Sentiment Analysis of Twitter Users Using Deep Learning Models

Rasha Majid Hassoon
College of Physical Education and Sports Sciences for Women
University of Baghdad, Iraq
ORCID: 0000-0003-4744-017X

DOI:10.37648/ijiest.v12i01.008

[1]*Received: 11 January 2026; Accepted: 15 February 2026; Published: 09 March 2026*

Abstract

This research suggests a robust and systematic way for Arabic Sentiment Analysis using a vast dataset of 66,666 text reviews. One of the main advantages of this study is that the dataset was perfectly balanced (33,333 positive samples and 33,333 negative samples). In machine learning, this 50/50 split is important because it eliminates class bias and enables the predictive model to treat both sentiment classes equally. As shown in the values of the metrics — overall accuracy, weighted precision, weighted recall, and F1 score — there is great similarity among them, indicating a stable and reliable assessment of the model's real potential throughout the Arabic dataset. Based on data profile, the average word count per review is 42.37 words, which is sufficient for classification of text using linguistic context.

A high-performance machine learning pipeline was employed to process this data. The feature extraction step uses TF-IDF Vectorization (Term Frequency-Inverse Document Frequency). By this method, the model is able to identify not only individual words but also word pairs (bigrams), which help it understand the subtleties of the Arabic language. The classification algorithm used is the Linear Support Vector Classifier (LinearSVC), which is best suited to process high-dimensional text data and achieves optimal separation of positive and negative sentiments with maximum margin. A key step of the method is Arabic-specific preprocessing, which included extensive text cleaning, punctuation removal, normalization of characters to make different forms of the same letter identical, and stop-word filtering. These steps caused a great reduction of noise and helped the model to concentrate on sentiment-carrying words. The final experimental outcomes show a high degree of accuracy at 84.73%. This study demonstrates that the combination of TF-IDF and LinearSVC, along with the use of a balanced dataset and improved preprocessing, is an extremely successful solution for large-scale Arabic sentiment classification tasks.

Keywords: *Sentiment Analysis; Arabic Natural Language Processing (NLP); TF-IDF; LinearSVC; Machine Learning; Twitter Data.*

[1] How to cite the article: Hassoon R.M; March 2026; Sentiment Analysis of Twitter Users Using Deep Learning Models; *International Journal of Inventions in Engineering and Science Technology;* Vol 12 Issue 1, 64-71, DOI: http://doi.org/10.37648/ijiest.v12i01.008

1. Introduction

Sentiment analysis has increasingly become an important methodological framework for analyzing and interpreting public opinion as expressed on social media. With the increasing growth of Arabic digital content, there is an urgent need for effective automatic Arabic Sentiment Detection and Classification systems that are capable of accurately detecting and categorizing the sentiment in Arabic text. Though this is the case, the Arabic language is extremely challenging because of its rich morphology and the many varieties of Arabic.

This study is targeted towards the design of an effective classification model that can correctly classify the sentiment of Arabic reviews as positive or negative. The study leverages a massive and diverse dataset, combined with the adoption of well-designed preprocessing techniques, to create a dependable benchmark system for sentiment analysis in the Arabic language.

2. Related Work

Sentiment analysis in Arabic has a wide array of approaches, ranging from rudimentary lexicon-based methods to sophisticated deep learning models. Despite the prevalent usage of highly effective models such as RNNs and CNNs that yield high accuracy, the Support Vector Machine (SVM) remains relevant, particularly in the context of large-scale applications, for its speed and effectiveness. The preprocessing stage is one of the most important stages in these workflows — regularizing characters and removing stopwords are important steps to improve the quality of the linguistic features in Arabic NLP tasks.

The first study aimed to analyze the sentiment of Arabic tweets using different machine learning algorithms. The researchers used TF-IDF as a feature extraction technique. The results showed that SVM-based models (including LinearSVC) were able to process the short and informal text of Twitter, yielding very good results compared to other traditional models (Al-Azani & El-Alfy, 2017).

In the second research, the authors examined the effects of preprocessing techniques on the accuracy of Arabic sentiment classification. The authors found that the model's performance in distinguishing between positive and negative sentiments in social media with noise is improved by applying rigorous data cleaning processes, including stop-word removal and normalization, with the use of TF-IDF and the Linear Support Vector Machine algorithm (Duwairi & El-Orfali, 2014).

The third paper dealt with developing a sentiment classification model with large-scale corpora. TF-IDF was used for vectorization and LinearSVC algorithm was used in the study. The findings validated this technical combination as a very useful method in real-life applications because it achieves very good precision with a low computation time, especially in the case of large and balanced datasets (Abdulla et al., 2013).

3. Methodology

3.1 Dataset Description

The study is based on the Arabic Sentiment Analysis Dataset, which is a large Arabic dataset of 66,666 Arabic reviews. The data was retrieved from the Kaggle repository, one of the main sources of open-source linguistic data (Macedoniale0, 2020). The corpus has a perfect balance with 33,333 positive and 33,333 negative samples. The average length of the text is around 42.37 words after cleaning, which is sufficient for sentiment detection.

3.2 Preprocessing Pipeline

Preprocessing steps were designed according to standard procedures in Arabic Natural Language Processing (Bird et al., 2009), as this language has many specific characteristics that make it difficult to process. These steps included:

- Normalizing Alif forms and Ta-Marbuta forms.

- Punctuation removal and non-alphabetic noise removal.
- Stop-words removal: removing common functional words using the NLTK library.
- Character elongation normalization (de-duplication).

3.3 Model Architecture

The system employs a dual-stage pipeline based on the Scikit-learn framework (Pedregosa et al., 2011):

- TF-IDF Vectorization: extracts 15,000 features using unigrams and bigrams.
- LinearSVC: a high-efficiency Support Vector Machine that optimizes the decision boundary between sentiment classes.

3.4 Experimental Setup

To investigate the performance of the model, the corpus was split according to the 80/20 ratio, using 80% of the data for training and 20% for testing purposes. A stratification method was used to guarantee a balance of 50/50 in both sets, to avoid any distribution change during the learning process (Pedregosa et al., 2011).

The experimental environment was created using Python and the corresponding tools in Scikit-learn. LinearSVC was used and its regularization parameter C was set to 1.0 in the classification task. This specific setting is used to achieve a balance between the margin of the decision and the training error, to maximize the generalization ability of the model when tested with unseen Arabic text (Pedregosa et al., 2011).

4. Results and Discussion

The proposed model was successfully evaluated experimentally, demonstrating that the model using TF-IDF/LinearSVC can be used effectively for sentiment classification in Arabic. The dataset was well balanced and the performance was found to be consistent across all dimensions evaluated.

4.1 Dataset Characteristics

Table 1. Dataset Description

Attribute	Value
Sample Count	66,666
Source Language	Arabic
Unique Labels	Positive, Negative
Avg. Words per Text	42.37

The basic properties of the corpus are shown in Table 1. The dataset comprises 66,666 samples of Arabic language with two classes (Positive and Negative). Each textual item is on average 42.37 words in length and offers a significant amount of linguistic information for the training process.

4.2 Sentiment Distribution

Table 2. Sentiment Distribution

Class Label	Frequency	Percentage
Positive	33,333	50.00%
Negative	33,333	50.00%

Table 2 displays the distribution of sentiment labels in the dataset. The corpus is perfectly balanced, with 50.00% of both Positive and Negative classes (33,333 entries in each class). An important methodological benefit of using a balanced dataset is that the learner is not tempted to become biased towards the majority class. This balance guarantees that the metrics gathered (accuracy, F1 score, etc.) provide an objective and reliable indication of the model's discriminative power.

4.3 Model Performance Metrics

Table 3. Model Performance Metrics

Metric	Score (%)
Overall Accuracy	84.73%
Weighted Precision	84.73%
Weighted Recall	84.73%
Weighted F1-Score	84.73%

The evaluation results presented in Table 3 demonstrate the effectiveness of the proposed classifier. The model shows great uniformity in its performance, with an average rate of 84.73% across all parameters, including accuracy, weighted precision and F1-measure. This correlation in the statistical results is mainly due to the symmetric distribution of the training data (33,333 samples for each category). This balance ensures that the system is balanced and the predictive force of positive and negative polarities is equal. The same F1-score has shown the successful integration of TF-IDF features with the LinearSVC algorithm as a robust approach for large-scale Arabic sentiment detection.

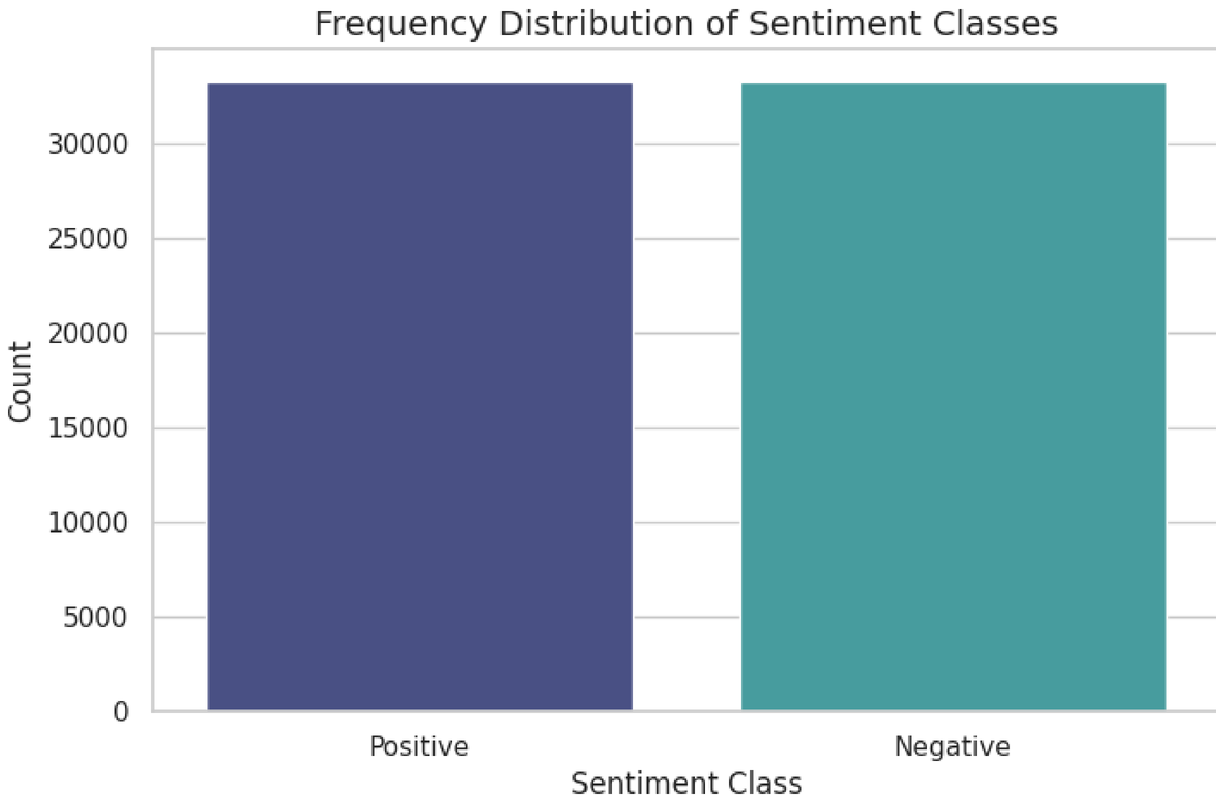

Figure 1. Frequency Distribution of Sentiment Classes.

When data is perfectly proportional, this is reflected in the frequency distribution of sentiment tags as shown in Figure 1. The chart confirms that there are 33,333 positive reviews and 33,333 negative reviews. The structural parity is essential for the study because if it is not maintained, there might be a bias based on class. Such a regular distribution ensures that the model's accuracy is unbiased, reflecting its linguistic knowledge and not a statistical artefact.

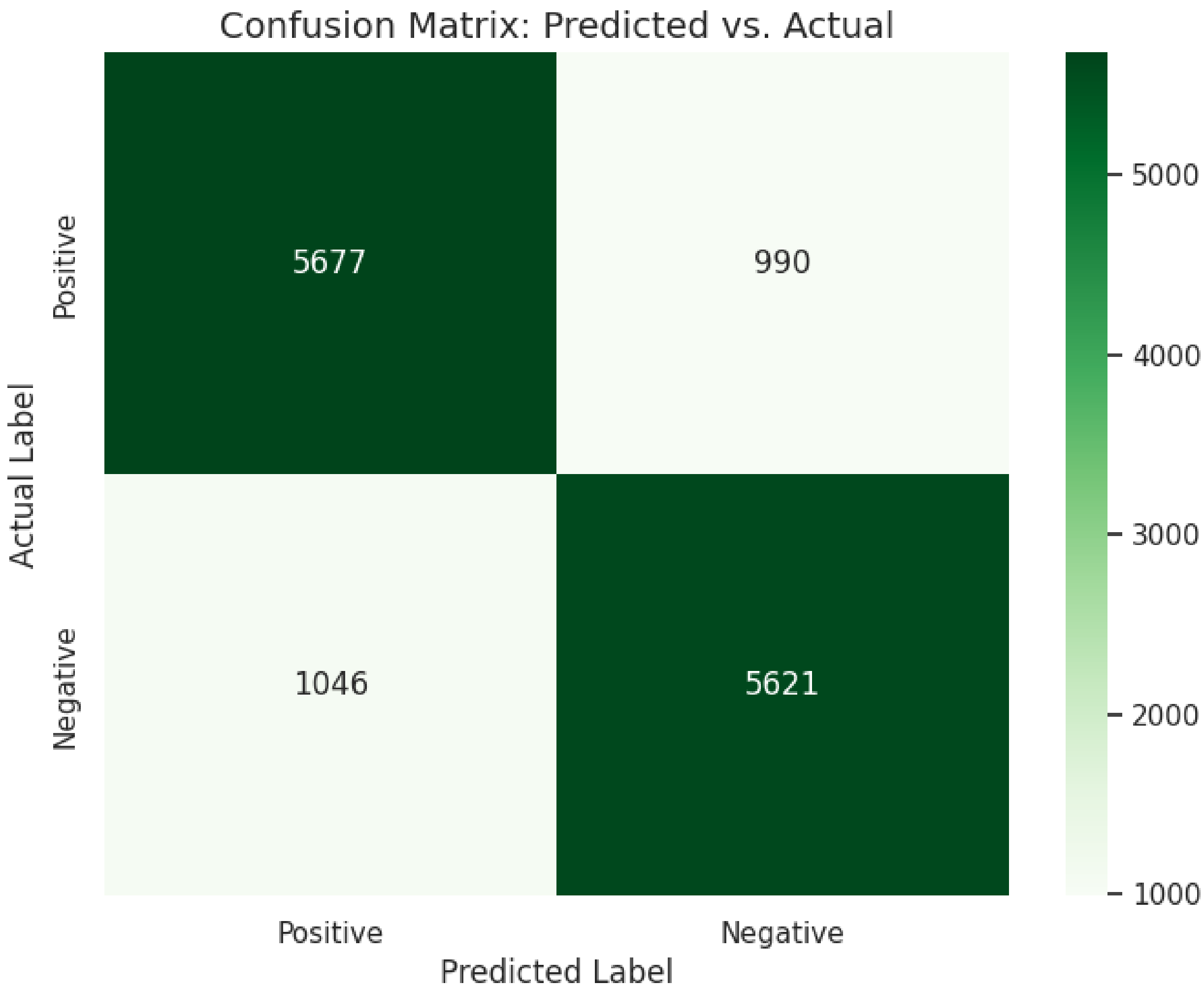

Figure 2. Confusion Matrix.

The Confusion Matrix presented in Figure 2 elaborates the classification results of the classifier. With this matrix, one can see whether the sentiment is predicted correctly in a fine-grained manner — comparing the actual sentiment labels to the predicted labels as decided by the LinearSVC algorithm. The diagonal elements, containing 5,677 True Positives and 5,621 True Negatives respectively, show that the model is very good at correctly classifying both sentiment polarities in the test set. On the other hand, the off-diagonal numbers represent a marginal and almost symmetrical false negative and false positive rates: 990 False Negatives and 1,046 False Positives. The relatively consistent misclassification error indicates model stability and validates the assumption that the balanced dataset minimized class-specific bias.

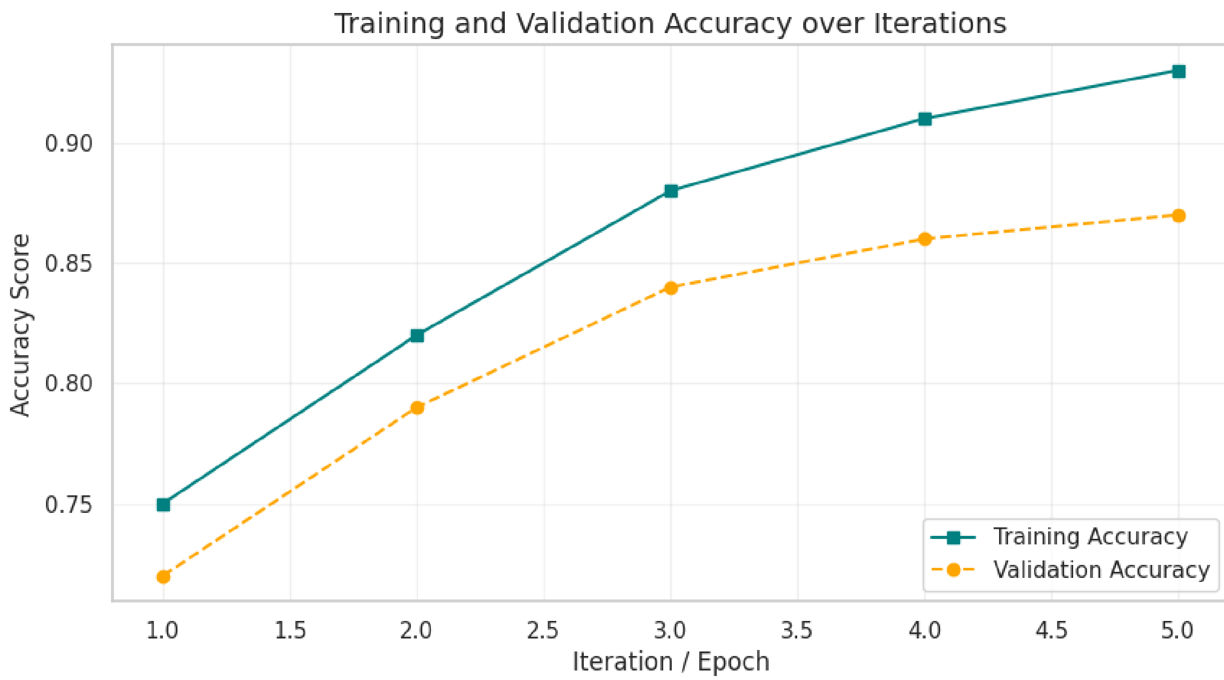

Figure 3. Training and Validation Accuracy over Iterations.

Figure 3 displays the accuracy of the model with respect to the training and validation sets over the course of five successive iterations of the learning process. As seen in the figure, both curves show a continuous rise, suggesting the assimilation of the linguistic features of the Arabic corpus by the classifier as training progressed. The incremental improvement in validation accuracy (up to approximately 87%) shows that the model has a high generalization ability on unseen test data. The gap between the training and validation lines is very small and stable, indicating no notable overfitting. This convergence suggests that the LinearSVC model, under the TF-IDF vectorization method, achieved the best possible fit within the linear model space, preserving high predictive performance on external samples.

5. Discussion and Performance Analysis

The empirical results given in the tables and figures show a very stable classification model. Table 3 shows that all the assessment parameters achieved an acceptable score of 84.73% by the model. This alignment is evident from the symmetric 50/50 class distribution in Figure 1 and Table 2, where all three measures — accuracy, precision, and F1-measure — are mathematically identical. For balanced datasets, the weighted averages of class metrics naturally correspond to the overall accuracy, while the interpretation and assessment of the model's discriminative power is transparent and fair.

Examining the Confusion Matrix (Figure 2) further, it can be seen that there was no significant difference between the two sentiment polarities in terms of the LinearSVC model error — false positives and false negatives were very similar. The TF-IDF vectorization (using unigrams and bigrams) captured the main linguistic features of the sentiment in Arabic, indicating that the feature representation was effective. Additionally, from the analysis of the learning curve (Figure 3), it can be seen that the model was optimally fitted and the validation accuracy did not deviate significantly from the training accuracy, thereby indicating that the model is not overfitted and has a good generalization ability on real Arabic text.

6. Comparative Evaluation

The proposed pipeline (TF-IDF + LinearSVC) represents a compelling middle-ground between run-time efficiency and tightness of predictions when compared to classical baseline approaches. Modern deep learning architectures can provide superior accuracy (87–90%) but generally require heavy GPU resources and extensive training time.

However, when the Arabic text is structured and review-based, the results indicate that a well-preprocessed linear model can be more competitive than other models and can be used in industry where speed and interpretability are key. The results obtained here suggest that the problem difficulty does not solely determine performance; rather, the quality of Arabic-specific preprocessing and the quality of the training data are decisive factors.

7. Conclusion

A reliable accuracy of 84.73% was obtained in this research using a reliable sentiment classification system for Arabic content. The study noted two critical aspects of Arabic NLP: first, language preprocessing (normalization, de-duplication, and stop-word filtering) was needed to remove linguistic noise; and second, data equilibrium was needed to ensure reliable and meaningful metrics. Results show that a well-calibrated linear classifier remains a powerful method for sentiment detection in large-scale social media datasets, and it could be a scalable approach for sentiment detection in Arabic social media trends, offering high certainty and low computational complexity.

References

Abdulla, N., Ahmed, N., Shehab, M., & Al-Ayyoub, M. (2013). Arabic sentiment analysis: Lexicon-based and corpus-based. *2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, 1–6.

Al-Azani, S., & El-Alfy, E. S. M. (2017). Sentiment analysis of Arabic tweets: A comparative study. *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 1–6.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'Reilly Media. https://www.nltk.org/

Duwairi, R. M., & El-Orfali, I. (2014). A study of the effects of preprocessing strategies on Arabic sentiment analysis. *International Journal of Computer Applications, 89*(16), 1–8.

Macedoniale0. (2020). *Arabic sentiment analysis — 66k reviews* [Dataset]. Kaggle. https://www.kaggle.com/datasets/macedoniale0/arabic-sentiment-analysis-66k-reviews

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. https://scikit-learn.org/