Wasit Journal for Pure Science

Journal Homepage: https://wjps.uowasit.edu.iq/index.php/wjps/index
e-ISSN: 2790-5241 p-ISSN: 2790-5233

Dual-Language Sentiment Analysis: A Comprehensive Evaluating SVM, Logistic Regression, XGBoost, and Decision Tree Using TF-IDF On Arabic and English Dataset

Hawraa Ali Taher *

Department of Computer Science, Faculty of Education for Girls, University of Kufa, Najaf, IRAQ

*Corresponding Author: Hawraa Ali Taher

DOI: https://doi.org/10.31185/wjps.549
Received 10 September 2024; Accepted 01 November 2024; Available online 30 December 2024

ABSTRACT: Sentiment analysis (SA) is a growing area of study that straddles a number of disciplines, including machine learning, data mining, and natural language processing. It is focused on the automatic extraction of viewpoints presented in a certain text. Many studies have been conducted in the area of sentiment analysis because to its broad uses, particularly on texts in English, whereas other languages like Arabic have gotten less attention. The Arabic language presents several difficulties, such as its rich morphology and the difficulty of tracing words back to their original roots. Arabic comments have been analyzed and categorized into good and negative attitudes using a framework. With the aim of evaluating any tweet, opinion, purpose, or reputation, such as a university, company, mobile, and others, the research analyzes the comments made by users of the social networking site Twitter. It does this by using classification technology and machine learning, which are among the fundamental tasks of the data mining process used in the larger process, which is to explore knowledge. This search helps the user to access the evaluation of other users through their tweets and comments on the social networking site for an opinion immediately and automatically, and then the process of uploading and evaluating the opinions using appropriate algorithms for this purpose as (Decision Tree classifier DTC, XGboost, Logistic Regression LR, Support Vector Machine SVM) with Term Frequency-Inverse Document Frequency TF_IDF.

Keywords: Sentiment Analysis, TF-IDF, Arabic Language, English Language, Machine Learning Algorithms

1. INTRODUCTION

User reviews, comments, ratings, and feedback are there as a result of the quick expansion of web and social media applications. These thoughts cover a variety of topics, including goods, politics, news, people, services, and events. To accurately determine what the user is thinking and feeling, these opinions must be processed and assessed. Prior to the development of automated sentiment analysis tools, gathering consumer feedback was a laborious and time-consuming procedure. where opinion analysis has grown in importance in the modern [1].

People tend to spend more time on social media as a result of the spread of crises, whether they are man-made or natural, as social media platforms like Facebook and Twitter turn into a popular source of information because they deliver news more quickly than official news channels and emergency response organizations. It is possible to write thoughts and feelings concerning the news being broadcast through these channels [2].

Arabic's rich morphology makes sentiment analysis (SA) a difficult undertaking because it is an analogue language. When using Twitter data, which is known to be very informal and noisy, the process becomes even more challenging.

Arabic faces difficulties with sentiment analysis. The usage of dialectal Arabic was one of these difficulties (DA). In Arabic, there are two official languages, although the one used in written form differs significantly from the one used in spoken form. The spoken language varies in different Arab nations, resulting in many Arabic strokes. The official language is Modern Standard Arabic (MSA) [3].

For related work, various types of Arabic sentiment analysis methods have been developed by various researchers, for example, Tubishat, M., et al. (2019) suggested two modifications to the WOA algorithm. In the initialization stage of WOA, the first one uses Elite Opposition-Based Learning (EOBL). At the conclusion of each WOA iteration, the second optimization integrates the evolutionary operators from the differential evolution algorithm, including mutation, crossing, and selection factors. In order to limit the search space that WOA explored, we additionally employed Information Gain (IG) as a filter feature selection strategy with a Support Vector Machine (SVM) classifier. Since there aren't as many studies on sentiment analysis for Arabic as there are for English, four Arabic reference datasets are utilized to test for improvement [4].

Also, Ombabi, A. H., et al. (2020) proposed a new deep learning model for sentiment analysis in Arabic on a one-layer CNN architecture to extract local features, and two layers of LSTM to retain long-term dependencies, a new deep learning model for sentiment analysis in Arabic has been suggested. The FastText word embedding form is compatible with this form. Numerous tests were conducted on a multi-domain array. 90.75% is the model's exceptional classification performance. The suggested model is also evaluated using various models and inclusion classifiers [5]. The findings demonstrate that faster and more useful alternatives to Arabic sentiment analysis include the FastText skip-gram model and SVM classifier. In addition, Jihad, & Abdalkafor(2019) Offer numerous techniques for determining a commenter's sentiment. The Arabic language presents several difficulties when dealing with it, including the difficulty of returning a word to its original root and the language's rich morphology. In this study, the difficulties of learning Arabic were also examined, and a framework for classifying comments made in Arabic as positive, negative, or neutral feelings was devised. After the framework has been trained and tested, quotes from his work are then drawn [6].

On the other hand, Alayba, et al. (2018) For Arabic sentiment analysis on various datasets, it was suggested to combine two techniques (CNNs and LSTMs) to see the benefits. Additionally, he attempted to take into account the morphological variety of some Arabic words utilizing various levels of sentiment classification [7].

2. MATERIALS AND METHOD

In this paper, we will apply machine learning techniques on data set Arabic and data set English. In figure (1), that display steps work of the Sentiment Analysis and figure (2) for preprocessing, as following:

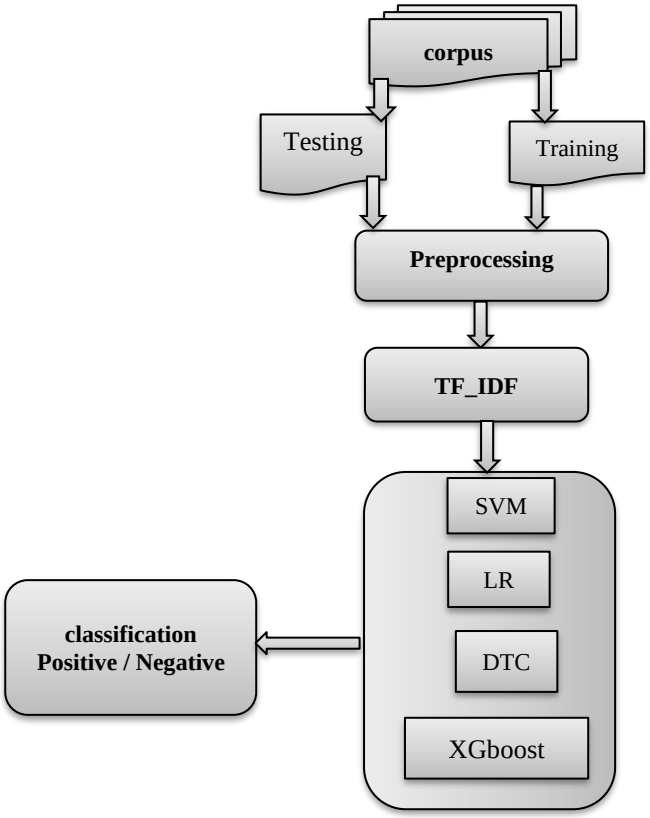

FIGURE 1. -. Generalized structure of Sentiment Analysis for (Arabic& English)

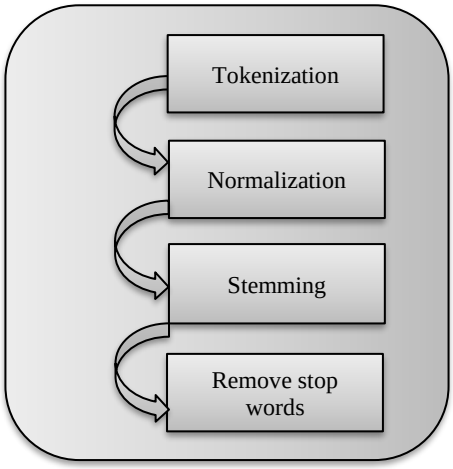

FIGURE 2. -.. Preprocessing

Algorithm of the Sentiment Analysis (Arabic & English)

Algorithm A:
Input :T//Text
Output: positive or negative
Step1:preprocessing**:**

- *Tokenization*
- *normalize*
- *Remove stopsword or with stop words*

- *Stemming (Get all word roots)*

Step2: TF_IDF
Step3: input the text to machine learning technique
Step4: return the sentiment analysis (P or N)

As following explain the steps for the system to work (Sentiment Analysis):

2.1 : preprocessing

Data preprocessing in sentiment analysis improves model accuracy by cleaning text of noise, reducing data size by removing unnecessary words, and standardizing formats. These steps reduce bias, increase analysis speed, and make the model more efficient and accurate in understanding sentiments. As following:

a) Tokenization: is an essential step in natural language processing. Tokenization is a crucial element in many studies because it is the first step in the processing of the text. [8].

b) Normalize: is the process of standardizing the letters and restoring a token to its original form. The inflectional form is removed, in normalization process to obtain the basic form.

c) stop words removal: It is the process of getting rid of any repetitions, like prepositions, that don't change the meaning of the sentence. In our search, stop words have been eliminated from all sites in the text [9].

d) Stemming :To solve the issue of lexical mismatch, it is the process of decreasing inflectional words and bringing them back to their origin, basic form, or root of the written word in general [10].

2.2 : Term Frequency-Inverse Document Frequency (TF_IDF)

a statistical method for determining the significance of a word for a document or category inside of a collection or set of files. The main rule is that if a term or phrase appears frequently in one article but infrequently in others, it is thought to have strong class-distinguishability and is appropriate for categorization. It is a function of choosing the weight of the most frequently used terms in the current vector space model. Its two fundamental elements are the repetition of words and the repeating of sentences with the directions reversed. The number of times a certain word appears in the file is known as its word frequency. A word's total relevance is determined by the frequency of the reverse coil. The number of documents containing the phrase is divided by the repetition of the word's inverse, which is then multiplied by the total number of documents, and the resulting logarithmic product. The following are the formulas for word frequency (TF) and inverse text frequency (IDF)[11][12]:

$$\text{tf ij} = \frac{n_{ij}}{\sum_k n_{k,j}} \quad (1)$$

$$\text{idfi} = log\ \frac{|D|}{|[\text{j} : t_i \in d_j} \quad (2)$$

$$\text{tfidf} = \frac{tf_{ij} * idf_i}{\sqrt{\sum_{t_i \in d_j} [tf_{ij} * idf_i]^2}} \quad (3)$$

2.3: Machine Learning Techniques:

The following are the techniques that were applied for sentiment analysis:

2.3.1. Support Vector Machine:

SVC is a member of the family of multifunctional schemas (SVM) called Support Vector Machines (SVM). Based on the idea of statistical learning theory, SVM is a new generation of supervised learning system. SVM works for classification by locating a super level in the potential input space. Determining the hyperplane that results in a complete separation between the two classes is the fundamental strategy used in support vector machines. In SVM, the superplane is often constructed using a subset of data called the training dataset, and its generalizability is verified using a separate subset of data called the test dataset. A high-dimensional plane (N - 1) is constructed in the SVC to categorize N-dimensional data. It is noticed that there can be an endless number of hyperlevels that can separate this data in the binary case of linearly separable data (Figure 4). The maximum margin is only present in one hyper level,

though. This is the super-optimal plane, and the support vectors are the vectors (points) that limit the margin's width. Figure 5 shows a geometric representation of the SVM margin. [13][14]. As following figure (3) show Geometrical Representation of the SVM Margin.

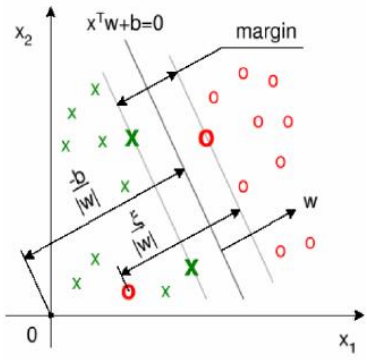

FIGURE 3. Geometrical Representation of the SVM Margin

Mathematical representation to be in the following form:

$$W^T . x = 0 \quad (4)$$

where w, x represents vectors, and vector w is usually called a weight vector. The training data can be represented in the following mathematical form:

$$\{(x_1 , y_1), (x_2, y_2), \ldots . (x_n , y_n)\} \in R^n$$

And the classification function will be in the following mathematical form:

$$y = f(x): R^n \rightarrow \{1, -1\}$$

The process of training will take place to find the maximum amount of the margin and represented in the following mathematical form:

$$m = \frac{2}{||w||} \quad (5)$$

Where we notice the inverse relationship between weights vector and the margin, the less the weight, the greater the amount of the margin. The cost function equation for SVM classifier:

$$min\, C \sum_{1=i}^{m} [y^{(i)}\, cost_1(\theta^T x^{(i)} + (1 - y^{(i)}) cost_0(\theta^T x^{(i)})] + \frac{1}{2} \sum_{i=1}^{n} \theta_j^2 \quad (6)$$

$$min(w, b) \frac{1}{2} ||W||^2 \; subject\; to\; y_i(w.x + b) \geq 1 \; for\; any\; i = 1, \ldots n \quad (7)$$

2.3.2 Logistic Regression (LR):

Logistic regression is a component of the supervised classification strategy. This algorithm's importance and usage have increased over the past few years. With this approach, individuals are divided into groups based on the logistic function. This model is used to evaluate the statistical importance of each independent variable in relation to probability. It is an effective technique for simulating the binomial outcome. [15]. Figure (4) show Logistic Regression:

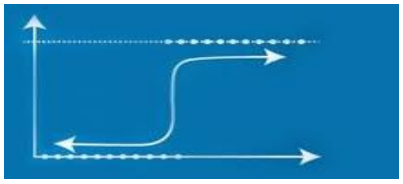

FIGURE 4. Logistic Regression

$$h(x_i) = \beta^T x_i \quad (8)$$

where $h(x_i)$ hypothesis, B is regression coefficient, x is data. The output will be in the form of the probability that the values are confined (0,1) and we apply the sigmoid function its mathematical formula:

$$h(x_i) = \beta^T x_i = \frac{1}{1 + e^{-\beta^T x_i}} \quad (9)$$

$$P(y_{i=1} | x_i; \beta) = h(x_i) \quad (10)$$

$$P(y_{i=0} | x_i; \beta) = 1 - h(x_i) \quad (11)$$

The two equations (11,12), are summarized by the following equation:

$$P(y_i | x_i; \beta) = (h(x_i))^{y_i} (1 - h(x_i))^{1-y_{i-1}} \quad (12)$$

the cost function for logistic regression is maximum likelihood estimation (MLE):

$$l(\beta) = \sum_{i=1}^{n} -y_i \log(h(x_i)) - (1 - y_i) \log(1 - h(x_i)) \quad (13)$$

To find the best value B that makes our error amount as little as possible in order for the prediction process to be accurate, we use the optimization algorithm. The most famous optimization algorithm that is used with logistic regression is Gradient Descent Algorithm:

$$\beta_j = \beta_j - \alpha \sum_{i=1}^{n} (h(x_i) - y_i) x_{ij} \quad (14)$$

2.3.3. Decision Tree Classifier (DTC):

A decision tree classifier is represented as a binary tree, where each leaf is rated 0 or 1 and each interior node is classified by a variable. The longest path from root to leaf of the decision tree is at its depth. Benefit: If the size of the tree is small, DT is an effective and potent method for filtering data. It is simple to comprehend and articulate. Decision tree models can be understood by everyone with a simple explanation. Disadvantage: When it comes to data upkeep, DT is a challenging and complex procedure. Because of its instability, the ideal decision tree structure could significantly vary with even a little change in the data [16]. Figure (5) show Decision Tree Classifier:

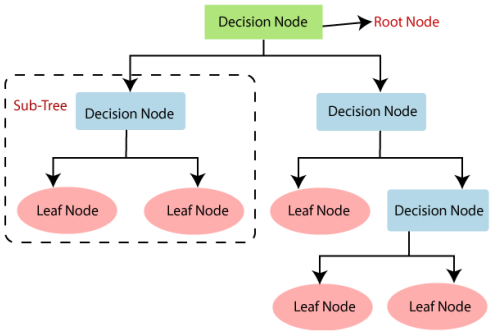

FIGURE 5. Decision Tree Classifier

The main problem arises which is how to choose the best attribute for root node and child nodes during the implementation of the decision tree. Therefore, to solve such problems there is a technique called Attribute Selection Measure or ASM. With this analogy, we can easily determine the best attribute for the tree nodes. There are two common methods of ASM, which are (Information Gain, Gini Index).

$$Information\ Gain = Entropy(S) - [(Weighted\ Avg) * Entropy(each\ feature) \quad (15)$$

$$Entropy(s) = -P(yes)log2\ P(yes) - P(no)log2\ P(no) \quad (16)$$

Where, S= Total number of samples, P(yes)= probability of yes, P(no)= probability of no.

$$Gini\ Index = 1 - \sum_j P_j^2 \quad (17)$$

2.3.4. XGBoost:

Extreme gradient boosting, or XGboost, is employed for sentiment embeddings classification. A more intricate version of the gradient boosting strategy is called XGboost. It comes with a linear model solver and tree learning algorithms. Its capacity to carry out parallel processing on a single processor account for its speed. It also includes cross-validation and critical variable detection tools. The model needs to have a few parameters changed in order to be optimal. Some of the main benefits of XGboost are as follows: Reducing overfitting is facilitated by regularization.

 Parallel processing: XGboost employs a much faster parallel processing system.

 Missing values: An integrated procedure is available to address missing values.

 Inbuilt cross-validation: this feature enables the user to carry out cross-validation at every stage of the boosting procedure. [16][17][18].

Algorithm of XGBoost for Sentiment Analysis:

Step1: Feature Extraction by TF-IDF

Setp2: Define the Loss Function:

$$L(y, \hat{y}) = -\sum_{i=1}^{n} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (18)$$

Step3: Building successive Trees:

- Calculate Derivations:

  -First derivation (gradient):

$$g_i = \frac{\partial L(y_i, \hat{y}_i)}{\partial \hat{y}_i} \quad (19)$$

  -Second derivation (hessian)

$$h_i = \frac{\partial^2 L(y_i, \hat{y}_i)}{\partial \hat{y}_i^2} \quad (20)$$

- Build a New Tree using the derivation to enhance the current model. The new tree aims to reduce the loss

Step4: Model Update

- Determine leaf weight

$$w_i = \frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \quad (21)$$

- Update Prediction

$$y_i^{(t+1)} = \hat{y}_i^{(t)} + \eta w_j \quad (22)$$

Step5: Model Regularization:

$$\Omega(f_t) = T + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2 \qquad (23)$$

Step6: Iterate Until Convergence:

Step7: Final Prediction:

$$\hat{y}_i = \sum_{t=0}^{T} f_t(x_i) \qquad (24)$$

3. THE RESULTS

These experiments aimed to sentiment analysis by used classifiers as (SVM, DT, XGboost, LR), also measured the accuracy each of the classifier before preprocessing, and with preprocessing with Arabic data have (1800 twitters), and English data have (748 twitters). The Following equations for accuracy:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \qquad (25)$$

$$Recall = \frac{Number\ \ of\ true\ positive\ samples}{Number\ \ of actual\ \ positive\ samples} \qquad (26)$$

$$precision\ = \frac{Number\ \ of\ true\ positive\ samples}{Number\ \ of predicted\ \ positive\ samples} \qquad (27)$$

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall} \qquad (28)$$

Table 1. - Results accuracy of All Algorithms for Arabic Sentiment Analysis (Without preprocessing)

	SVM	LR	DT	XGBoost
Precision	0.82	0.84	0.77	0.73
Recall	0.82	0.83	0.77	0.89
F1_score	0.82	0.84	0.77	0.80

Table 2. - Results accuracy of All Algorithms for Arabic Sentiment Analysis (With preprocessing)

	SVM	LR	DT	XGBoost
Precision	0.82	0.84	0.77	0.71
Recall	0.82	0.84	0.77	0.87
F1_scor	0.82	0.84	0.77	0.78

Table 3. - Results accuracy of All Algorithms for Arabic Sentiment Analysis (With Preprocessing & TF-IDF)

	SVM	LR	DT	XGBoost
Precision	0.84	0.83	0.76	0.70
Recall	0.84	0.83	0.76	0.84
F1_score	0.84	0.83	0.76	0.77

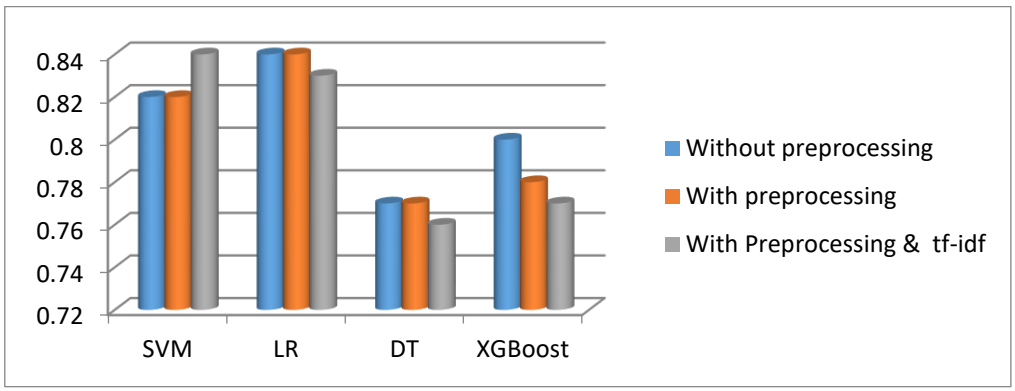

FIGURE 6. Results of Accuracy for Arabic Language

Table 4. - Results accuracy of All Algorithms for English Sentiment Analysis (Without preprocessing)

	SVM	LR	DT	XGBoost
Precision	0.74	0.74	0.63	0.67
Recall	0.75	0.73	0.67	0.75
F1_score	0.75	0.73	0.65	0.71

Table 5. - Results accuracy of All Algorithms for English Sentiment Analysis (With preprocessing)

	SVM	LR	DT	XGBoost
Precision	0.66	0.65	0.61	0.72
Recall	0.79	0.82	0.84	0.76
F1_score	0.72	0.72	0.71	0.74

Table 6. - Results accuracy of All Algorithms for English Sentiment Analysis (With Preprocessing & TF-IDF)

	SVM	LR	DT	XGBoost
Precision	0.77	0.75	0.67	0.65
Recall	0.80	0.86	0.75	0.76
F1_score	0.79	0.81	0.71	0.70

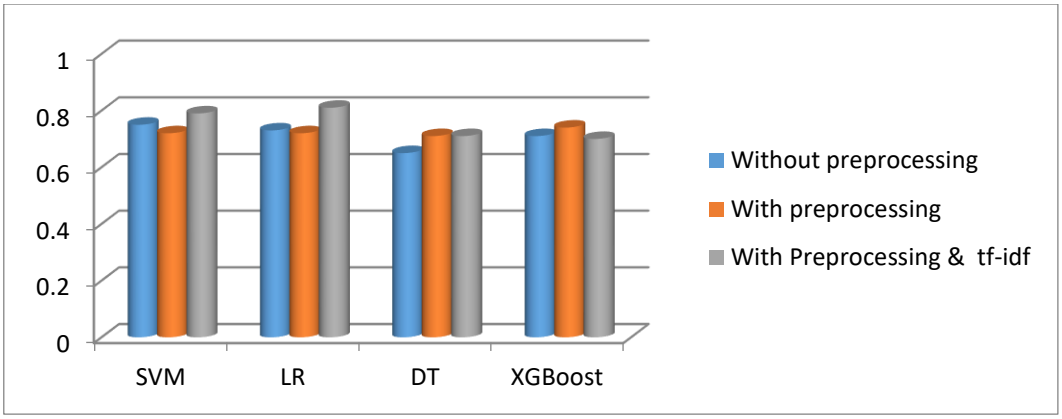

FIGURE 7. Results of Accuracy for English Language

We conclude, that the results are different between machine learning algorithms (logistic regression, support vector machine, decision tree classifier, neural network) for both Arabic and English languages, and the highest accuracy of the algorithms is with preprocessing and with Term Frequency-Inverse Document Frequency (TF_IDF). **TF-IDF** helps improve the accuracy of sentiment analysis by identifying the most important words and filtering out common words, making the data more focused and reducing noise.

4. CONCLUSION

Sentiment analysis is a technique within Natural Language Processing (NLP) that aims to identify and interpret the emotions or sentiments expressed in written texts. This technique classifies texts into categories such as positive, negative, or neutral emotions, with the goal of understanding how the writer or participants in digital conversations feel. Due to the importance of analyzing feelings in all areas of life, a group of tweets was collected and machine learning algorithms were applied to them. This paper has addressed sentiment analysis in Arabic tweets and English tweets. To this end, we applied on 180000 tweets for Arabic Language and 748 tweets for English Language. Preprocessing was applied for Arabic, English tweets and TFIDF for improve the accuracy of sentiment analysis. Four classifiers were used to assess our framework are (SVM, DT, XGboost, LR). The results obtained are promising. The good results, that prove the accuracy of the work, which shows the importance of pre-treatment of tweets, which increases the accuracy of the work. Sentiment analysis can be improved through several future works, including expanding the dataset to include more tweets in different languages and experimenting with advanced deep learning models such as LSTM and BERT. Developing models to analyze mixed sentiments and better understand dialects can also be beneficial. There should also be a focus on analyzing the broader context of tweets to improve result accuracy and enhancing preprocessing operations for the texts. Adding new features such as Word2Vec or GloVe, as well as analyzing topics related to sentiments, can help in gaining a deeper understanding of patterns.

5. ACKNOWLEDGEMENT

The lab and required documents were provided by the Department of Computer Science, Faculty of Education for Girls, University of Kufa, Najaf, Iraq, for which the authors are grateful.

REFERENCES

[1]. Duwairi, R. M., Marji, R., Sha'ban, N., & Rushaidat, S. (2014, April). Sentiment analysis in arabic tweets. In 2014 5th international conference on information and communication systems (ICICS) (pp. 1-6). IEEE.

[2]. Lamsal, R. (2021). Design and analysis of a large-scale COVID-19 tweets dataset. Applied Intelligence, 51(5), 2790-2804.

[3]. Al-Twairesh, N., Al-Khalifa, H., Alsalman, A., & Al-Ohali, Y. (2018). Sentiment analysis of arabic tweets: Feature engineering and a hybrid approach. arXiv preprint arXiv:1805.08533.

[4]. Tubishat, M., Abushariah, M. A., Idris, N., & Aljarah, I. (2019). Improved whale optimization algorithm for feature selection in Arabic sentiment analysis. Applied Intelligence, 49(5), 1688-1707.

[5]. Ombabi, A. H., Ouarda, W., & Alimi, A. M. (2020). Deep learning CNN–LSTM framework for Arabic sentiment analysis using textual information shared in social networks. Social Network Analysis and Mining, 10(1), 1-13.

[6]. Jihad, A. A., & Abdalkafor, A. S. (2019). A Framework for Sentiment Analysis in Arabic Text. Indonesian Journal of Electrical Engineering and Computer Science, 16(3), 1482-1489.

[7]. Alayba, A. M., Palade, V., England, M., & Iqbal, R. (2018, August). A combined CNN and LSTM model for arabic sentiment analysis. In International cross-domain conference for machine learning and knowledge extraction (pp. 179-191). Springer, Cham.

[8]. Attia, M. (2007, June). Arabic tokenization system. In Proceedings of the 2007 workshop on computational approaches to semitic languages: Common issues and resources (pp. 65-72).

[9]. Taher, H. A., Abdulameer, M. H., & Mahdi, B. (2022). Information Retrieval Scheme Via Similarity Technique. International Journal on Technical and Physical Problems of Engineering (IJTPE), (51), 375-379.

[10]. Larkey, L. S., Ballesteros, L., & Connell, M. E. (2007). Light stemming for Arabic information retrieval. In Arabic computational morphology (pp. 221-243). Springer, Dordrecht.

[11]. Liu, C. Z., Sheng, Y. X., Wei, Z. Q., & Yang, Y. Q. (2018, August). Research of text classification based on improved TF-IDF algorithm. In 2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE) (pp. 218-222). IEEE.

[12]. Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF* IDF, LSI and multi-words for text classification. Expert systems with applications, 38(3), 2758-2765.

[13]. Oommen, T., Misra, D., Twarakavi, N. K., Prakash, A., Sahoo, B., & Bandopadhyay, S. (2008). An objective analysis of support vector machine-based classification for remote sensing. Mathematical geosciences, 40(4), 409-424.

[14]. Auria, L., & Moro, R. A. (2008). Support vector machines (SVM) as a technique for solvency analysis.

[15]. Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A comparative analysis of logistic regression, random forest and KNN models for the text classification. Augmented Human Research, 5(1), 1-16.

[16]. Samih, A., Ghadi, A., & Fennan, A. (2023). Enhanced sentiment analysis based on improved word embeddings and XGboost. International Journal of Electrical & Computer Engineering (2088-8708), 13(2).

[17]. Afifah, K., Yulita, I. N., & Sarathan, I. (2021, October). Sentiment analysis on telemedicine app reviews using xgboost classifier. In 2021 International Conference on Artificial Intelligence and Big Data Analytics (pp. 22-27). IEEE.

[18]. Nsaif, A. A., & Abd, D. H. (2022, July). Sentiment analysis of political post classification based on XGBoost. In Proceedings of International Conference on Computing and Communication Networks: ICCCN 2021 (pp. 177-188). Singapore: Springer Nature Singapore.