as ensuring that the firms can indeed derive tangible benefits from these big data sources [6]. A recent research study [7] focused on features-oriented sentiment analysis which is also known as aspect-oriented sentiment analysis. This type of analysis mainly does not focus on document. Due to the growing era of technology, different methods of real time sentiment tracking have been enhanced to enable organizations to measure the flow of public sentiment. Programs such as Brand24 and Sprout Social employment machine learning [8] [9] to identify the sentiment of text in different and even emojis, thereby giving a fuller understanding of customer feelings [10]. Moreover, aspect-based sentiment analysis (ABSA) has become another important approach which is to identify certain characteristics of the product or service and allows companies to assess feedback from customers on specific characteristics, such as, for example, battery life or usability, separately [11]. Considering deep learning models, attention-based model haven been used in a recent study which mainly proposes multi-channel gated recurrent RNN algorithms for aspect-based sentiment classification purpose. The work proof that the proposal of multi-channels in the existing RNN model [12].

As these methodologies are progressing there are still some issues arising in sentiment analysis because of factors like sarcasm, cross cultural differences and—regarding social media—frequent changes of language [13]. These challenges have been pointed out in recent literature reviews and the community has called for more research to enhance the reliability of sentiment analysis approaches [14]. Also, the market for sentiment analysis tools is expected to expand rapidly owing to the rising need for timely analysis of the customers' sentiments and behavior [15]. Aspect-based sentiment analysis, the model with good contextual information, namely, Attention-based Bidirectional LSTM (BiLSTM) networks, is more suitable when it comes to fine-grained tasks [16]. Similarly, other recent works by [17] suggested a convolutional neural network and BiLSTM with attention mechanisms to deliver higher accuracy than conventional methods of sentiment classification in product reviews. Transformer based models have dramatically influenced approaches used in sentiment analysis. Subsequently, [18] used gradient boosting algorithms for the sentiment analysis tasks and to their finding, it outperformed other previous models for identifying the complicated sentiment patterns in large contextual data. For the sentiment classification in particular domains, for example, financial or health care domains, and have shown that the domain-wise improvement of the classifier performance is possible in this case that combine sentiment analysis with other NLP tasks [19].

Altogether, rhetoric trends dynamic, and new developments in tools and methods are expected for better application efficiency and higher predictive results of sentiment detection in different environments. This suggests that, as organizations continue to use these insights for strategic decision making [20], expanded research will be required to respond to the limitations of current methodologies and to investigate new employment contexts in this rapidly expanding domain. In this paper, instead of investigating and comparing traditional machine learning methods to sentiment analysis of textual data as previous studies have done, the current study employs advanced supervised learning models that Calibrated Classifier CV, Passive Aggressive Classifier, and Quadratic Discriminant Analysis (QDA) networks. These strategies are intended to improve the reliability and stability of sentiment predicting which was mentioned to be a weakness in the prior researches.

III. Proposed Research Methodology

The following sections provide the details of the methodological approach, as illustrated in Fig. 1, used in this sentiment analysis study, by following steps of data preprocessing, feature extraction, model training and experiment.

A. Data Preprocessing

Data cleaning is very important to ensure that pre-processing on data is well done and well checked before applying any machine learning. Initially, for noise removal, following elimination of special characters, URLs, any numbers, all the stop words, including 'is', 'the', etc. To minimize model bias, data entries with redundancy or duality were spotted and disregarded. After this, preprocessing undertaken to the text included conversion of text to lower case to eliminate redundancy concerning the sensitivity of the upper and lower 'cases. For more refinement, lemmatization was applied to stem words, where it uses the smallest root for a word to avoid any complexity in text data [21]. Lastly, to improve text vectorization in the next steps, each sentence was broken down to individual words (tokens). This ensures that data is preprocessed and ready for model training.

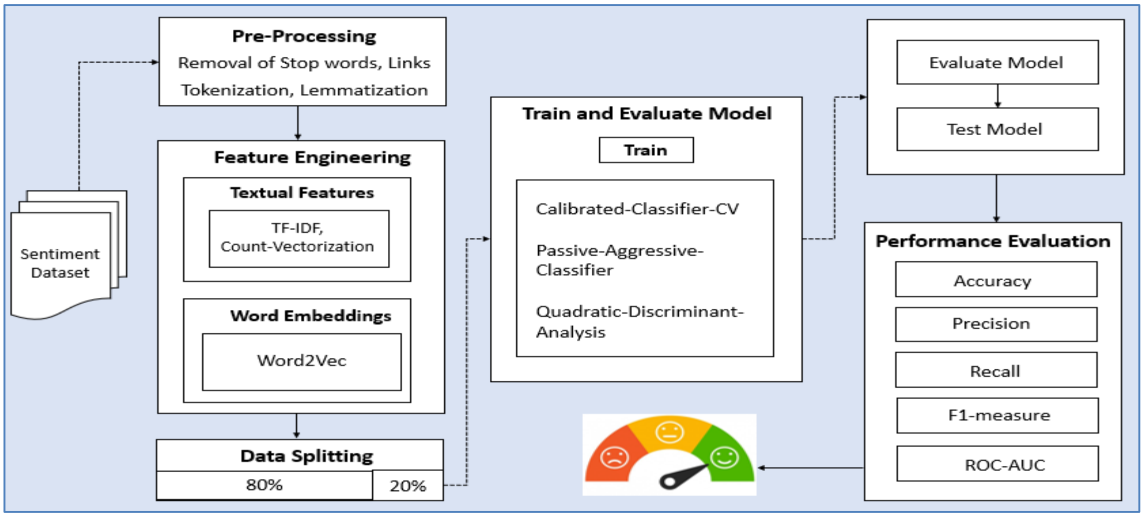


Fig. 1. Steps of proposed research methodology.

B. Feature Engineering

Feature engineering means getting the preprocessed textual data into forms directly understandable to the machine learning algorithms. In this study, three techniques were employed: Term frequency-Inverse document frequency (TF-IDF), Count Vectorization and Word2Vec. This paper provides the following overview of the mathematical basis and application of these methods.

1) TFIDF: TF-IDF refers to a technique of weighing words in a document against a corpus to determine the importance of the term in the document. It is the product of two components: Term Frequency (TF), and Inverse Document Frequency (IDF). To measure how much the term is exclusive or specific to a corresponding document. In this study, TF-IDF vectors were calculated, using Eq. (1) on the textual data and were sparse and of high dimensionality in representation of the documents [22]. Table I displays the description of symbols used in equations.

$$Document\ vector_d = [TF - IDF(t_1, d, D), \dots \dots [TF - IDF(t_n, d, D)] \quad (1)$$

2) Count vectorizer: Frequency based vectoring or word-frequency vectorization derives a numerical value for each word based on the number of times the term appears in that document relative to a fixed list of terms. Every document is then converted to vector, whose elements are the vocabulary of the subjects and the values being the frequency of each of the terms used using Eq. (2). Although noncomplex, it does an excellent job of encoding the distribution of words in the dataset, which is represented as a sparse matrix for input into the machine learning algorithms.

$$Document\ vector_d = [x_1, x_2, \dots\dots, x_n] \quad (2)$$

3) Word2Vec representation: Word2Vec gives dense words embedding in the continuum vector space to capture semantic relationship in between words, it learns word embeddings from a large corpus. These embeddings capture context meaning to make the words that have similar contexts to have similar representations. For document-level representation, generally take the average of the word vector in applying machine learning models, so it is compact as well as semantically rich.

$$Document\ vector_d = \frac{1}{n} \sum_{i=1}^{n} Word2Vec\,(t_i) \quad (3)$$

C. Model Engineering

Algorithm selection, setting, and model training on the preprocessed dataset form the model engineering process based on three algorithms were employed: CalibratedClassifierCV, PassiveAggressiveClassifier, and Quadratic Discriminant Analysis (QDA). Equations defining each and principles which underline each are provided in the following:

CalibratedClassifierCV (CACV) is a meta-algorithm aimed towards increasing the accuracy of a base classifier when using probability estimates for decision making. It functions by using the raw outputs of the classifier, to which logistic regression model ensures to the raw outputs of the classifier, that ensures a monotonic relationship between probabilities and true outcomes, computed as in Eq. (4). This algorithm comes very handy especially when the base classifier gives unformatted probabilities or raw scores.

$$p(y = 1|x) = \frac{1}{1+\exp(-(a.f(x)+b))} \quad (4)$$

PassiveAggressiveClassifier (PAC) is an online learning algorithm which is suitable for scaling and efficient classification paradigm. It adapts its model weights only when predictions are wrong or the decision margin is less than specified, which makes it reactive "aggressively or slightly" to mistakes. The model optimizes a hinge loss function that has been well applied in binary and multi-class classification and supports linear kernel-based learning, computed objective function as in Eq. (5). The model is especially useful for the cases of working with high dimensions and big data, like text classification tasks, at which it balances the speed of adaptation to new data and necessary computational resources.

$$L(w,x,y) = \max(0,1 - y(w.x)) \quad (5)$$

The model updates w iteratively as in Eq. (6):

$$w_{t+1} = w_t + \tau y x \quad (6)$$

Where $\tau = \frac{1-(w_t.x)}{||x||^2}$ is the learning rate to ensure convergence while remaining sample of passive for correctly classified.

Quadratic Discriminant Analysis (QDA) is another generative classification algorithm, which implies that the model assumes features are normally distributed within the classes. It is an extension to Linear Discriminant Analysis (LDA) where covariances within each class may differ and therefore produces quadratic decision boundaries. The current implementation of QDA is based on the Bayes' theorem, where the likelihood of each class is the multivariate Gaussian probability density, as in Eq. (7). Unlike other machine learning algorithms which may not be well suited when dealing with non-linear feature-class space. In general, QDA is more complex than LDA, but it is more flexible; thus, it is preferable when the classes have different variance.

$$Q(d=k|g) = \frac{Q(g|d=k)Q(d=k)}{Q(g)} \quad (7)$$

This is subjected to Eq. (8):

$$Q(g|d=k) = \frac{1}{(2\lambda)^{e/2}|\sum k|^{1/2}} \exp(-\frac{1}{2}\,(g-\mu_k)^T \textstyle\sum_K^{-1}(g-\mu_k) \quad (8)$$

The decision boundary for QDA is quadratic, computed as (9).

$$\delta_k(g) = -\frac{1}{2}\ln|\textstyle\sum k| - \frac{1}{2}\,(g-\mu_k)^T \sum_K^{-1}(g-\mu_k) + \ln Q(d=k) \quad (9)$$

Class predictions are calculated by maximizing the posterior probability using Eq. (10).

$$\hat{d} = arg.max_k Q(d=k|g) \quad (10)$$

D. Dataset

This study aims at the creation of a sentiment analysis system specializing in the analysis of emotional and opinionated posts in social media. Social media is quite popular and creates large textual data daily with useful knowledge of the public's perception of products, services, events, and social issues. This system is expected to utilize NLP tools to identify and sort sentiments of bilateral content, for example, positive sentiment or negative sentiment or even no sentiment. The dataset employed in this study is collected from open platform Kaggle, sourced from authentic social media data comprising of different styles of writing and different contexts such as brand tracking, in a crisis, for opinion mining and social trend analysis.

TABLE I. DESCRIPTION OF SYMBOLS USING IN EQUATIONS

Symbols	Description
$\boldsymbol{t}$	Term in a document
$\boldsymbol{d}$	Document
$\boldsymbol{n}$	Total number of terms in a document
$\boldsymbol{x}$	Frequency based on each word
$\boldsymbol{f(x)}$	Raw output of the base classifier
$\boldsymbol{a\ and\ b}$	Parameters optimized via logistics regression.
$\boldsymbol{w, x}$	Weight and feature vector
$\boldsymbol{y}$	True label (+1 or -1)
$\boldsymbol{\tau}$	Learning rate
$\boldsymbol{\mu_k}$	Mean vector of class k
$\boldsymbol{\sum k}$	Covariance matrix of class k
$\boldsymbol{e}$	Dimensionality of feature space
$\boldsymbol{TP, TN}$	True Positive and Negative
$\boldsymbol{FP, FN}$	False Positive and Negative

E. Evaluation Measures

Measures of performance evaluation are metrics used in machine learning that provide a way of qualifying the several aspects of the model's predictions, as shown in Table II. Accuracy tends to give a broad view of the accurateness of the model since it quantifies the actual number of properly classified samples to the overall number of samples. Recall measures the ratio of true positives among all actual positive observations, or the ability to avoid false negative predictions. Recall (Sensitivity) shows how many actual positives were correctly identified, which focuses on minimizing the number of negative cases that are positive. F1-Score, this metric is the harmonic meaning between Precision and Recall, which is better when used when the distribution is uneven.

TABLE II. EQUATIONS OF PERFORMANCE MEASURES

Metrics	Equation
Accuracy	$\frac{TP+TN}{TF+FN+FP+TP}$
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
F1-score	$\frac{2(Precision*Recall)}{Precision+Recall}$

AUC-ROC means Area Under the Receiver Operating Characteristic Curve, and it measures the model's conditional probability of correctly identifying a negative case using all the thresholds. Hence, these four metrics offer an uninterrupted way of evaluating the performance of the model such that the model's reliability and efficiency would be achieved.

IV. RESULTS

The results of the sentiment analysis experiments using three models, based on three different feature extraction techniques including TF-IDF, Count Vectorizer, and Word2Vec are summarized through confusion matrices and corresponding performance metrics. These results, as displayed in Table III help in directing focus to the model's strength and weakness aspect of correctly predicting sentiment labels namely negative, neutral and positive. The exploratory data analysis (EDA) visualizations summarize key textual patterns within the dataset:

A. Distribution of Text Length

From this histogram as shown in Fig. 2 (a), this graph describes the frequency of texts within the data set according to their length. The frequency distribution shows most texts are of lengths between 10 and 40 Words, although as text length increases, the number of texts decreases. We see that distribution is right-skewed, which means that there are more texts of shorter length than texts of very long length. This bar chart in (b) shows 10 most frequently appearing words in the dataset and the frequency of each of these words. Among these, “I’m”, “day”, “like”, “know” are few of the most frequent words used in day-to-day conversation. These often-used words seem to indicate that the dataset samples a daily or personal interaction-oriented environment.

B. Word Frequency Distribution for Words with Frequency >10

The bar chart shown in (c), will give a detailed analysis of the words that appear more than 10 times in the dataset. Specifically, words that can be found in the list of 10 most frequent words like the “I’m,” “like,” and “know” are in the middle. Like ‘Interests’, ‘Excitement’, some extra word like ‘amazing’, ‘day’, ‘today’, ‘tomorrow’ seems to point toward sentiment –rich contexts or likely sentiment temporal relatedness in the data set.

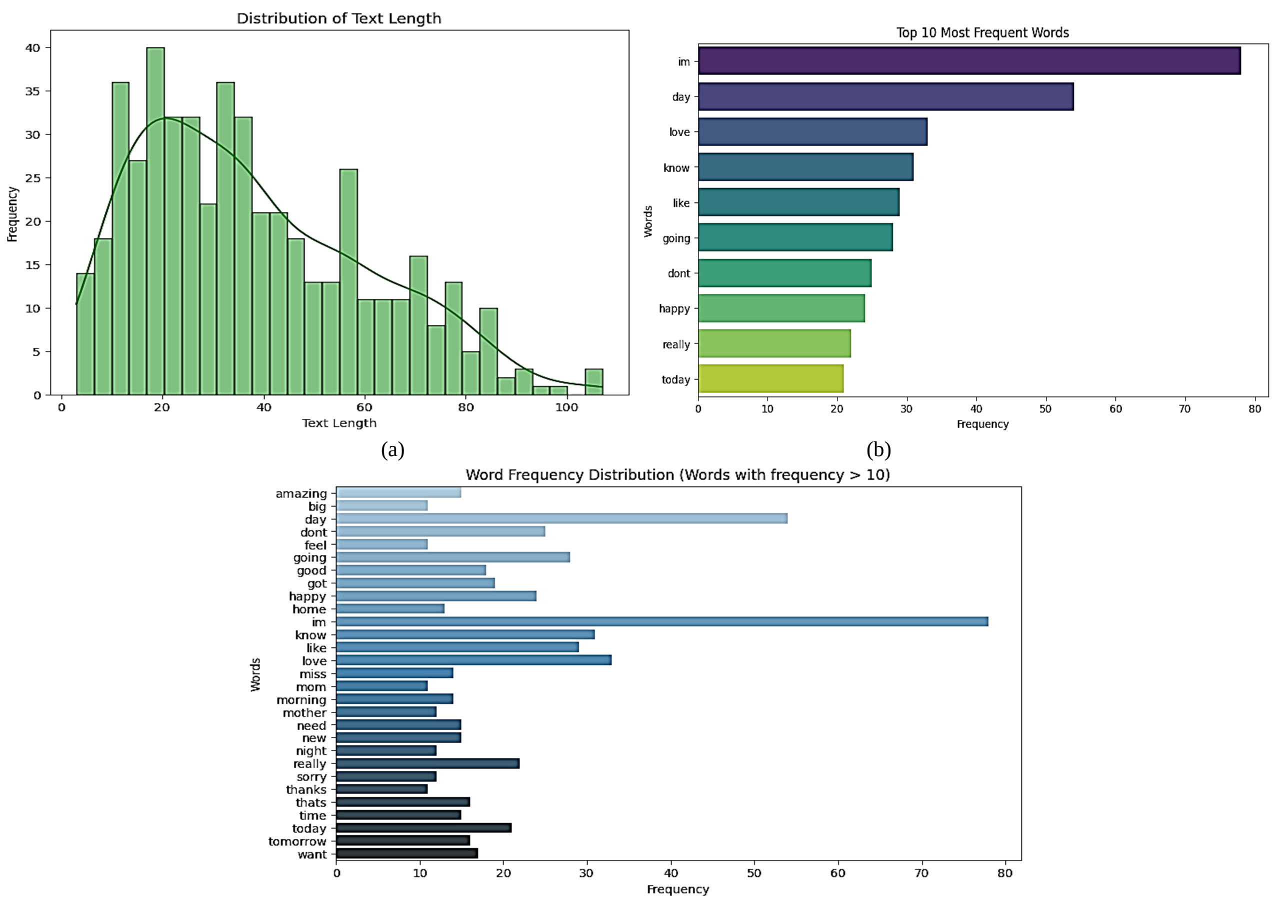
C. Distribution of Labels

In Fig. 3 displaying bar chart (a) capturing the current distribution of labels in sentiment analysis. Overall, the data split over a broad range with the sentiment of neutral prevailing over positive and negative sentiments, though with decreased number. The negative and positive sentiments are similar in the number of corresponding features, and they are between 125-175; however, the most prominently observed sentiments are the neutral ones with over 200 samples.

D. Word Cloud:

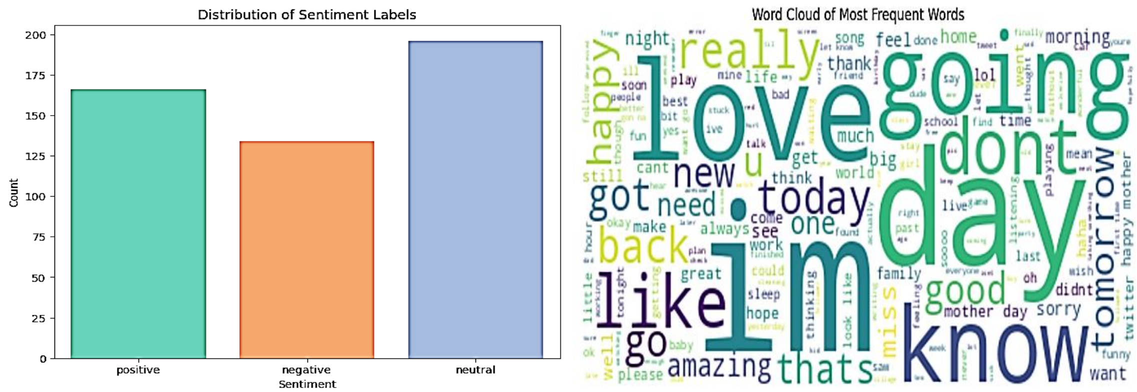
Fig. 3 preview (b) of the most often occurring words in the data set as a whole. The word cloud illustrates the frequency of words most often repeated in the dataset; it includes words such as love, going, day, I’m, know, among others. This shows the word frequency in sample texts, with possible positive words for the choice of ‘love’ and ‘day’ against possible negative words ‘don’t ‘and ‘can’t’.

This analysis underlines the fact that the dataset has conversational and sentiment-related properties, has more short texts, and uses more often and more frequently the most common sentiment-related words. It is useful in understanding the structure and contents of the text, therefore assists in the preprocessing and feature extraction steps that may be followed in other downstream tasks such as sentiment analysis attempting to build methodologies for classification models based on textual characteristics.



(a) (b) (c)

Fig. 2. Analysis of dataset text length along with frequency.



(a) (b)

Fig. 3. Analysis of ratio of sentiments along with most frequent words.

E. TF-IDF Results

By using feature extractor TF-IDF, PassiveAggressiveClassifier (PAC) demonstrated the highest accuracy of 76% and F1-score of 76%. The confusion matrix as shown in Fig. 4 presents a relatively good performance for all the different sentiment labels, while there is misclassification where positive sentiments are mistaken for merely neutral sentiments. The second model the CalibratedClassifierCV (CACV) brought an accuracy of 70% that clearly in the confusion matrix shows fairly good results but more false positive in the neutral class. When using QuadraticDiscriminantAnalysis (QDA), an accuracy of 72% was achieved, however from the confusion matrix, it reveals noticeable errors where neutral and positive were mistaken for negatives.

F. Count Vectorizer Results

While doing Count Vectorization, QDA provided the highest accuracy of 74% as well as F1- score of 71%. A confusion matrix, illustrated in Fig. 5 showing a few misclassifications of the current sentiment as neutral compared to other models, but sometimes positive sentiment was predicted as neutral. The use of the CACV resulted in an accuracy of 72 % and a confusion matrix indicating improved ability to identify between the neutral classes but slight concerns as to the distinction between the positive and neutral classes. The classifier implemented is the PAC which had an accuracy of 70% and the confusion matrix showed a considerable overlapping between positive and negative classifications.

G. Word2Vec Results

Word2Vec proved to be the most suitable for QDA obtaining 80% accuracy and 80% F1-score. Its confusion matrix in Fig. 6 shows good classification of the classes particularly for the neutral sentiment with very few inter changes. CACV yielded accuracy of 74 %, and the confusion matrix also shows that there is some mislabeling between positive and negative sentiments. At the same time PAC, which achieved 70% accuracy faced the highest difficulty in classification of neutral and positive sentiments where they were being mixed up most of the time. These confusion matrices give detailed information computed on the base of percentage values about how these models behave. Hence, the proposed QDA model with Word2Vec outperformed the others across metrics with minimal confusion among the sentiment classes, demonstrating its ability to leverage Word2Vec embeddings effectively. On the other hand, both TF-IDF and Count Vectorizer performed equally well with different classifiers, in which PAC provided the best result when using TF-IDF but worst result when using Word2Vec.

TABLE III. ANALYSIS OF APPLIED MODEL RESULTS WITH FEATURES

TF-IDF											
CalibratedClassifierCV				PassiveAggressiveClassifier				QuadraticDiscriminantAnalysis			
Acc	Pre	Recall	F1	Acc	Pre	Recall	F1	Acc	Pre	Recall	F1
70	70	70	70	76	78	75	76	72	71	72	71
Count Vectorization											
72	73	72	72	70	69	70	70	74	68	70	71
Word2Vec											
74	72	74	74	70	68	71	70	80	79	78	80

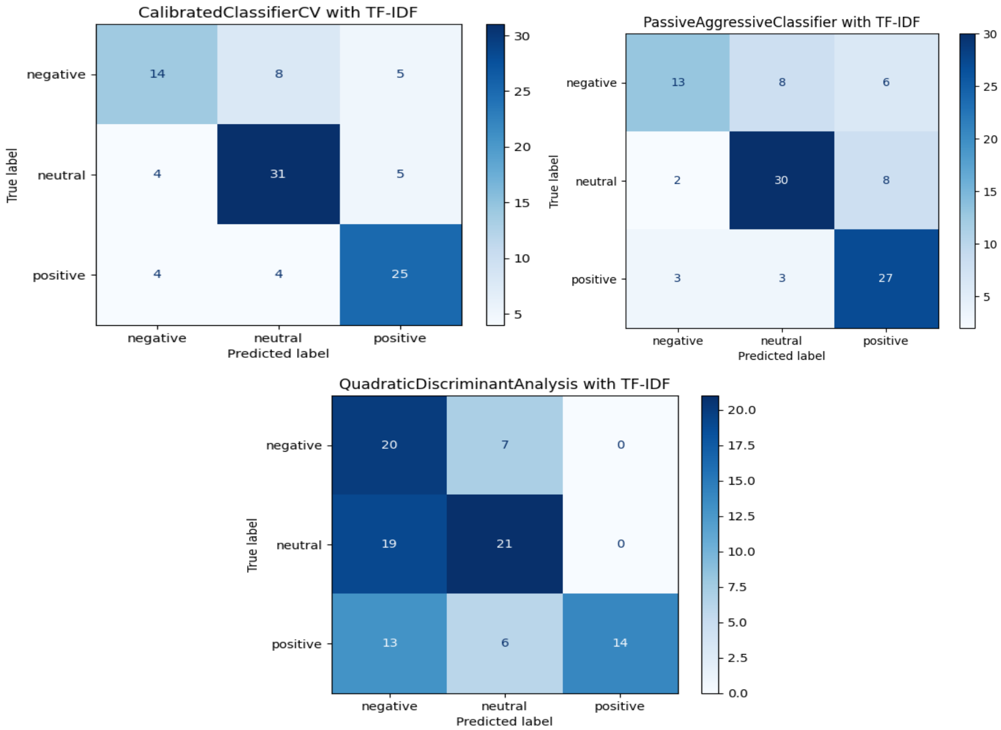


Fig. 4. Confusion Matrix of model performance using TF-IDF features.

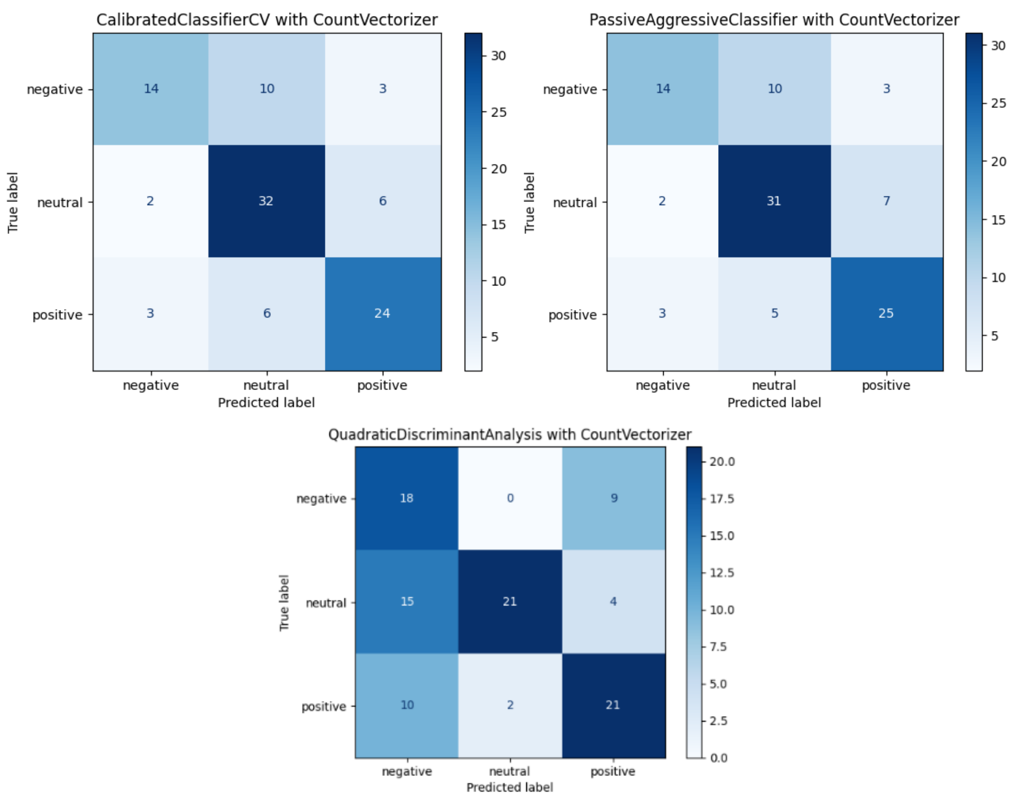


Fig. 5. Confusion matrix of model performance using count vectorizer features.

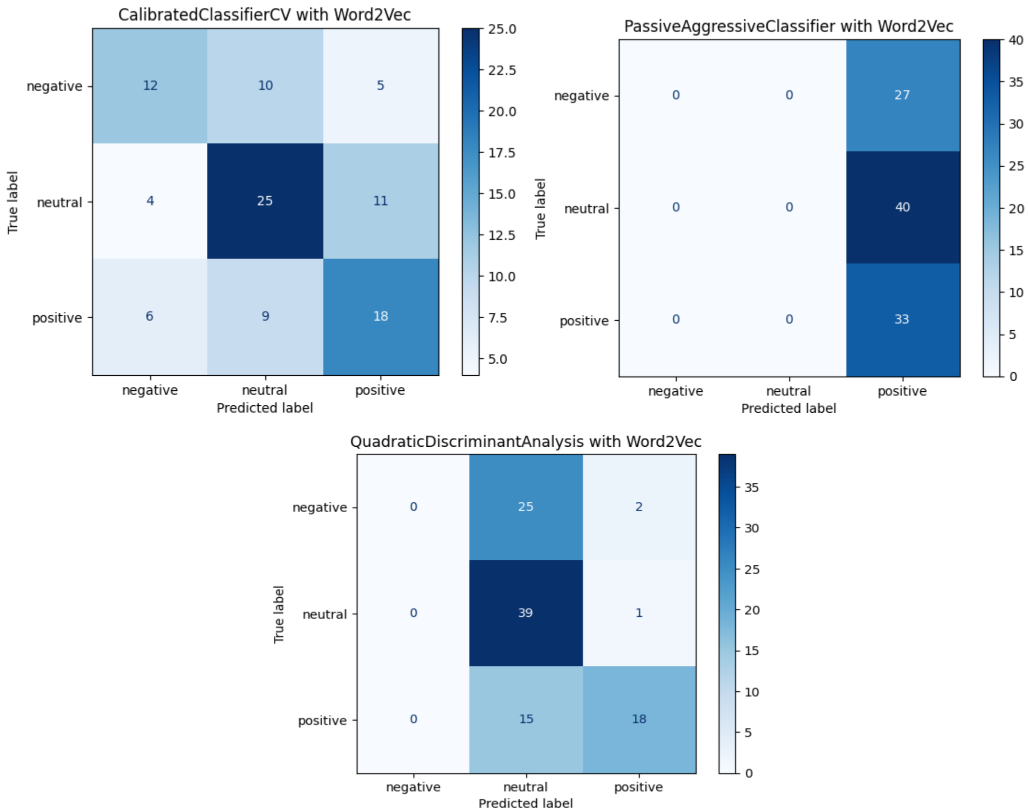
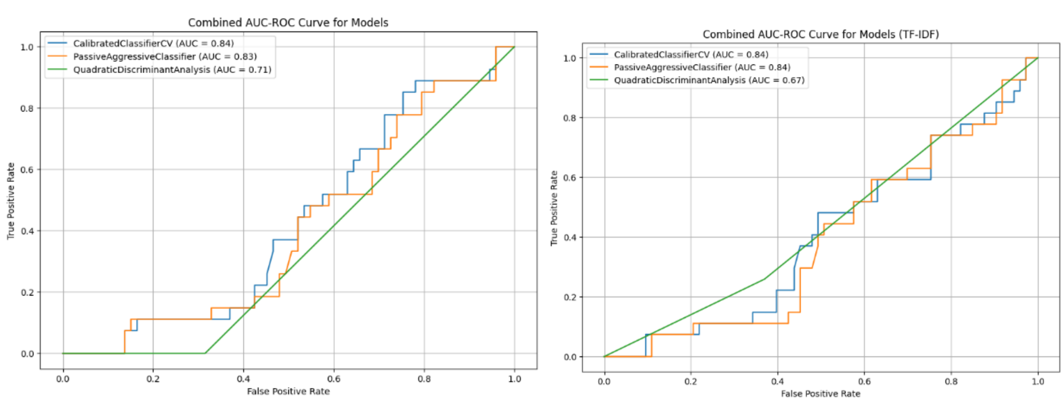


Fig. 6. Confusion matrix of model performance using Word2Vec features.

From the presented ROC curves, in Fig. 7, infer the overall classification performance of the models along with the examined feature extraction methods. For TF-IDF and Count Vectorizer, PAC and CACV have AUC of 0.76 – 0.81, which means the models are good in scenarios in which the separation between classes is clear; however, QDA has a problem with AUC values 0.66 – 0.72, suggesting a lower ability to distinguish between classes is lower. However, Word2Vec considerably enhances effectiveness; the best outcome is given by CACV (AUC = 0.73), followed by QDA (AUC = 0.66). These curves show that even though few classifiers like PAC works well with traditional features such as TF-IDF, the embedding techniques like Word2Vec yielded a better class separation in most of the classifiers as supported by higher AUC scores throughout most of the curves. Moreover, the ROC diagrams show that some models are more efficient in terms of true positive and false positive rates which is evident when comparing Word2Vec representations showing that feature embeddings affect the classification performance.



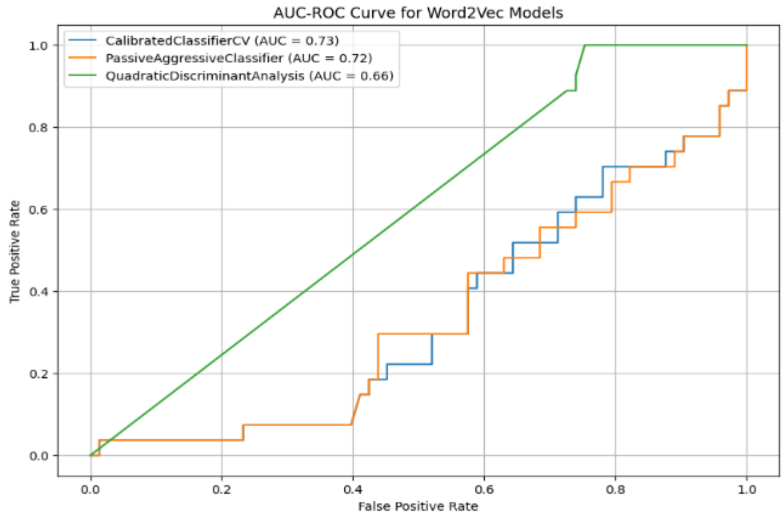


Fig. 7. Analysis of combined AUC-ROC model performance using features.

H. Discussion

To sum up, different feature extraction techniques have been effective in model performances of different degrees. Overall, Word2Vec was the best performer while QDA achieved the best accuracy of 80% and the best confusion matrix, further illustrating superior ability to deal with semantic resemblance for textual data, showing comparative analysis in term of accuracy measures illustrated as in Fig. 8. TF was more useful for PAC, and it scored the highest accuracy of 76% within its features, although it lower in performance with Word2Vec. Count Vectorization was moderate in its performance, in this field QDA was the most effective out of all (74% accuracy), however it didn't come close to the Word2Vec results. These trends therefore show how features interact with model type and suggest that advanced method such as Word2Vec work best for sentiment analysis models because they are best suited for capturing context and relation between words. These results suggest that extraction methods and classifiers employ significant roles in sentiment analysis.

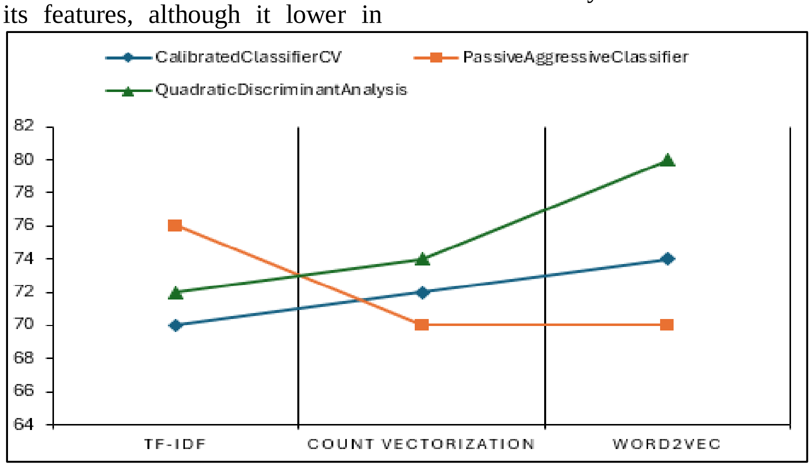


Fig. 8. Comparative analysis of accuracy measure along with models.

The contrast of the suggested model with other similar researches for sentiment prediction from textual information shows enhanced performances, as display in table IV. The current approaches, such as NB with 73%, RF with 74%, and SVC at 71%, have been computed in Twitter and IMBD datasets. On the other hand, the suggested Quadratic Discriminant Classifier in the Kaggle Social Media dataset just attained an 80% of accuracy. This considerable improvement demonstrates that the idea of the proposed model can learn various sentiment patterns, proving that the proposed model is capable of being a more feasible solution for the SA task as compared to the traditional machine learning models.

TABLE IV. COMPARISON WITH EXISTING STUDIES

Ref	Year	Models	Dataset	Results (%)
[18]	2020	RF	IMBD	74
[22]	2021	NB	Twitter	73
[23]	2023	SVC	Twitter	71
Proposed	**2025**	**QDC**	**Social Media**	**80**

V. Conclusion

Social media has become a digital world for users to share their opinions, views and interact through posts, messages, comments, and reviews, making it central mode for interpreting public sentiment. The role of AI in sentiment analysis is growing significantly, as it allows automated detection of emotions and opinions from UGC. In this study, we examined three feature extraction methods combined with various advanced AI models for sentiment classification. Among the models, QDA with Word2Vec embeddings achieved the highest accuracy of 80%, demonstrating its superior ability to capture semantic relationships and patterns in text. These findings show the effectiveness of integrating advanced feature representations with appropriate classifiers. Although the research study is helpful for predicting the sentiment analysis from online content, the limitation of the study that these findings are limited to textual data only. Future work highlights the significant exploration of deeper neural architectures and larger datasets to boost performance further. This study provides a gateway for developing more robust AI-driven tools for sentiment analysis, contributing to better understanding and leveraging user opinions in diverse applications.

References

[1] Zhang, L., Wang, S. and Liu, B., 2018. Deep learning for sentiment analysis: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), p.e1253.

[2] Iqbal, S., Khan, F., Khan, H.U., Iqbal, T. and Shah, J.H., 2022. Sentiment analysis of social media content in pashto language using deep learning algorithms. Journal of Internet Technology, 23(7), pp.1669-1677.

[3] Suryawanshi, N.S., 2024. Sentiment analysis with machine learning and deep learning: A survey of techniques and applications. International Journal of Science and Research Archive, 12(2), pp.005-015.

[4] Mahmood, A., Khan, H.U. and Ramzan, M., 2020. On modelling for bias-aware sentiment analysis and its impact in Twitter. Journal of Web Engineering, 19(1), pp.1-27.

[5] Krugmann, J.O. and Hartmann, J., 2024. Sentiment Analysis in the Age of Generative AI. Customer Needs and Solutions, 11(1), p.3.

[6] Convin.ai. (2024). Top Sentiment Analysis Tools to Watch in 2024 And Further. Retrieved from https://convin.ai/blog/sentiment-analysis-tools-2024.

[7] Ahmad, W., Khan, H.U., Iqbal, T. and Iqbal, S., 2023. Attention-based multi-channel gated recurrent neural networks: a novel feature-centric approach for aspect-based sentiment classification. IEEE Access, 11, pp.54408-54427.

[8] Kebede, D. and Tesfai, N., 2023. Ai-powered Text Analysis Tool for Sentiment Analysis.

[9] Al-Otaibi, S.T. and Al-Rasheed, A.A., 2022. A review and comparative analysis of sentiment analysis techniques. Informatica, 46(6).

[10] Ahmad, M., Aftab, S., Muhammad, S.S. and Ahmad, S., 2017. Machine learning techniques for sentiment analysis: A review. Int. J. Multidiscip. Sci. Eng, 8(3), p.27.

[11] Eliot, L., 2020. Legal Sentiment Analysis and Opinion Mining (LSAOM): Assimilating Advances in Autonomous AI Legal Reasoning. arXiv preprint arXiv:2010.02726.

[12] Ahmad, W., Khan, H.U., Iqbal, T., Khan, M.A., Tariq, U. and Cha, J.H., 2023. Hybrid multichannel-based deep models using deep features for feature-oriented sentiment analysis. Sustainability, 15(9), p.7213.

[13] Ishfaq, U., Khan, H.U. and Iqbal, K., 2016, April. Modeling to find the top bloggers using sentiment features. In 2016 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube) (pp. 227-233). IEEE.

[14] Mao, Y., Liu, Q. and Zhang, Y., 2024. Sentiment analysis methods, applications, and challenges: A systematic literature review. Journal of King Saud University-Computer and Information Sciences, p.102048.

[15] Chakriswaran, P., Vincent, D.R., Srinivasan, K., Sharma, V., Chang, C.Y. and Reina, D.G., 2019. Emotion AI-driven sentiment analysis: A survey, future research directions, and open issues. Applied Sciences, 9(24), p.5462.

[16] Wang, Y., Huang, M., Zhu, X. and Zhao, L., 2016, November. Attention-based LSTM for aspect-level sentiment classification. In Proceedings of the 2016 conference on empirical methods in natural language processing (pp. 606-615).

[17] Zhang, J., Liu, F.A., Xu, W. and Yu, H., 2019. Feature fusion text classification model combining CNN and BiGRU with multi-attention mechanism. Future Internet, 11(11), p.237.

[18] Neelakandan, S. and Paulraj, D., 2020. A gradient boosted decision tree-based sentiment classification of twitter data. International Journal of Wavelets, Multiresolution and Information Processing, 18(04), p.2050027.

[19] Ishfaq, U., Khan, H.U. and Iqbal, K., 2017. Identifying the influential bloggers: a modular approach based on sentiment analysis. Journal of Web Engineering, pp.505-523.

[20] Naz, A., Khan, H.U., Alesawi, S., Abouola, O.I., Daud, A. and Ramzan, M., 2024. AI Knows You: Deep Learning Model for Prediction of Extroversion Personality Trait. IEEE Access.

[21] Alsini, R., Naz, A., Khan, H.U., Bukhari, A., Daud, A. and Ramzan, M., 2024. Using deep learning and word embeddings for predicting human agreeableness behavior. Scientific Reports, 14(1), p.29875.

[22] Qi, Y. and Shabrina, Z., 2023. Sentiment analysis using Twitter data: a comparative application of lexicon-and machine-learning-based approach. Social Network Analysis and Mining, 13(1), p.31.

[23] Yadav, N., Kudale, O., Rao, A., Gupta, S. and Shitole, A., 2021. Twitter sentiment analysis using supervised machine learning. In Intelligent data communication technologies and internet of things: Proceedings of ICICI 2020 (pp. 631-642). Springer Singapore.