

# Arabic offensive text classification using emojis: Including emoji data in Arabic natural language processing

Amal Albalawi<sup>1,2</sup>, Wael M. S. Yafooz<sup>1</sup>

<sup>1</sup>Department of Computer Science, College of Computer Science and Engineering, Taibah University, Madinah, Saudi Arabia

<sup>2</sup>Department of Computer Science College of Computer and Cyber Sciences, University of Prince Mughrin, Madinah, Saudi Arabia

## Article Info

### Article history:

Received Apr 21, 2024

Revised Jan 15, 2025

Accepted Mar 3, 2025

### Keywords:

Arabic text classification  
Deep learning emojis analysis  
Machine learning  
Natural language processing  
Offensive language detection

## ABSTRACT

In the digital social media ecosystem, controlling offensive language requires advanced algorithmic tools. This study examines the influence of including emojis translation in the text preprocessing stage of the classification of offensive Arabic text. A novel dataset of 10,000 Arabic tweets was developed, with rigorous annotations to classify content as offensive or non-offensive. The dataset was meticulously annotated and validated using Cohen's kappa (CK) and Krippendorff's Alpha ( $\alpha$ ) to ensure consistency and accuracy. Several experiments evaluated the dataset with the most common text classification models: seven machine learning (ML) classifiers and three deep learning (DL) models. Two experimental sets were conducted: one with emoji translation in preprocessing to enrich text input and another without emoji translation to directly assess the impact of emojis on classification accuracy. The findings indicate that emojis significantly affect text classification models, with advanced DL models showing higher sensitivity to contextual nuances conveyed by emojis compared to traditional ML classifiers. This research highlights the dual role of emojis, which are often linked to positive emotions and offensive contexts, adding complexity to digital communication. It contributes to the development of more accurate and context-sensitive natural language processing (NLP) tools.

*This is an open access article under the [CC BY-SA](#) license.*



## Corresponding Author:

Amal Albalawi

Department of Computer Science, College of Computer Science and Engineering, Taibah University  
Madinah, Saudi Arabia

Email: AmalMohAlbalawi@gmail.com

## 1. INTRODUCTION

Due to the growth of social networks in recent years, people are now more connected to one another, allowing for large-scale, real-time communication on a worldwide scale [1], [2]. This unrestricted communication has facilitated the exchange of diverse ideas, emotions, and information. However, the anonymity offered by these platforms also raises concerns about potential misuse and the spread of offensive language [2]. This has led to an increase in hate speech and cyberbullying on social media platforms [3]. To address this problem, some nations have taken measures to prohibit the proliferation of hate speech on social media platforms. In 2007, Saudi Arabia implemented legal measures to regulate information mediums and technologies used for insulting, defaming, and slandering others, particularly when such actions harm individuals or violate their privacy. This framework addresses the use of contemporary information dissemination technologies in these contexts [4]. In another example, the Network Enforcement Act was enacted in Germany in 2017 [5]. Moreover, offensive language has been the target of ongoing legislative changes, and cutting-edge technical solutions are being investigated to help social media platforms and other organizations to enforce these laws [6]. The literature on offensive language detection in Arabic texts includes a wide range of studies, each of

which contributes unique insights on the preprocessing, analysis, and classification of this content on various digital platforms. A significant limitation has been the lack of robust public datasets for Arabic, which are essential for effective computational analysis [6]. This gap highlights the critical need for robust datasets that can better represent the linguistic complexities of Arabic. Several studies have proposed diverse datasets, starting with small collections, such as the 1,000 tweets analyzed by AlGhamdi and Khan [7] to larger sets, such as the 15,050-comment dataset used by Alakrot *et al.* [8]. These datasets cover multiple languages, including Arabic, and originate from diverse sources such as Twitter, YouTube, and other social media platforms, reflecting the global, multilingual challenges of moderating offensive content AlGhamdi and Khan [7] and Alakrot *et al.* [8]. The preprocessing methods used vary widely in these studies, with techniques such as root-based stemming, n-gram models, light stemming, and lemmatization commonly used to optimize text data for subsequent analysis Al-Saif and Al-Dossari [9] and Founta *et al.* [10]. This initial step is essential for effective model training and the accurate classification of content. In terms of model deployment, several word representation technologies were used, including GloVe, fastText, bag of words (BoW) and term frequency-inverse document frequency (TF-IDF), which facilitate the transfer of text into formats that can be efficiently processed by machine learning (ML) classifiers and deep learning (DL) models. The ultimate goal of these research efforts is to create automated systems that are able to effectively detect and classify various forms of offensive content, such as cyberbullying, hate speech, and other suspicious messages. Some studies suggest enhancing classifier performance by adopting DL techniques or expanding the training dataset, which is evidence of ongoing innovations in this field Fkih *et al.* [11].

A variety of computational techniques have been used for offensive language detection and classification. Among these approaches, both traditional ML algorithms and advanced DL techniques, including transformer models, have proven to be of great benefit. Studies such as Al-Saif and Al-Dossari [9] and AlGhamdi and Khan [7] leveraged traditional ML methods such as support vector machines (SVM) and decision trees to classify different types of Arabic tweets into categories such as hate speech and cybercrime. These approaches highlight the adaptability of ML in dealing with the subtle complexities of digital communication data [7], [9]. On the DL front, Founta *et al.* [10] used a unified architecture that integrates character-level convolutional neural networks (CNNs) and word-level recurrent neural networks (RNNs) to robustly detect offensive content on various social media platforms, demonstrating the depth with which DL models can extract and learn from text data. Significant progress in the application of DL in this field was highlighted by the work of Al-Shaalan and Al-Khalifa [12], who used the bidirectional encoder representations from transformers (BERT) model, a state-of-the-art transformer model. BERT, which stands for bidirectional encoding representation transformers, excels at understanding the full context of a sentence by looking at the text surrounding each word. This is particularly useful for detecting hate speech because the context in which words are used can fundamentally change their meaning [13]. Their approach exemplifies the state-of-the-art capabilities of transformer models in capturing complex linguistic nuances and improving classification accuracy.

The work of Mubarak *et al.* [14] presents an innovative approach to using emojis as anchors to effectively aggregate large amounts of offensive tweets and hate speech in Arabic. This method significantly improves the efficiency of data collection by focusing on tweets that contain specific emojis associated with offensive content. However, it is important to note that emojis can have dual meanings, and their interpretation can largely depend on the context in which they are used. For example, emojis like “rolling on the floor laughing” 🤣 or “heart” ❤️, which are typically associated with positive emotions, can also be used in a sarcastic manner or in a way that implies a specific meaning. This dual use highlights the complexity of emojis use in digital communication.

In light of these challenges, this paper aims to advance the field of offensive language detection in Arabic text. This research evaluates the effectiveness of current methodologies in ML classifiers and DL models for identifying offensive content on the X (Twitter) platform. Additionally, explore the transformation of emojis into text and its impact on detection accuracy. By comparing traditional ML models with advanced DL models, it provides strategic insights for improving content moderation. The contribution of this paper can be summarized as follows:

- a. A novel investigation of how the inclusion of emojis impacts the performance of natural language processing (NLP) models in classifying objectionable content. By incorporating emojis into text analysis, the research explores how these symbols modify the interpretive dynamics of content moderation systems, underscoring the complex role of emojis in digital communication.
- b. Develop a novel Arabic dataset collected from the X (Twitter) platform comprised of 10,000 annotated tweets of offensive content by three Arabic-native speakers.
- c. Evaluate the effectiveness of traditional ML classifiers and DL models on the proposed dataset to measure their performance and ability to detect offensive text.

- d. The research establishes rigorous standards that not only advance the understanding of content recognition techniques but also pave the way for the development of these systems in dealing with complex and culturally varying texts.
- e. A comprehensive comparison between traditional ML methods and DL techniques in classifying offensive content. This analysis provides strategic insights that are important for improving content moderation frameworks and guiding the development of more effective and accurate moderation tools.

The rest of this paper is structured as follows. Section 2 explain the methods and framework of the study. The results and discussion are shown in section 3. Section 4 concludes the paper.

## 2. METHODS AND MATERIALS

This section presents the proposed methodology for this study, emphasizing the model design stage as shown in Figure 1. The proposed model has five key phases: data collection, data pre-processing, model construction, model performance evaluation, and comparison of the results. The methodology is divided into two different paths during the data preprocessing phase to rigorously evaluate the impact of emojis on classification performance. The first path involves processing data using emojis, translated into their textual meanings, with the aim of preserving and analyzing the emotions and information they convey. On the other hand, the second path, involves preprocessing the data with all emojis removed, to serve as a check to evaluate the necessity and impact of including emojis on the model's predictive accuracy. After building the model and evaluating it under these two conditions, the methodology culminates in a comparative analysis of the results. This comparison critically evaluates whether the model's performance is improved or worsened by the inclusion of emojis, providing a clear indication of the importance of emojis in text classification tasks within NLP.

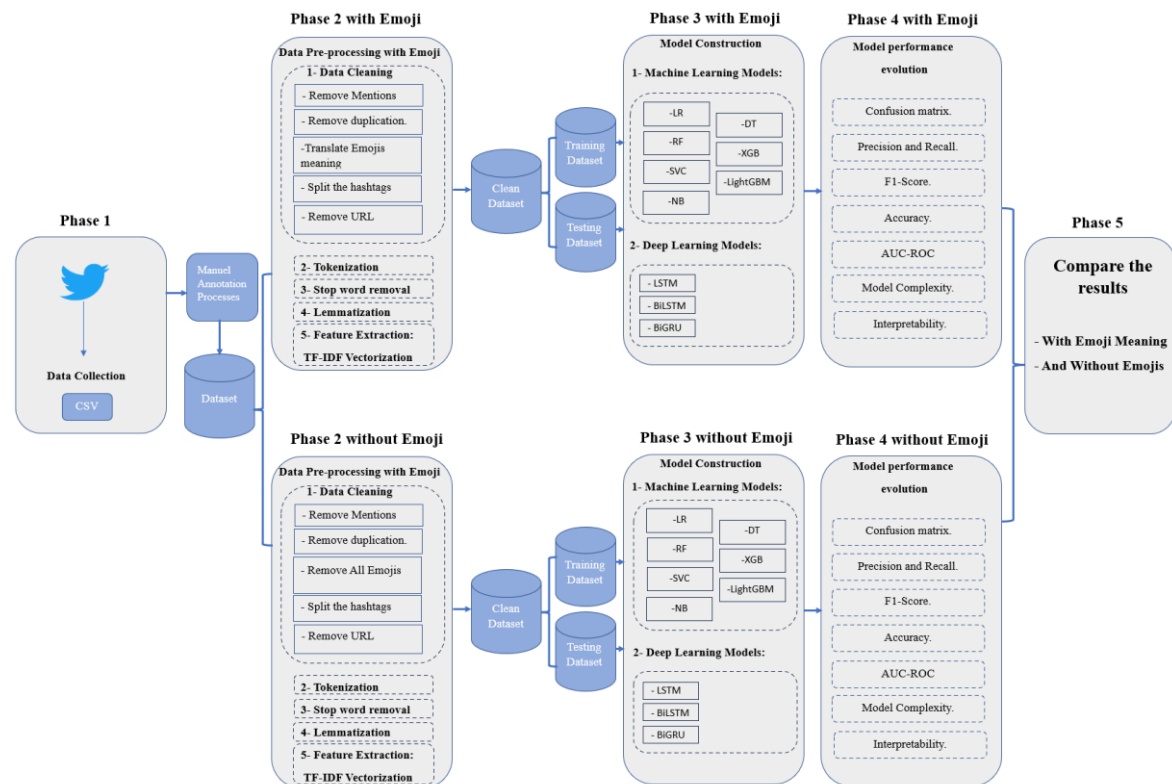


Figure 1. The proposed approach

### 2.1. Dataset creation phase

The novel dataset proposed in this study was collected from the X (Twitter) API. Furthermore, the dataset was manually labelled by three Arabic native speakers. The dataset size is 10,000 Tweets. Table 1 provides a sample list of key terms used in the data collection process, highlighting the various linguistic cues

and expressions that characterize offensive language in Arabic. Table 2 provides representative examples from each category. This table not only serves as a guide to understanding the types of content included within each category but also emphasizes the comprehensive scope of the dataset in covering a wide range of offensive expressions. Through this detailed categorization and provision of examples, the dataset represents a valuable resource for researchers aiming to explore the dynamics of offensive language in Arabic social media, providing insights into the linguistic strategies used in different forms of offensive communication.

Table 1. Example for key words used to collect tweets for the dataset for each type of tweet

Category	Key word in Arabic	Translation of the Key word in English
Hate speech	حشره، أكرهه، استحققت، ابلع خونه،	hate, despise, detest, swallow, bug, traitor
Harassment	غبى، أحمق، حمار، حقير	Stupid, idiot, donkey, despicable
Explicit content	مكانك المطبخ، مغسول دماغك	Your place is in the kitchen, you are brainwashed"
Religious insult	شيعي، سني، وهابي	Shia, Sunni, Wahhabi
Offensive language relating to gender.	مكانك المطبخ، الرجال مالهم فائدة	Your place is in the kitchen, men are useless
Offensive language relating to ethnicity or race.	عبد، فلبيني، سعودي، يمني، مصري	Slave, Filipino, Saudi, Yamani, Egyptian
Body shaming	طويل انت دب، راسك مثلث، قصير،	Fat, triangle head, ugly, short, tall
Classism	خادمي، فقير	Servant, poor
Ageism	شباب بزر	Kid, old man
Ableism	المعاقين عبي، مريض عبي	disabled are a burden; sick are a burden

Table 2. Example for tweets from each category that are considered offensive in the dataset

Category	Tweet (in Arabic)	English translation	Explanation
Hate speech	" يعني انت غبي او اهل "	Means you are stupid or dumb.	Direct insult
Harassment	" ياطاقيه يا اهل... "	You are a dumb hat	Indirect insult that bullies a specific football team
Race	" كل ماهو اسود غبي وأحمق... "	Everything black is stupid and foolish	Direct insult
Ethnicity	" ... لا بارك الله فيكم من فلين "	May God does not bless you, Filipinos	Indirect insult for a specific nationality
Ableism	" المعاقين عقليا وحركيا مساكن لاه يخلوهم أحياء؟ ...مش معيشة عبي.. "	Mentally and physically disabled people are forlorn why do they let them live? They are a burden.	Direct insult
Sexual	" عامل شبة المرأة الشر**طة... "	You are like a bi**h woman	Direct insult
Religious hate	" من تكون شيعي بكل تأكيد ستكون حمار... "	If you are Shia, you are definitely a donkey	Direct insult
Gender	" على المطبخ "	Go to the kitchen.	Indirect insult that means women can't do anything other than cooking
Body Shaming	" اخي كل دب في العالم غبي... "	Every fat person is stupid...	Direct insult
Classism	" ..صحفي وفقير اهل و مرتزق.. "	.... poor journalist, foolish and mercenary...	Direct insult
Ageism	" ..مدرب شباب فاسد.. "	Overage, corrupt coach	Direct insult

The proposed dataset with 10,000 tweets was manually annotated by three human native Arabic speakers. In this study, each tweet was evaluated by three annotators to ensure the robustness of our offensive language detection dataset. To measure inter-annotator agreement, we calculated the CK [15] for each pair, resulting in values of 0.854, 0.885, and 0.882, which indicate a high level of consensus among the annotators. Additionally, we computed the Krippendorff's Alpha [16], obtaining a value of 0.874, further confirming the reliability of our annotations across all three evaluators. This strong agreement underscores the high quality of our dataset, which is essential for developing effective ML classifiers for NLP. By demonstrating consistent annotations, our research lays a solid foundation for algorithms that accurately interpret complex language nuances on social media. The annotation process represents a crucial step in creating a gold-standard corpus. To ensure the integrity and consistency of this process, all annotators adhered to the following guidelines and instructions:

- Annotators were strongly encouraged to approach the annotation process objectively, making sure that their backgrounds or personal prejudices, such as their cultural or religious beliefs, did not interfere with their objectivity. The accuracy and dependability of the dataset depends heavily on this adherence to neutrality.
- The standards for classifying tweets as offensive were well-defined and covered a broad spectrum of topics, including hate speech, cyberbullying, racial/ethnic discrimination, disabilities, sexual orientation, and religion. Annotators had to use their thorough knowledge of inappropriate content to identify tweets that fit into any of these categories.

- c. Annotators were trained not just to identify overtly offensive tweets but also to recognize indirectly offensive tweets by analyzing the overt and covert meanings contained within the tweets. This methodology guarantees that the dataset encompasses the entire range of offensive language, including subtleties and implications that might not be quickly discerned. To aid the annotators in this difficult endeavour, examples of indirectly offensive tweets were given.

Figures 2 and 3 provide examples of the dataset, showing tweets with quotes from the Quran and Arabic Bible, respectively. We considered these tweets as non-offensive in our dataset, which are respected in their respective religious contexts. This classification demonstrates our dedication to contextual sensitivity as well as our understanding of the writings' cultural and religious relevance. By presenting these instances, we hope to draw attention to the careful process used to classify content, making sure that information that is fundamental to cultural and religious debate and is by its very nature respectful, appropriately labeled as non-offensive. Moreover, the dataset consists of different tweets that are expressed in different Arabic dialects but have similar meanings. This highlights the intricacy and nuance involved in accurately classifying and interpreting tweets and serves as an example of the linguistic diversity and rich tapestry of dialects within the Arabic language. The dataset's thorough coverage and the careful consideration given to the nuances of regional variations and colloquial idioms common in social media discourse are highlighted by the inclusion of such samples. This is a crucial factor to take into account when creating sophisticated, and culturally aware analytic systems that can correctly understand and categorize content across the vast Arabic-speaking world.

Non-Offensive قال تعالى ( فمن ثقلت موازينه فأولئك هم المفلحون ومن خفت موازينه فأولئك الذين خسروا أنفسهم في جهنم خالدون ) سورة المؤمنون "111-102"

Figure 2. Non-offensive tweets that have a quote from Quran

Non-Offensive إنجيل يوحنا (16-13) : "لكني أقول لكم الحق، إنه خير لكم أن أطلق؛ لأنه إن لم أطلق لن ياتيكم ببركيت، إن لي أمورا كثيرة أيضا لأقول لكم، ولكن لا تستطيعون أن تحمّلوا الأن، وأما متى جاء ذاك روح الحق، فهو يرشدكم إلى جميع الحق؛ لأنه لا يتكلم من نفسه، بل كل ما يسمع يتكلم به"

Figure 3. Non-offensive tweets that have a quote from Arabic Bible

A Figures 4 and 5 present the most frequently used words in the offensive and non-offensive categories of the dataset, respectively. Figure 4 highlights common words found in offensive tweets, shedding light on specific terms or patterns that frequently appear in objectionable content. Understanding these word distributions can enhance classification algorithms by identifying key indicators of offensive language. Similarly, Figure 5 displays the most used words in the non-offensive category, providing insight into the vocabulary typically found in neutral or appropriate tweets. By comparing the word distributions in both categories, we can better distinguish offensive from non-offensive content, improving filtering and detection techniques.



Figure 4. Offensive data



Figure 5. Non-offensive data

In the analysis shown in Figure 6, a variety of emojis appear within the dataset, highlighting the diversity of the tweets examined. In particular, the most popular emojis in this dataset include several laughing emojis. This occurrence of laughing emojis can likely be attributed to the frequent appearance of these emojis multiple times in individual tweets. Particular emojis which are linked to positive feelings, such as hearts, laughing, and smiling emojis, are among those that appear most frequently in the offensive text. This reveals a certain cultural difference in Arabic tweets in which laughing emojis are often used in sarcastic contexts, even



when the underlying message is offensive. This pattern indicates a culturally distinct application of emojis, underscoring the complexity of interpreting emojis in different cultural backgrounds. Although the initial collection of the proposed dataset did not focus specifically on emojis, their influence was considered in the classification process, indicating the importance of emojis in conveying subtle communication signals within interactions.

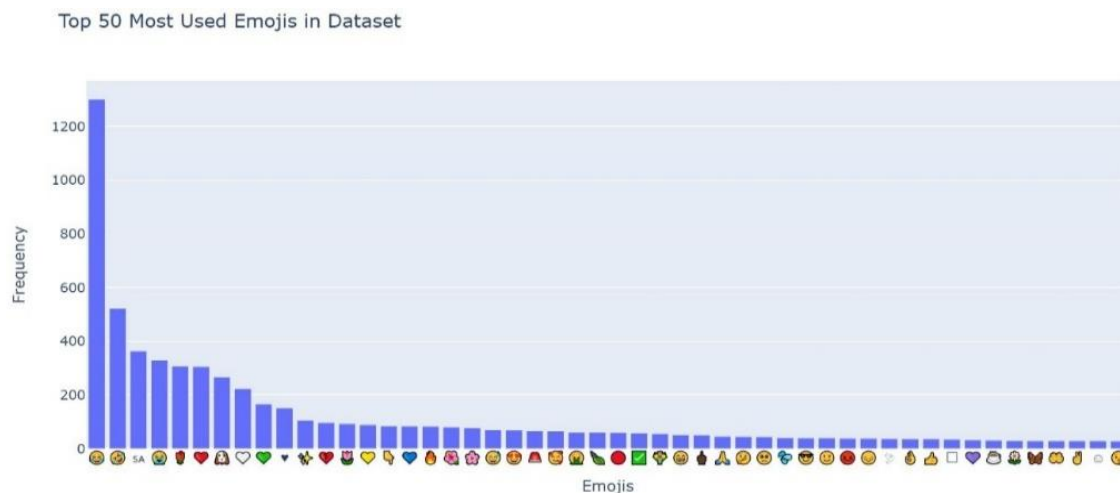


Figure 6. Most used emojis in the dataset

## 2.2. Data pre-processing phase

Cleaning the dataset is a crucial step when dealing with NLP tasks. This study used two methods to clean the dataset. The main difference in these methods is that the first method translated the emojis to include the meaning of the emojis in the tweet as further discussed in the following section. The second method used the same cleaning process but removed all the emojis.

### 2.2.1. Data cleaning

This section explains the data cleaning process implemented in both paths; with emoji acronyms and without emoji acronyms. In the first path, the cleaning process split the hashtags; removed URLs; removed usernames; removed the numbers, special characters and diacritics (Tashkeel); translated emojis' meanings; removed English text; and removed stop words.

This study leveraged a set of emojis identified by Mubarak *et al.* [14] as common in offensive communication. This selection has been carefully curated to capture the types of emojis that are most often used to express negative emotions such as disrespect, anger, or disgust. These often include various animal symbols and other types of symbolism used metaphorically or directly to insult individuals. On this basis, this research not only included these well-known offensive emojis but also expanded the set to include emojis generally considered non-offensive. This broader spectrum of emojis was included to improve the accuracy of our analysis and increase the robustness of our dataset.

The preprocessing phase embarked on translating the semantic content of each emoji into textual representations. This process involved interpreting the meanings of emojis within tweets to ensure that their contextual significance was preserved and accurately reflected in text-based classification tasks. In doing so, we aimed to capture the subtle ways in which emojis contribute to the tone and intent of online communication, thus enriching models' ability to identify and classify tweets more effectively. This method is detailed below and is illustrated through Table 3, which shows an example of how emojis can be transformed into their textual counterparts to enrich contextual analysis in tweet classification.

In the construction of the preprocessing pipeline for this study, a second cleaning process is meticulously designed focusing on the core linguistic components of the analysis. This process includes splitting hashtags and removing URLs, usernames, numbers, special characters, diacritics, emojis, English text, and stop words, thus focusing exclusively on Arabic text elements. This strategy differs from the initial approach by excluding emojis to allow for a clear comparison with the first cleanup, which included emojis. This ensures that the dataset remains focused only on Arabic text, allowing a direct assessment of the impact of different preprocessing strategies on the results of our analysis.

Table 3. Example for cleaning tweets process with emojis

Cleaning steps	Text after Applying the step
Original text	@Adam انا أحب اللعب ❤️ وال gaming في هذه اللعبة <3 www.gamex.com
Split hashtags	@Adam انا أحب اللعب ❤️ وال gaming في هذه اللعبة <3 www.gamex.com
Removed URLs	@Adam انا أحب اللعب ❤️ وال gaming في هذه اللعبة <3
Removed usernames.	انا أحب اللعب ❤️ وال gaming في هذه اللعبة <3
Removed numbers	انا أحب اللعب ❤️ وال gaming في هذه اللعبة <
Removed special characters	انا أحب اللعب ❤️ وال gaming في هذه اللعبة
Removed diacritics (Tashkeel)	انا أحب اللعب ❤️ وال gaming في هذه اللعبة
Translate emojis meaning	انا أحب اللعب حب وال gaming في هذه اللعبة
Removed English	انا أحب اللعب حب وال في هذه اللعبة
Remove stop word	أحب اللعب حب اللعبة
Clean Tweet	أحب اللعب حب اللعبة

### 2.2.2. Tokenization

This section explains the tokenization step used in this study. In the initial stages of preprocessing, encoding was applied to split the text into its component codes, mainly to allow for more detailed analysis. The word tokenize function from the natural language Toolkit (NLTK), a well-recognized instrument in the field of NLP, was used. This step is necessary to convert raw text data into a structured form that can be easily analyzed and processed in the following stages.

### 2.2.3. Stemming

Following the tokenization, stemming was applied to return words to their root form, combining variations of the same word into one representative form. The Arab light stemmer from the NLTK library was used, which is specifically designed to handle the morphological richness of the Arabic language. This emission process is integral to reducing the complexity of the dataset and improving the efficiency of the feature extraction process.

### 2.2.4. Feature extraction phase

The next crucial stage in the preprocessing was feature extraction. Various types of features were explored, leading to a thorough search to identify the most effective combination. The results indicated that the optimal performance was achieved by exclusively utilizing TF-IDF. TF-IDF is a statistical measure used to evaluate the importance of a word in a document relative to a set of documents or texts. This technique highlights words that are repeated within a document but not across documents, identifying several features useful for ML classifiers.

### 2.2.5. Dataset split

An essential part of preparing a dataset for the modeling process is dividing the data into training and test sets for all models used in this study. Allocating 70% of the data to training ensures that models learn from a substantial portion of the dataset, capturing important patterns and features. The remaining 30% is used for testing, allowing an independent evaluation of the models' predictive performance and generalization ability. This separation of data into training and testing sets allows for an effective assessment of the model's performance on unseen inputs.

## 2.3. Model construction phase

The model construction phase of this study took a comprehensive approach, integrating a variety of models including ML classifiers and DL models. This strategic selection of models aims to accurately evaluate and compare the effectiveness of different computational techniques in processing and classifying Arabic tweets, especially in the context of detecting offensive language. The following sections provide an in-depth exploration of the models used in this study. The computational experiments were conducted using a system with an Intel(R) Core(TM) i5-1035G4 CPU running at 1.10 GHz. The system had 8.00 GB of RAM installed and was running a 64-bit version of Windows 10 on a x64 based processor.

### 2.3.1. Machine learning models:

The study initially used traditional ML classifiers for their efficiency, interpretability, and long-term success in various NLP tasks. Logistic regression (LR) is a supervised learning classification algorithm that models the probability of an event using a logistic function based on input features [17]. Support vector classifier (SVC) is a supervised learning approach used for both classification and regression challenges. It works by finding the optimal hyperplane that best separates the data points into different classes. Naive Bayes (NB) is chosen for

its simplicity and speed in making predictions; it assumes feature independence, using it to predict the class with the highest likelihood. NB is commonly applied in text categorization, employing binary or frequency-based vectors to represent data [18]–[20]. Random forest (RF) and decision trees (DT) are clustering methods known for their robustness and ability to handle non-linear data. RF is an ensemble learning method that builds multiple decision trees and aggregates their outputs to improve the accuracy and stability of classification or regression models [21]. DT is a supervised learning technique used for classification and regression tasks. It segments data into subsets based on feature values, constructing a model that resembles a tree structure with labels at the leaf nodes [17]. XGBoost (XGB) is an efficient and scalable implementation of the gradient boosting framework that provides fast and accurate methods for regression, and classification tasks [22]. Light gradient boosting machine (LightGBM) is a fast, distributed gradient boosting framework that uses tree-based learning algorithms; it is optimized for large datasets and highly efficient in computational performance [23].

### 2.3.2. Deep learning models

Advances in DL have vastly improved the applications of NLP, leading to the inclusion of several DL models. Long short-term memory (LSTM) and bidirectional gated recurrent unit network (BiGRU) models are able to capture long-term dependencies in sequential data, a common feature of text [24]. Bidirectional LSTMs: An extension of traditional LSTMs that include forward and backward passes to better understand context and semantics in textual data. BiGRU processes data in both forward and backward directions, enhancing sequence modeling by incorporating the past and future context [25].

## 2.4. Model evaluation phase

The model evaluation phase is a crucial phase of the study and aims to rigorously evaluate the performance of the generated models in different dimensions. This evaluation not only sheds light on the effectiveness of the models in classifying Arabic tweets but also examines the impact of including emojis on their predictive capabilities. To ensure a comprehensive evaluation of the models, the following metrics and aspects were considered: Precision, Recall, F1-Score, area under the receiver operating characteristic curve (AUC-ROC), Model complexity, interpretability, and confusion matrix. The comparative analysis focuses on the performance of the models in two different scenarios: with emojis translation and without emojis as described in the proposed approach. The formula for classification accuracy is the ratio of correct predictions to total predictions as shown below the formulas for accuracy, F-Score, and precision [26], [27].

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (1)$$

$$F1\text{-Score} = 2 \times \frac{\text{precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2)$$

$$\text{Precision} = \frac{TP}{(TP + FP)} \quad (3)$$

$$\text{Recall} = \frac{TP}{(TP + FN)} \quad (4)$$

where in (1)-(4). TP = true positives correctly classified as positive. TN = true negatives: These are instances correctly classified as negative. FP = false positives: These are instances incorrectly classified as positive. FN = false negatives: These are instances incorrectly classified as negative. Validation accuracy and training accuracy metrics describe how well models perform on both training data and unseen validation data, highlighting the effectiveness of learning and the ability to generalize. The AUC-ROC evaluates the model's discrimination ability between classes at different threshold settings, providing an overall measure of performance regardless of class distribution. Model complexity studies the structural complexity of each model, including the number of parameters and computational requirements, which can affect deployment and scalability. Interpretability is the extent to which model decisions can be understood and interpreted, which is important for confidence and applicability in real-world scenarios, especially for sensitive applications. Confusion Matrix is useful in identifying specific strengths and weaknesses of the classification, providing a detailed analysis of model predictions across different categories [28].

## 3. RESULTS AND DISCUSSION

This section presents the experiment results for each model, starting with ML classifiers, followed by DL models. The study explores the classification of Arabic text into offensive and non-offensive categories, augmented with emojis, to investigate the precise role of these symbols in communication. Emojis that are



typically associated with positive emotions often have ambiguous meanings in Arabic texts and may convey negativity or sarcasm. These complexities pose significant challenges to NLP models, which are analyzed across a variety of computational frameworks, including ML classifiers and DL models.

3.1. ML experiments

Traditional ML classifiers’ results are shown in Table 4 with emojis and in Table 5 without emojis. LR and NB show limitations in dealing with the fine details of emojis-enhanced text. The intricate implications of emojis in sentiment analysis are difficult for their underlying algorithms to handle because they are made to evaluate straightforward signals. This limitation is evident in the performance discrepancy observed between texts processed with and without emojis, highlighting the need for models with enhanced contextual awareness. On the other hand, XGBoost, RF, and LightGBM showed significant accuracy and recall in non-emojis scenarios. In particular, XGB showed high accuracy, indicating its effectiveness in scenarios with well-defined feature sets as shown in Figure 7 for the XGB confusion matrix and Figure 8 for XGB AUC-ROC.

Table 4. ML experiment results with emojis acronyms

Model name	Precision	Recall	F1-Score	Accuracy	Training accuracy	AUC-ROC	Model complexity	Interpretability
LR	0.95204	0.966531	0.959235	0.958862	0.981716	0.989566	Low	High
SVC	0.960645	0.965517	0.963075	0.962925	0.99707	0.989253	Medium	Medium
NB	0.882460	0.974645	0.926265	0.92229	0.972701	0.984454	Low	High
RF	0.953861	0.964503	0.959152	0.958862	0.999492	0.990271	High	Medium
DT	0.952763	0.961460	0.957092	0.956830	0.999492	0.957312	High	High
XGB	0.968399	0.963488	0.965937	0.965972	0.981462	0.991874	High	Low to Medium
LightGBM	0.959555	0.962474	0.961012	0.9571	0.960893	0.991383	High	Low to Medium

Table 5. ML experiment results without emojis acronyms

Model Name	Precision	Recall	F1-Score	Accuracy	Training Accuracy	AUC-ROC	Model Complexity	Interpretability
LR	0.952048	0.96653	0.95923	0.958862	0.981717	0.989566	Low	High
SVC	0.960646	0.965517	0.96307		0.962925	0.997080	0.989253	Medium
NB	0.892857	0.974240	0.93177		0.928096	0.979449	0.987908	Low
RF	0.969496	0.965654	0.96757	0.967377	0.998716	0.989710	High	Medium
DT	0.964215	0.961030	0.96262	0.962383	0.998858	0.963334	High	High
XGB	0.970822	0.966975	0.96889	0.968708	0.979877	0.993604	High	Low to Medium
LightGBM	0.966158	0.961691	0.96391	0.963715	0.985443	0.993214	High	Low to Medium

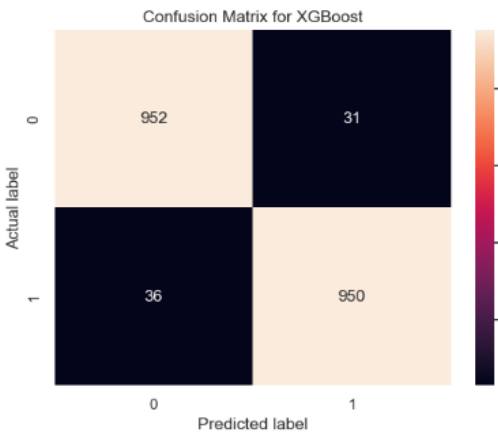


Figure 7. XGB confusion matrix

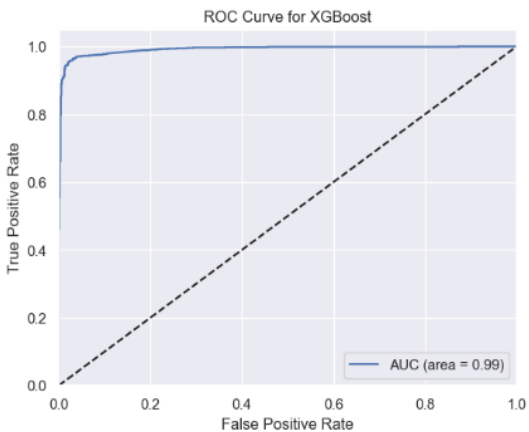


Figure 8. AUC-ROC for XGB

3.2. DL experiments results

This section discusses the results for the DL model. This study considers three state-of-the-art architectures: LSTM, BiLSTM, and BiGRU. The construction of the models is the same in both experiments: using emojis and without emojis. Table 6 provides the model in detail.

Each model was evaluated based on the number of layers, neurons, batch size, dropout, learning rate, precision, recall, F1-Score, AUC-ROC, model complexity, and interpretability. The models' performance results with emoji acronyms are provided in Table 7, and the models' results without emoji acronyms are provided in Table 8. These architectures excel at capturing long-term dependencies and show high sensitivity to contextual and emotional cues provided by emojis. As evidenced by their continually better performance in texts with a high emoji content, these models frequently successfully identify the nuanced emotional undertones present in emojis. This underlines the important role of emojis in conveying complex emotional states, especially in a linguistically rich language like Arabic.

BiLSTM demonstrated the highest levels of accuracy and excellent overall performance across all metrics in DL models. Figure 9 shows the training and validation accuracy, in addition to the training and validation loss for BiLSTM with emojis. Also, Figure 10 shows the training and validation accuracy, in addition to the training and validation loss for BiLSTM but without emojis.

Table 6. Architectural details of DL models used across both experiments

Model name	Number of layers	Neurons	Batch size	Dropout	Learning rate	Epoch	Optimizer
LSTM	2	128	32	0.5	0.0001	10	Adam
BiLSTM	2	128	32	Spatial: 0.3, BiLSTM: 0.25, Additional: 0.5	0.0005	10	Adam
BiGRU	2	64	32	0.3	Default of Adam	10	Adam

Table 7. DL experiment results with emojis acronyms

Model Name	Precision	Recall	F1-Score	Validation accuracy	Training accuracy	AUC-ROC	Model complexity	Interpretability
LSTM	0.9631	0.9568	0.9599	0.9587	0.9935	0.9587	High	Low
BiLSTM	0.9731	0.9502	0.9615	0.9607	0.9911	0.9611	High	Low
BiGRU	0.9462	0.9418	0.9440	0.9434	0.9803	0.9434	High	Low

Table 8. DL experiment results without emojis acronyms

Model Name	Precision	Recall	F1-Score	Validation accuracy	Training accuracy	AUC-ROC	Model complexity	Interpretability
LSTM	0.9703	0.9486	0.9593	0.9593	0.9913	0.95950	High	Low
BiLSTM	0.9593	0.9644	0.9618	0.9613	0.9971	0.96135	High	Low
BiGRU	0.9361	0.9511	0.9435	0.9424	0.9898	0.94234	High	Low

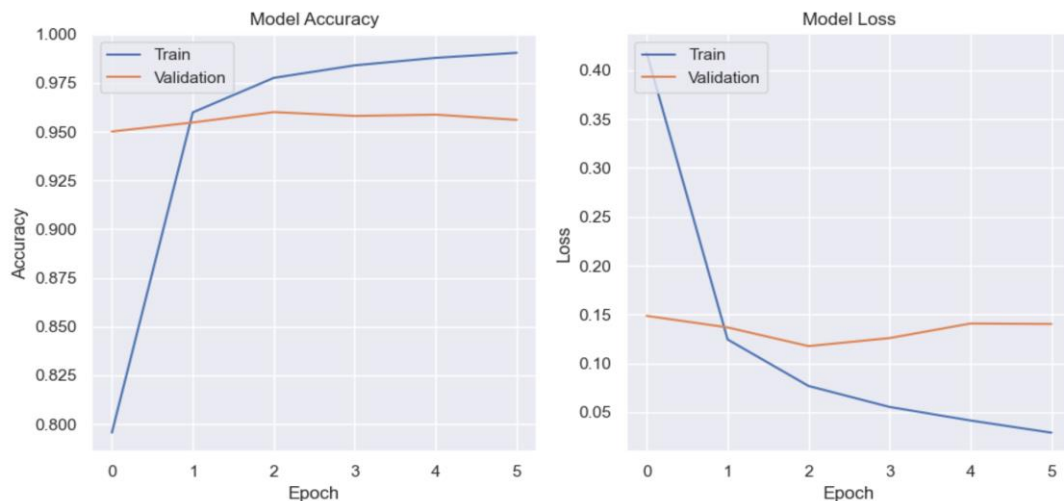


Figure 9. Test accuracy vs validation accuracy and test loss vs validation loss for BiLSTM model without emojis

### 3.3. Discussion

ML tree-based classifiers, especially XGBoost (XGB), performed well, particularly in the absence of emojis. XGB outperformed both DL and transformer models in terms of accuracy in this dataset. This

suggests that XGB was effective at classifying offensive text based on well-defined feature sets, even without the additional complexity introduced by emojis. This raises an important question: Why did XGB outperform more complex models like transformers?

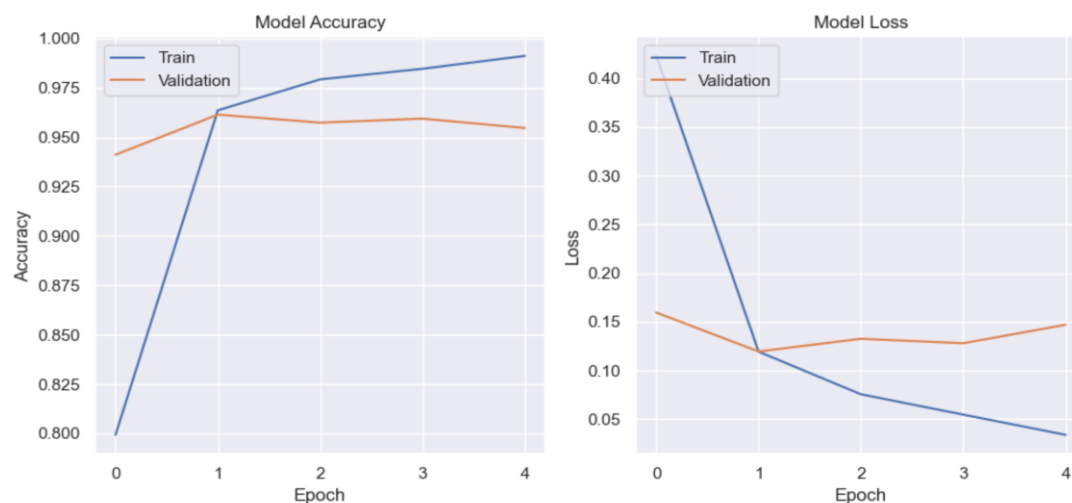


Figure 10. Test accuracy vs validation accuracy and test loss vs validation loss for BiLSTM model

There are three possible explanation is that XGB's reliance on feature engineering and decision trees may allow it to focus more directly on specific features without the noise introduced by emojis. However, despite this superior performance in some cases, XGB lacks the deeper contextual understanding that models like transformers offer. In this context, interpretability becomes crucial. XGB's advantage in terms of interpretability allows for a clearer understanding of which features drive decisions. This invites further investigation into how emojis could be better incorporated into ML pipelines. While this study used text representations of emojis, future research could explore the use of emoji embeddings or even more advanced processing techniques that capture the cultural nuances of emojis in Arabic text.

In addition, the size of the dataset plays an important role in determining the performance of a DL model. In a comparative context, small datasets have been shown to limit the effectiveness of DL methods. For example, AlGhamdi and Khan [7] found that with only 1,555 tweets, ML classifiers such as SVM performed better than LSTM, which was affected by the limited scale of the dataset and the nature of binary classification. Similarly, a study by Mubarak *et al.* [14] with a dataset size of 12,698 SVM outperformed some advanced transformer models such as MBERT and XLM-RoBERTa. Additional studies reinforce the impact of dataset size on model performance: RF and DT outperformed NN in a dataset of 6,964 tweets in Fkih *et al.* [11]. In another scenario including 9,316 tweets in Alshalan and Al-Khalifa [12], SVM outperformed a BERT model. In contrast, a study by Omar *et al.* [28] showed that a DL model, RNN, outperforms ML classifiers on a very large dataset of 20,000 social media posts, highlighting the scalability and learnability of DL under sufficient data conditions.

Moreover, this research utilized uniform preprocessing across all experiments to ensure a fair comparison. In the preprocessing phase, we employed an Arabic light stemmer. This approach might impact the transformers and reduce their capabilities. Transformers rely heavily on contextual integration to understand the complex meanings of words within sentences. By reducing words to their roots, stemming can eliminate important morphological details and subtle nuances that are essential to understanding deep context. This reduction can undermine the model's ability to accurately capture the full range of linguistic information necessary to accurately understand and generate text [29]. Future research could explore more sophisticated methods for integrating emojis, expand the dataset size, and test across a wider range of transformer model variants.

### 3.4. Error analysis

This section presents the error analysis for the best model. False negatives occur when offensive tweets are classified as non-offensive, and false positives occur when non-offensive tweets are classified as

offensive. Key challenges, as shown in Table 9, include implicit offensive content, a lack of contextual data, cultural expressions, keyword misunderstandings, and sarcasm. Insults combined with positive emojis also pose significant difficulties, as positive elements can mask negative undertones.

Table 9. Example of common error type in offensive tweets detection

Class	Error type	Example
False negative	Cultural nuances and indirect insult.	"مكانك المطبخ وليس الفحول" (Classified as non-offensive)
	Different dialect	"عمر دفع الله ده زول يقول ليهو انت اهيل زمان كر كثير غبي" (Classified as non-offensive)
	Cultural background	"يلعن ابو شبيتك شايب عايب" (Classified as non-offensive)
	Cultural background and lack of contextual data.	"هذا مقدارك" (Classified as non-offensive)
False positive	Implicit insult and positive emoji	"المطبخ مكان طاهر ع الاقل انت مكانك تدري وينه اكيد" (Classified as non-offensive)
	Incorrect flagging due to presence of specific keywords (صاحي) often misinterpreted.	"بيتنا كله صاحي اذان الفجر ننتظر شروق الشمس" (Classified as offensive)
	Terms of cultural expression of concern and well wishes such as "تبطي سنه" or "يشفي والدك" can be confusing.	"مشناق اسال الله يشفي والدك العالي والحبيب كحيلان تبطي سنه والله يكثر خير خيره الخير" (Classified as offensive)
	Non-human target	"يا ابو عزوز" (Classified as offensive)
	Misunderstanding of the clinical or educational context.	"احلي شعور عندي اخلص تنظيف دروج المطبخ الداخل احسن براحه مو طبيعيه وحاليا" (Classified as offensive)
		"انا داخله اكتب لانها مركبه وحسبي الله المواعين" (Classified as offensive)
		"ثريد سمعت الهوس الجنسي تعرف اسبابه وكيفيه علاجه تعالوا نتعرف الاجابات" (Classified as offensive)

#### 4. CONCLUSION

Effectively managing offensive language on social media is a difficult task. In this study, we examine the effect of emoji translation during text preprocessing on the classification of offensive Arabic tweets, using a carefully annotated dataset of 10,000 tweets. Our findings suggest that emojis play an important role in text categorization. This research makes several key contributions. It provides new insights into how emojis affect the performance of NLP models in detecting offensive content. We developed a new Arabic dataset of offensive tweets, which were annotated by three Arabic speakers to ensure cultural relevance and accuracy. The study establishes strict criteria for addressing complex and culturally diverse texts in NLP tasks. Furthermore, it provides a comprehensive comparison of traditional ML classifiers and DL techniques in offensive content classification and evaluates their performance on the newly proposed dataset. By providing a detailed dataset and comprehensive analysis, this study lays the foundation for improving content moderation tools. It also suggests avenues for future research into developing more context-aware NLP models capable of capturing the cultural subtleties and nuances inherent in digital communication.

#### FUNDING INFORMATION

The authors state that no funding was received for this work.

#### AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Amal Albalawi	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Wael M. S. Yafooz	✓	✓		✓	✓	✓	✓		✓	✓	✓	✓		

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

#### CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.




## DATA AVAILABILITY

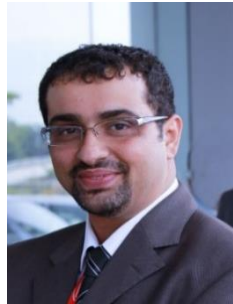
The data supporting the findings of this study can be obtained upon request from the corresponding author, AA.




## REFERENCES

- [1] A. E. Schlosser, "Self-disclosure versus self-presentation on social media," *Current Opinion in Psychology*, vol. 31, pp. 1–6, Feb. 2020, doi: 10.1016/j.copsyc.2019.06.025.
- [2] K. Cortis and B. Davis, "Over a decade of social opinion mining: a systematic review," *Artificial Intelligence Review*, vol. 54, no. 7, pp. 4873–4965, Oct. 2021, doi: 10.1007/s10462-021-10030-2.
- [3] W. M. S. Yafooz, A. Al-Dhaqm, and A. Alsaedi, "Detecting kids cyberbullying using transfer learning approach: Transformer fine-tuning models," in *Kids Cybersecurity Using Computational Intelligence Techniques*, 2023, pp. 255–267. doi: 10.1007/978-3-031-21199-7\_18.
- [4] "Anti-cyber crime law," *The Saudi Board of Experts at the Council of Ministers*. <https://laws.boe.gov.sa/BoeLaws/Laws/LawDetails/25df73d6-0f49-4dc5-b010-a9a700f2ec1d/2> (accessed Apr 21, 2024).
- [5] J. Ruppendorfer, M. Siegel, and M. Wiegand, *Konvens 2018 - GermEval proceedings*. Verlag der Österreichischen Akademie der Wissenschaften, 2018. doi: 10.1553/0x003a105d.
- [6] H. Elzayady, M. S. Mohamed, K. M. Badran, and G. I. Salama, "Detecting Arabic textual threats in social media using artificial intelligence: An overview," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 25, no. 3, pp. 1712–1722, Mar. 2022, doi: 10.11591/ijeecs.v25.i3.pp1712-1722.
- [7] M. A. AlGhamdi and M. A. Khan, "Intelligent analysis of Arabic tweets for detection of suspicious messages," *Arabian Journal for Science and Engineering*, vol. 45, no. 8, pp. 6021–6032, Aug. 2020, doi: 10.1007/s13369-020-04447-0.
- [8] A. Alakrot, L. Murray, and N. S. Nikolov, "Dataset construction for the detection of anti-social behaviour in online communication in Arabic," *Procedia Computer Science*, vol. 142, pp. 174–181, 2018, doi: 10.1016/j.procs.2018.10.473.
- [9] H. Al-Saif and H. Al-Dossari, "Detecting and classifying crimes from arabic twitter posts using text mining techniques," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 10, pp. 377–387, 2018, doi: 10.14569/IJACSA.2018.091046.
- [10] A. M. Founta, D. Chatzakou, N. Kourtellis, J. Blackburn, A. Vakali, and I. Leontiadis, "A unified deep learning architecture for abuse detection," in *Proceedings of the 10th ACM Conference on Web Science*, Jun. 2019, pp. 105–114. doi: 10.1145/3292522.3326028.
- [11] F. Fkih, T. Moulahi, and A. Alabdulatif, "Machine learning model for offensive speech detection in online social networks slang content," *WSEAS Transactions on Information Science and Applications*, vol. 20, pp. 7–15, Jan. 2023, doi: 10.37394/23209.2023.20.2.
- [12] R. Alshalan and H. Al-Khalifa, "A deep learning approach for automatic hate speech detection in the Saudi Twittersphere," *Applied Sciences*, vol. 10, no. 23, Dec. 2020, doi: 10.3390/app10238614.
- [13] W. M. S. Yafooz, "Enhancing Arabic dialect detection on social media: A hybrid model with an attention mechanism," *Information*, vol. 15, no. 6, May 2024, doi: 10.3390/info15060316.
- [14] H. Mubarak, S. Hassan, and S. A. Chowdhury, "Emojis as anchors to detect Arabic offensive language and hate speech," *Natural Language Engineering*, vol. 29, no. 6, pp. 1436–1457, Nov. 2023, doi: 10.1017/S1351324923000402.
- [15] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, Mar. 1977, doi: 10.2307/2529310.
- [16] K. Krippendorff, *Content analysis: An introduction to its methodology*. 2455 Teller Road, Thousand Oaks California 91320: SAGE Publications, Inc., 2019. doi: 10.4135/9781071878781.
- [17] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT Press, 2016.
- [18] V. Korde and C. N. Mahender, "Text classification and classifiers: A survey," *International Journal of Artificial Intelligence & Applications*, vol. 3, no. 2, pp. 85–99, Mar. 2012, doi: 10.5121/ijaa.2012.3208.
- [19] A. McCallum and K. Nigam, "A comparison of event models for naive bayes text classification," in *Proceeding AAAI-98 Workshop Learn. Text Categorization*, 1998, pp. 41–48.
- [20] S. Mandala, A. I. Ramadhan, M. Rosalinda, W. M. S. Yafooz, and R. H. Khoar, "DDoS detection by using information gain-Naïve Bayes," in *2022 2nd International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA)*, Dec. 2022, pp. 283–288. doi: 10.1109/ICICyTA57421.2022.10038054.
- [21] H. Mubarak, K. Darwish, and W. Magdy, "Abusive language detection on Arabic social media," in *Proceedings of the First Workshop on Abusive Language Online*, 2017, pp. 52–56. doi: 10.18653/v1/W17-3008.
- [22] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
- [23] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017, pp. 1–9.
- [24] I. H. Sarker, "Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions," *SN Computer Science*, vol. 2, no. 6, Nov. 2021, doi: 10.1007/s42979-021-00815-1.
- [25] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *Prepr. arXiv.1412.3555*, 2014.
- [26] W. M. S. Yafooz, E. A. Hizam, and W. A. Alromema, "Arabic sentiment analysis on chewing khat leaves using machine learning and ensemble methods," *Engineering, Technology & Applied Science Research*, vol. 11, no. 2, pp. 6845–6848, Apr. 2021, doi: 10.48084/etasr.4026.
- [27] J. Jose and D. V. Jose, "Deep learning algorithms for intrusion detection systems in internet of things using CIC-IDS 2017 dataset," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 1, pp. 1134–1141, Feb. 2023, doi: 10.11591/ijece.v13i1.pp1134-1141.
- [28] A. Omar, T. M. Mahmoud, and T. Abd-El-Hafeez, "Comparative performance of machine learning and deep learning algorithms for Arabic hate speech detection in OSNs," in *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020)*, 2020, pp. 247–257. doi: 10.1007/978-3-030-44289-7\_24.
- [29] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

**BIOGRAPHIES OF AUTHORS**

**Amal Albalawi**    is a Lecturer at the College of Computer and Cyber Sciences, University of Prince Mugrin. She holds a Master's degree in Computer Science (First Honors) and a Bachelor's degree in Information Systems from Taibah University. Her research includes web services, data mining, machine learning, and deep learning. She has published multiple papers in leading journals and presented at international conferences. She was honored to receive the Dr. Hussein Al-Sayed Award for Scientific Research, presented under the patronage of Prince Salman bin Sultan during the Excellence Awards ceremony in Medina, in recognition of her contributions to scientific research. She can be contacted at email: AmalMohAlbalawi@gmail.com.



**Wael M. S. Yafooz**    is a professor in the computer Science Department, Taibah University, Saudi Arabia. He received his bachelor's degree in the area of computer science from Egypt in 2002 while a Master of Science in computer Science from the University of MARA Technology (UiTM)- Malaysia 2010 as well as a PhD in Computer Science in 2014 from UiTM. He is an IEEE Senior Member and has obtained the Fellow of the Higher Education Academy (FHEA) recognition. He was awarded many Gold and Silver Medals for his contribution to a local and international expo of innovation and invention in the area of computer science. Besides, he was awarded the Excellent Research Award from UiTM. He served as a member of various committees in many international conferences. Additionally, he chaired IEEE international conferences in Malaysia and China. Besides, He is a volunteer reviewer with different peer-review journals. Moreover, he supervised number of students at the master and PhD levels. Furthermore, He was invited as a speaker in many international conferences held in Bangladesh, Thailand, India, China and Russia. His research interest includes, data mining, machine learning, deep learning, natural language processing, social network analytics and data management. He can be contacted at email: waelmohammed@hotmail.com, wyafouz@taibahu.edu.sa.