Semantic Analysis on Twitter User Data using the Hybrid Embedding Model: A Traditional Machine Learning Approach

Banjamin IBOK - 1532543

Abstract

Hybrid Embedding Models integrate statistical importance from TF-IDF with semantic contextual meaning from Word2Vec, and since the output of word vectors are numbers, these numbers can be fitted into Traditional Machine Learning models to make predictions. This paper explains the process of implementing each of the architectures on a Twitter dataset and compares them with the Hybrid Architecture on Logistic Regression, Random Forest, SVM, and XGBoost, analyzes the results, and makes recommendations based on findings to compare with other works done on the subject. The experiment used appropriate visualization techniques like Histograms, Word Cloud, Confusion Matrix and ROC Curves. Metrices like F-1 score, Precision, Support, Recall and Accuracy were used for evaluation, and from the outcome, TF-IDF was seen to outperform Word2Vec and even the Hybrid counterpart on all models with the highest Accuracy being 83% for SVM, 74% for Logistic Regression and 82% for Logistic Regression on the Hybrid Embedded architecture respectively.

1 Introduction

Sentiment analysis is a critical task in Natural Language Processing (NLP) that involves determining the sentiment polarity of textual data (Shaik et al., 2022). Traditional sentiment analysis models employ techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) and Word to Vector (Word2Vec) representations to transform text into numerical features (Dey & Das, 2023). While these approaches have demonstrated success in various applications, they also present challenges such as handling ambiguous language, sarcasm, and domain-specific sentiment variations (O. Slim et al., 2024). This study aims to evaluate the effectiveness of traditional sentiment analysis models based on a dataset from one of the Sentiment Evaluation Competitions held in 2017 (SemEval 2017). The task includes a multiclass classification problem on a dataset churned from the Twitter social network within the period of one month from December 2016 to January 2017, a subtask consisting mainly of user responses and the supposed sentiment (Rosenthal et al., 2017), a text length distribution of the dataset is shown in Figure 1.

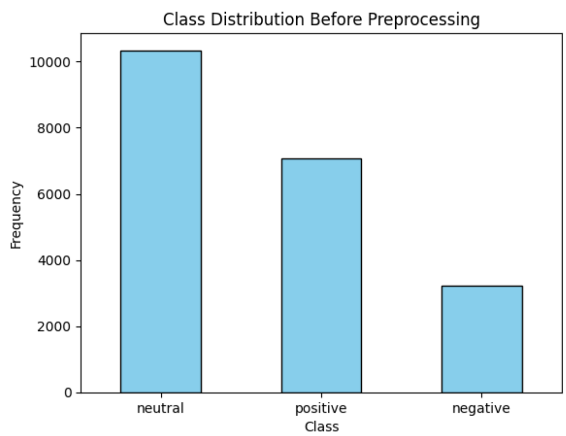

Figure 3: A figure showing data distribution of the target variable.

2 Related Research

According to a recent study (Nurhaliza Agustina et al., 2024), the role of feature extraction in sentiment analysis was emphasized, highlighting TF-IDF and Word2Vec as effective techniques on booster vaccine sentiment analysis, experiments were conducted using a Support Vector Machine (SVM) classifier. The dataset was preprocessed, and sentiment polarity was determined using TF-IDF and Word2Vec representations. The results showed that the SVM model achieved an accuracy of 89.5% with TF-IDF and 91.2% with Word2Vec. When both techniques were combined, the accuracy improved to 93.4%. The findings suggest that hybrid text representations enhance sentiment prediction, especially in vaccine-related discourse.

In another study by (Zhou et al., 2024), the efficiency of TF-IDF and Word2vec in extracting response behavior features from computer-based problem-solving evaluation was analyzed. The study compared the predictive, analytical, and clustering effects of classical machine learning methods on response behavior. According to the study, Random Forest model based on TF-IDF performed the best, followed by the SVM model based on Word2vec. Word2vec-based models outperformed TF-IDF-based ones in F1-score, accuracy, and recall.

Subsequently, (Zhan, 2025) compares the performance of TF-IDF and Word2Vec in sentiment analysis of food reviews using 560,000 food review data, the study focuses on the accuracy and generalization ability of the two methods under different dataset sizes. A previous study compared the performance of TF-IDF and Word2Vec under different dataset sized, and concluded that TF-IDF had better performance when the dataset was small, while Word2Vec showed better semantic capturing ability when the dataset increased (Bai et al., 2017).

3 Dataset Description

Dataset used contains 20,632 entries and no missing values. However, there were found 52 duplicating **Text** entries with exact matching details which were removed to reduce bias, misleading patterns and even overfitting (Goodfellow et al., 2016). Also, it was observed that the column named **Labels** contained 10 instances of timestamps which was not required for the prediction and as such was removed. The **Sentiment** column contains 3 unique values as follows, Negative, Neutral and Positive, with a frequency of 3221, 10313 and 7046 respectfully as shown in the distribution in Figure 2. This column was later encoded and mapped back to **Labels** column for use as target variable in the prediction.

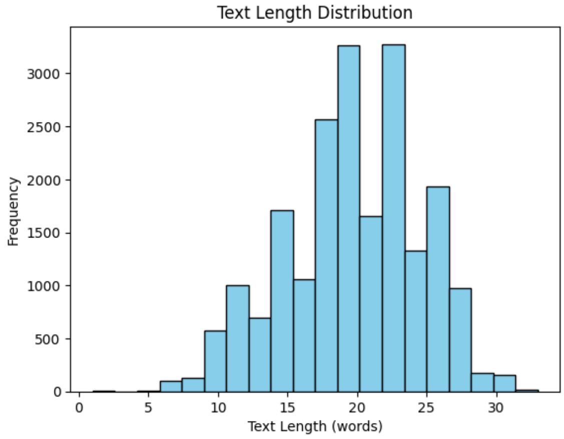

Figure 2: A figure showing text length distribution of the dataset.

4 Machine Learning Methods

4.1 Text Pre-Processing

Text preprocessing is the application of data mining techniques for cleaning text data before fitting into a ML model to allow for proper analysis and ensure accurate results. In this experiment, a text preprocessing pipeline was used, lowercasing and text contraction resolution was performed to convert words like “he’s” to “he is”, removal of URLs, mentions, hashtags, numbers, and care was taken to ensure that removal of punctuation was performed before tokenization to prevent unwanted splitting operations (Jurafsky & Martin., 2021). Removal of single letter words, extra-spaces and emoticons was also performed, check for presence of emojis was also conducted since the dataset is of social media origin.

4.2 Text Decomposition

An important text preprocessing operation is Tokenization. Tokenization is the segmentation of text into smaller units called tokens, and this operation this was carried out in the preprocessing pipeline right before Lemmatization.
Lemmatization is a text preprocessing technique in NLP that is used for decomposing words to their base form (lemma), by removing any form of suffixes and reducing a word to its root. This method helps structure the data for the computing algorithm, so words like “flying”, “gifted”, and “understood” get reduced to “fly”, “gift”, and “understand” respectively. There are two different methods for structuring text data, stemmatization and lemmatization. While the former just chops off the suffix without considering the context of the

word in the sentence, lemmatization considers the context of the word in the sentence, because it uses an inbuilt dictionary to compare the words. Lemmatization guarantees more accurate predictions (Chai, 2023), makes it the choice for this experiment.

4.3 TF-IDF

Feature importance is very essential in sentiment analysis (Nurhaliza Agustina et al., 2024), and TF-IDF is one of the techniques used to achieve this. TF-IDF scans through dataset corpus for the most occurring words (Term Frequency), and attaches weighted values using these metrics, high frequency words got lowest weights vice-versa (Inverse Document Frequency). TF-IDF uses a vectorization function to convert text to vectors for computation and maintains a fixed length output. An illustration of the TF-IDF words score in Figure 3.

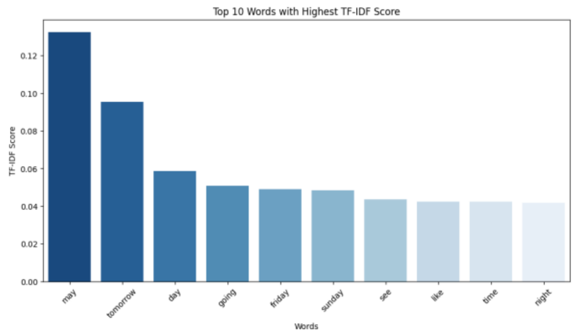

Figure 3: Showing the TF-IDF Top 10 words score.

4.4 Word2Vec

Word2Vec is another feature extraction technique which uses word embeddings to convert text to vectors for computation (Johnson et al., 2023). It comprises the features the Continuous bag of Words (COW), which predicts a targeted word given the surrounding words, and Skip-gram which predicts surrounding words given a target word. CBOW alternative was used in this experiment because it works better with Traditional ML models in Sentiment analysis and boasts of improved speed in training the model and performs better with smaller datasets.

4.5 Random Forest

Random Forest is an ensemble Traditional ML model that uses the decision tree algorithm to perform predictions. It is referred to as an ensemble model because it combines the output of multiple inbuilt decision trees to form an output. The key idea behind ensemble learning is that a group of individual trees combine to create a more efficient model with better accuracy and generalization by averaging their outputs. Random Forest algorithm is used in this experiment because it is known for handling text classification problems efficiently when combined with Word2Vec or TF-IDF (Hitesh et al., 2019).

4.6 SVM

SVM is a powerful supervised learning model that can be used for text classification problems. It is particularly effective for high dimensional spaces, such as text data represented through TF-IDF or Word Embeddings. SVM works by finding a hyperplane to separate multiple datapoints from classes in the vector space. The goal is to maximize the margin between the closest points (support vectors) of different classes. This makes it useful in NLP for text classification (Lilleberg et al., 2015). In this work, SVM will be used with TF-IDF, Word2Vec and a combination of both to solve the classification problem on our dataset.

4.7 XGBoost

XGBoost (Extreme Gradient Boosting) is a high-performance machine learning algorithm that is part of the gradient boosting family, it is another ensemble method that is based on decision trees, but has a gradient boosting feature that runs in a sequential manner and takes the output of one decision tree and feeds it to the next one down the line, each new tree is trained to correct the errors of the previous one. XGBoost also has a lot of parameters that can be fine tuned to improve model performance. In this work, XGBoost is used to classify text due to its high accuracy and versatility.

4.8 Logistic Regression

Logistic Regression is another supervised learning algorithm used primarily for binary classification problems, although it can be extended to multi-class classification. Even though Logistic regression requires modification for use with multi-class classification, it is widely used because of its simplicity and linearity. Logistic Regression was chosen for this experiment because it assumes linearity in the feature space, this is advantageous because in some NLP text representations like TF-

IDF, text data are seen as independent items in the feature space.

5 Experiments

5.1 Data Preparation

Data preparation for the multiclass classification started with previewing the dataset to understand the nature of the dataset conducting Exploratory Data Analysis, visualizing word cloud of the dataset before and after text preprocessing as seen in Figure 4 and 5. The word cloud in Figure has more meaningful and recognizable text than the image in Figure 4.

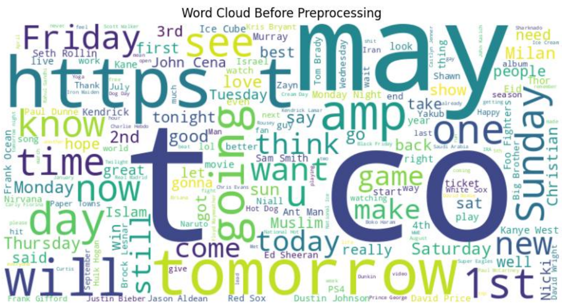

Figure 4: Showing Word Cloud before Preprocessing.

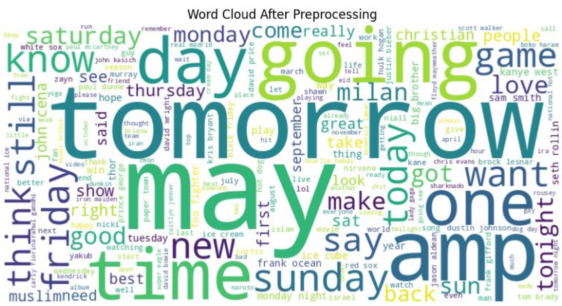

Figure 5: Showing Word Cloud after Preprocessing.

5.2 Experiment Workflow

The experiment accesses the performance of the selected models Logistic Regression, Randon Forest, SVM and XGBoost on TF-IDF, Word2Vec using CBOW and finally, the models are accessed on the Hybrid architecture which combines TF-IDF with Word2Vec and the results of the evaluations are collected and compared.

5.3 Discussion of Results

The results of the evaluations can be seen in Table 1, TF-IDF outperforms the Word2Vec and Hybrid Text Representation Architectures both in Accuracy and ROC evaluations. A graphical visualization of the ROC score on all the models are displayed in Figure 6 and 7 for the TF-IDF, Word2Vec and the Hybrid architecture.

5.4 Recommendation for Further Research

To further the study, the Word2Vec Skip-gram feature can be combined with TF-IDF to see if the performance can improve. Alternatively, Naïve Bayes could be used to test the Hybrid architecture to see the performance, another very remarkable

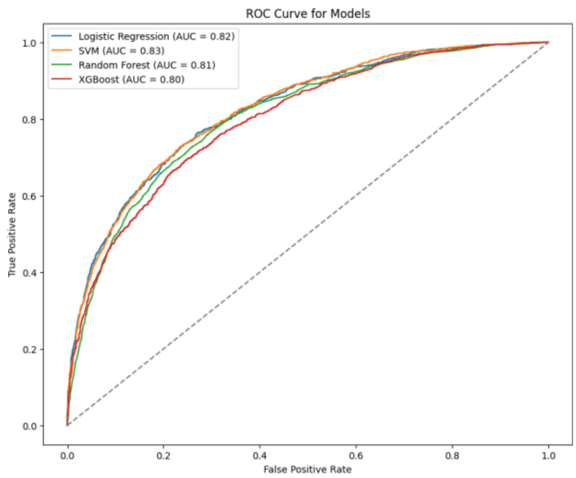

Figure 6: Showing ROC curve for TF-IDF.

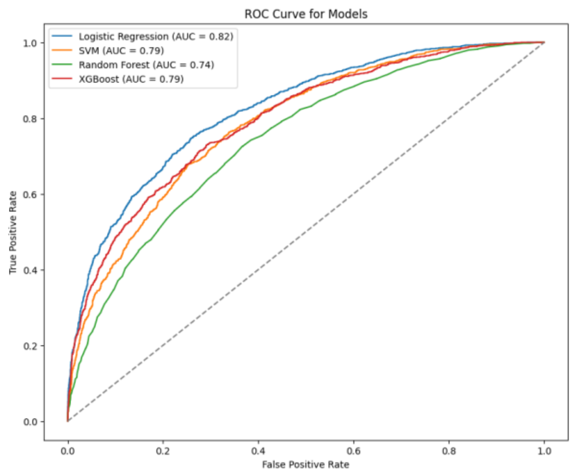

Figure 7: Showing ROC curve for the Hybrid Architecture.

ROC			
MODEL	**TF-IDF**	**W2V**	**HYBRID**
Log. Reg	82%	74%	82%
Rand. F.	81%	73%	74%
SVM	83%	74%	79%
XGB	80%	73%	79%

Table 1: ROC.

recommendation would be to run the Hybrid model with a Neural Network Model.

Accuracy			
MODEL	**TF-IDF**	**W2V**	**HYBRID**
Log. Reg	65%	58%	65%
Rand. F.	65%	58%	60%
SVM	64%	57%	58%
XGB	63%	56%	64%

Table 2: Accuracy.

6 References

Bai, T., Dou, H.-J., Zhao, W. X., Yang, D.-Y., & Wen, J.-R. (2017). An Experimental Study of Text Representation Methods for Cross-Site Purchase Preference Prediction Using the Social Text Data. *Journal of Computer Science and Technology*, *32*(4), 828–842. https://doi.org/10.1007/s11390-017-1763-6

Chai, C. P. (2023). Comparison of text preprocessing methods. *Natural Language Engineering*, *29*(3), 509–553. https://doi.org/10.1017/S1351324922000213

Dey, R. K., & Das, A. K. (2023). Modified term frequency-inverse document frequency based deep hybrid framework for sentiment analysis. *Multimedia Tools and Applications*, *82*(21), 32967–32990. https://doi.org/10.1007/s11042-023-14653-1

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT press.

Hitesh, M., Vaibhav, V., Kalki, Y. J. A., Kamtam, S. H., & Kumari, S. (2019). Real-Time Sentiment Analysis of 2019 Election Tweets using Word2vec and Random Forest Model. *2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT)*, 146–151. https://doi.org/10.1109/ICCT46177.2019.8969049

Johnson, S. J., Murty, M. R., & Navakanth, I. (2023). A detailed review on word embedding techniques with emphasis on word2vec. *Multimedia Tools and Applications*, *83*(13), 37979–38007. https://doi.org/10.1007/s11042-023-17007-z

Lilleberg, J., Zhu, Y., & Zhang, Y. (2015). Support vector machines and Word2vec for text classification with semantic features. *2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC)*, 136–140. https://doi.org/10.1109/ICCI-CC.2015.7259377

Nurhaliza Agustina, C. A., Novita, R., Mustakim, & Rozanda, N. E. (2024). The Implementation of TF-IDF and Word2Vec on Booster Vaccine Sentiment Analysis Using Support Vector Machine Algorithm. *Procedia Computer Science*, *234*, 156–163. https://doi.org/10.1016/j.procs.2024.02.162

O. Slim, S., Elsayed Aboutabl, A., & Derbala Yacoub, A. (2024). A Survey of Sentiment Analysis and Sarcasm Detection: Challenges, Techniques, and Trends. *International Journal of Electrical and Computer Engineering Systems*, *15*(1), 69–78. https://doi.org/10.32985/ijeces.15.1.7

Rosenthal, S., Farra, N., & Nakov, P. (2017). *SemEval-2017 Task 4: Sentiment Analysis in Twitter*.

Shaik, T., Tao, X., Li, Y., Dann, C., McDonald, J., Redmond, P., & Galligan, L. (2022). A Review of the Trends and Challenges in Adopting Natural Language Processing Methods for Education Feedback Analysis. *IEEE Access*, *10*, 56720–56739. https://doi.org/10.1109/ACCESS.2022.3177752

Zhan, Z. (2025). Comparative Analysis of TF-IDF and Word2Vec in Sentiment Analysis: A Case of Food Reviews. *ITM Web of Conferences*, *70*, 02013. https://doi.org/10.1051/itmconf/20257002013

Zhou, J., Ye, Z., Zhang, S., Geng, Z., Han, N., & Yang, T. (2024). Investigating response behavior through TF-IDF and Word2vec text analysis: A case study of PISA 2012 problem-solving process data. *Heliyon*, *10*(16), e35945. https://doi.org/10.1016/j.heliyon.2024.e35945

Jurafsky, D., & Martin, J. H. (2021). Transfer learning with pretrained language models and contextual embeddings. Speech and language processing (3rd ed. Draft). https://web. stanford. edu/~ jurafsky/slp3/11. pdf.

7 Appendix

Find link for the code at: https://github.com/JaminUbuntu/NLP-Coursework-Benjamin/blob/main/NLP_Coursework_Benjamin.ipynb