International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Improving the Identification of Hate Speech in Arabic Social Media Content Using Emojis Translation

Khadidja Zerrouki [*1], Nadjia Benblidia [2], Omar Boussaid [3]

Submitted: 13/01/2024 **Revised**: 05/05/2024 **Accepted**: 12/05/2024

Abstract: The presence of hate speech on the internet substantially threatens the well-being and safety of individuals using online platforms, hence requiring sophisticated approaches to detect and maintain a constructive atmosphere within social networks. However, extracting information from Arabic text posted on social networking platforms poses considerable challenges. This research paper presents a novel approach that utilizes artificial intelligence techniques to detect instances of hate speech in Arabic-language content disseminated through social media platforms. A supervised deep learning model is developed using the Bi-LSTM (Bidirectional Long Short-Term Memory) architecture and employing Arabic text pre-processing techniques to improve the model's overall performance. The model has undergone training and evaluation using a compilation of four public Arabic datasets containing instances of hate speech, which have been sourced from various social media platforms. The empirical results illustrate that the deep learning model proposed in this study demonstrates exceptional precision, with an accuracy rate of 98.4. The model demonstrates robust generalization skills, efficiently identifying instances of hate speech in Arabic text from several sources with varying degrees of complexity. Moreover, our study provides empirical evidence to support the claim that pre-processing emojis rather than removing them improves the effectiveness of deep learning models in detecting hate speech in Arabic text on social media.

Keywords: *Hate speech; Offensive; Arabic Text pre-processing; Emojis; Deep Learning; Bi-LSTM*

I. INTRODUCTION

The advent of social media platforms has emerged as a prominent and influential phenomenon in contemporary society. Digital platforms act as intermediaries for exchanging information, ideas, and content among users, allowing for the extensive dissemination of messages to a diverse and varied audience. There has been a substantial rise in social media usage, especially during the COVID-19 pandemic, as indicated by the more than 4.95 billion users[1], which accounts for over half of the world's population. Notable examples of social media networks include Facebook, Twitter, Instagram, and YouTube. The advent of social media has led to a substantial transformation in how individuals communicate, share information, and engage with one another. Nevertheless, the extensive spreading of hate speech, offensive language, and other forms of damaging content has been facilitated by the accessibility and anonymity offered by social media platforms.

Hate speech refers to a type of communication or expression that causes fear, encourages prejudice, or provokes violence against persons or groups based on their race, religion, gender, or other identifiable characteristics [1]. The topic being examined encompasses a wide range of manifestations, such as racism, sexism, extremism, and religious intolerance, among others. Furthermore, it can rapidly spread throughout many social media platforms.

A. Background

Due to its complex nature, identifying "hate speech" presents issues for individuals and artificial intelligence systems. The challenge in precisely explaining this concept arises from the widespread need for more clarity regarding the terminology employed to express it. Language terms such as "offensive", "abusive", "toxic", "harmful", or "obscene" contribute to the inherent complexity of the issue, as they often overlap and can be subjectively interpreted [2].

Recognizing and reducing hate speech is a substantial obstacle, as it necessitates a profound comprehension of the intricacies and subtleties of language and the capacity to differentiate between hate speech and the principle of freedom of expression. Numerous studies, approaches, and tools have been developed in order to tackle this difficulty within this particular subject. Previous research has provided evidence of the potential outcomes achieved by machine learning and deep learning algorithms [3]. The creation of this system can be approached through many deployment tactics, such as text classification and sentiment analysis. Additional methods, including natural language processing (NLP), data pre-processing, and feature engineering, can be

[1]*Department of Computer Science, LRDSI Laboratory, Saad Dahlab University, Blida, Algeria*
[2]*Department of Electronics, LRDSI Laboratory, Saad Dahlab University, Blida, Algeria*
[3]*Department of Computer Science and Statistics, ERIC Laboratory, Lumière University Lyon 2, Lyon, French*
** Corresponding Author Email: k_zerrouki@etu.univ-blida.dz*
ORCID Id: 0000-0001-8037-9038

[1] https://techjury.net/blog/how-much-data-is-created-every-day

incorporated to enhance the effectiveness of these processes [4]. The incorporation of Arabic text within social media platforms holds substantial importance due to its widespread usage.

Arabic is commonly recognized as the fourth most frequently utilized language in internet communication, exhibiting a significant concentration on social media platforms across the Middle East and North Africa regions.

The analysis of Arabic content in social media presents several distinct issues [5], which encompass:

- Arabic is an intricate and diverse language with numerous dialects, posing challenges for text mining and natural language processing.
- Script and orthography: Arabic script follows a right-to-left writing system and has several diacritical marks, making automatically extracting and processing textual material difficult.
- Limited availability of annotated corpora: In contrast to languages like English, Arabic has a scarcity of annotated corpora, posing difficulties in training and assessing machine learning models.
- Arabic exhibits a significant level of morphological ambiguity, posing difficulties in precisely determining the source of a word and its intended semantic interpretation.
- Arabic is spoken in numerous countries and areas, each characterized by distinct cultural and linguistic variations. Consequently, developing text classification models that can be universally applied across the Arabic-speaking world poses a significant challenge.

B. Contribution

Notwithstanding the constraints above, recent advancements have been made in Arabic text mining. Numerous methodologies and resources have been devised to tackle these challenges and acquire a more profound comprehension of the perspectives, inclinations, and behaviors of individuals proficient in the Arabic language [6].

The current study aims to provide a significant contribution to the existing literature on this specific subject through the implementation of the following activities:

- The present study involved the development of supervised machine learning and deep learning models to categorize hate speech material. The purpose of these models was to integrate multiple categorization techniques.
- A range of Arabic text pre-processing techniques has been developed to efficiently manage the data and improve the effectiveness of categorization tasks.
- Moreover, we have developed a function "Emojis-To-Test" that translate the meaning of emojis into Arabic text, guaranteeing the retention of essential information while maintaining the data's integrity through the pre-processing phase.

Our study used four preexisting datasets containing Arabic hate speech and offensive language classes. The Bi-LSTM model achieved a maximum accuracy of 98.4% when applied to the OSACT4 dataset [7]. This outcome was achieved by integrating Arabic text pre-processing technologies plus the emoji translation function with our deep learning model.

II. RELATED WORK

Substantial academic research and practical efforts have been made in recent years to identify and acknowledge hate speech and its related themes. A significant cohort of scholars and professionals has devoted their efforts to the utilization and augmentation of machine learning, deep learning, and natural language processing (NLP) methodologies tailored explicitly for examining Arabic textual data. Numerous investigations have been conducted within the particular field.

Albadi et al. [8] developed a deep learning framework to identify posts within the Arabic Twitter community that propagate hostility and aggression towards individuals due to their religious affiliation. The researchers created a dataset and Arabic vocabulary specifically designed to identify hate speech related to religious topics and made them accessible to the public. The researchers developed a series of categorization models utilizing lexicon-based, n-gram-based, and deep-learning methodologies. The research employed a fundamental recurrent neural network (RNN) that included gated recurrent units (GRU) and AraVec pre-trained word embeddings [9] as the most effective paradigm. The approach demonstrated a notable accuracy, reaching 84%, in accurately identifying occurrences of religious hate speech. These findings are the principal endeavor to tackle the issue of identifying instances of religious hate speech on the Arabic Twitter platform.

In their study, Aluru et al. [10] thoroughly evaluated nine different languages. In order to substantiate their findings, the researchers gathered data from a total of sixteen distinct sources. The study also encompassed a comparative evaluation of several methodologies employed to identify hate speech. The study findings indicated that the amalgamation of LASER embedding [11] with logistic regression yielded the most effective models for low-resource languages. On the contrary, empirical evidence demonstrated that multilingual BERT-based models had superior performance in the context of high-resource languages, especially achieving an accuracy rate of 83.65% when applied to Arabic datasets.

A multitask learning (MTL) strategy was employed by Aldjanabi et al. [12] to create a text classification system capable of identifying offensive and hate speech in the Arabic language. In order to properly incorporate contextual representations from global and specialized datasets, the MTL model was trained to utilize cross-corpora data. The development of this training procedure was based on the AraBERT model. The findings demonstrate a notable enhancement, as seen by a precision rate of 95.20% across four distinct datasets about the identification and classification of offensive and prejudiced words in Arabic.

Ousidhoum and Diab [13] provided a novel dataset for hate speech encompassing many languages and features, including Arabic. The researchers evaluated the efficacy of modern multilingual multitask learning methods [14] in analyzing this dataset. The results showcased the dataset's potential to improve the ability to detect and classify hate speech. In addition, the researchers investigated the influence of labeled data on improving the identification and classification of hate speech. The Arabic datasets achieved a macro-F1 score of 84%.

The multitask text classification model developed by Shapiro et al. [15] was explicitly intended to detect hate speech in Arabic tweets. The task that has been assigned consists of three distinct sub-tasks. The main subtask is evaluating unacceptable content inside a given tweet. Once offensive content is detected, the next step is determining if the tweet fits the requirements for being classified as hate speech. Given that the tweet can be accurately classified as hate speech. In order to address the problem of overfitting, the researchers employed transformer models in conjunction with contrastive learning and multitasking learning techniques. In sub-tasks A, B, and C, the proposed solution demonstrated macro-F1-average scores of 0.841, 0.817, and 0.476, respectively.

In [16], Althobaiti built a BERT model to identify abusive language and hate speech in Arabic tweets. They also compared Support Vector Machines (SVM) and logistic regression, two well-established machine learning methods. Additionally, they explored the use of sentiments and textual descriptions of emojis. The findings indicated that using data derived from emojis as a feature marginally enhances the efficacy of the models, with an accuracy of 0.933, the BERT model outperformed the other models in detecting offensive language and fine-grained hate speech from Arabic tweets.

III. MATERIALS AND METHODS

A. Pre-processing

Text pre-processing refers to transforming raw textual input into a format more suitable for analysis or use in machine learning applications [17]. The primary aim of text pre-processing is to produce a coherent and structured representation of textual material that may be effectively employed for text classification. The approaches employed for text preparation in our paper are as follows:

a) Text cleaning: The process involves removing extraneous and irrelevant information from the raw text, such as Arabic stop words, duplicate values, digits, punctuation marks, additional spaces, English letters, usernames, URLs, HTML tags, and other special characters.

b) Lemmatization: It involves identifying and reducing words to their lemma form, which reflects the uninflected and base form of the word. As an illustration, the lemma of the word "سيفتحونها" is "فتح". Lemmatization is essential because Arabic words have several inflected forms, and its use helps standardize textual information for analysis and understanding.

c) Unicode normalization: Refers to converting a Unicode string into a standardized format. In Python programming, text consists of Unicode characters, some of which may have varying forms based on their location inside a word. For instance, the letter "ع" can be represented as { ع, ﻋ ,ﻋ}. The canonical form refers to the principal manifestation of a character. Moreover, certain characters consist of many characters or phrases, potentially posing challenges in text processing. Unicode deconstruction entails dissecting these assembled characters into their constituent elements. In order to resolve these concerns, the Normalize Unicode function can be utilized to transform the text into a more appropriate format, such as the example "ﷺ" returning "صلى الله عليه وسلم" after undergoing normalization.

d) Dediacritization: Refers to removing diacritical marks from Arabic text. These marks are small symbols appended to a letter to indicate a particular pronunciation or tone. The presence of diacritics can lead to data sparsity. Therefore, it is advisable to employ the Dediacritization function to eliminate diacritics from Arabic text. This is exemplified by the following example: "هَلْ ذَهَبْتَ إِلَى المَكْتَبَةِ؟" return "هل ذهبت إلى المكتبة؟".

e) Converts emojis into text: Emojis are visual representations or symbols frequently employed in social media content to effectively communicate emotions, responses, or concepts succinctly and expressively. Due to the potential difficulties in text processing and sentiment analysis caused by emojis in tweets, most scholars strive to eliminate them from the text during the pre-processing stage. Emojis can augment the significance of a written word, provide additional information to a message, or even substitute text entirely. Hence, it is crucial to consider the emojis. For instance: "كل 👩 🐷" return "كل وجه امرأة خنزير", if the emojis were removed, the tweet would lose its meaning. However, by translating the emojis, the tweet would retain its intended meaning.

f) Tokenization: This fundamental process involves deconstructing a given text into distinct components referred

to as tokens. A range of linguistic units, including individual words, phrases, and complete sentences, may be included within these tokens. The activity of text processing and analysis is an essential stage that enables a more efficient and simplified investigation of textual information. In Natural Language Processing (NLP), tokenization assumes a crucial role as it enables the execution of diverse tasks such as sentiment analysis, text classification, and text clustering. The procedure above facilitates the recognition and understanding of the contextual details and semantic significance communicated in the text [18].

B. Features extraction

The feature extraction process involves converting data into a collection of distinct properties that may be utilized as input for a machine learning system. Trials employing various methodologies are frequently required to identify the most efficient feature extraction techniques for a specific task [19,20].

Word vectorization is crucial in text classification because machine learning models cannot comprehend standard text. The core principle involves depicting individual words as vectors with a pre-established length within a space of elevated dimensions. Within this domain, the closeness of vectors indicates the semantic resemblance between the corresponding sentences. There are numerous ways and strategies available for text vectorization. In this work, three distinct methods of word representation were utilized to construct our approach.

In our study, for the conventional machine learning models we employed:

a) Term frequency-inverse document frequency (TF-IDF): A quantitative measure employed to assess the significance of a word within a specific set of documents. To calculate the TF-IDF score for a term in a document, one must multiply the term frequency (TF) of the word in the document by the inverse document frequency (IDF) of the word that appears in the document. "frequency" refers to the numerical count of occurrences of a particular term inside a given written composition. On the other hand, the concept of "inverse document frequency" pertains to a logarithmically scaled measure that denotes the proportion of the overall number of documents inside a corpus that encompass a particular word [21].

b) Bag Of Words (BOW): A NLP technique employed to represent textual input as a set of individual words. This approach entails first tokenizing a collection of documents into individual words. Following this, a unique identifier is assigned to each word, and the frequency of each word's occurrence in each text is calculated. A vectorized representation of the document is generated by frequency counting, where each vector dimension corresponds to a unique word within the corpus [22].

For the deep learning model, we used

c) Word embedding: This is a computational technique for representing words as compact numerical vectors within a system with many dimensions. This methodology develops a correlation between words and their vectors, wherein words possessing identical meanings are situated nearby within the vector space. Word embeddings are typically obtained by training a neural network utilizing a large corpus of textual data [22].

C. Text classification

The scope of our activity is divided into two main components:

a) The conventional machine learning Algorithms: Involves the utilization of the following machine learning algorithms: In this particular context, the algorithms frequently utilized encompass Linear SVC (Support Vector Classifier), Logistic Regression (LR), Random Forest (RF), Stochastic Gradient Descent (SGD), and eXtreme Gradient Boosting (XGB). These algorithms commonly utilize statistics to identify patterns and correlations in the data, facilitating the ability to make predictions based on these detected patterns [24].

b) Deep learning model: We utilized Bi-LSTM deep neural networks, employing word embedding techniques to generate word vectors. These vectors were then fed into the LSTM layers to extract semantic information. The Bi-LSTM is an adapted variant of the LSTM model that can sequentially process input data in both the forward and backward directions. Using this approach, the Bi-LSTM can understand the sequence's contextual information from previous and subsequent time steps. As a result, it emerges as a prominent deep learning model that demonstrates proficiency in processing sequential data, namely textual information [25].

In this study, we employ Keras's sequential model, incorporating embedding, BI-LSTM, flattening, and dense layers into our model. Subsequently, we train, test, and validate the model using preprocessed data sets. The model is trained for 15 epochs and trained for 132 batches. The primary objective of this study was to examine the binary classification of hate speech or offensive language. The dataset was partitioned into two unique categories. The subject under consideration addresses the differentiation between hate speech and non-hate speech, as well as offensive and non-offensive speech

D. Architecture

Figure 1 depicts the comprehensive framework of the proposed system designed to identify hate speech in the Arabic language. Figure 2 depicts the architecture of the Bi-LSTM deep learning model. The system has two main components: "text pre-processing" and "text classification".

Every section offers a thorough elucidation of its corresponding material.

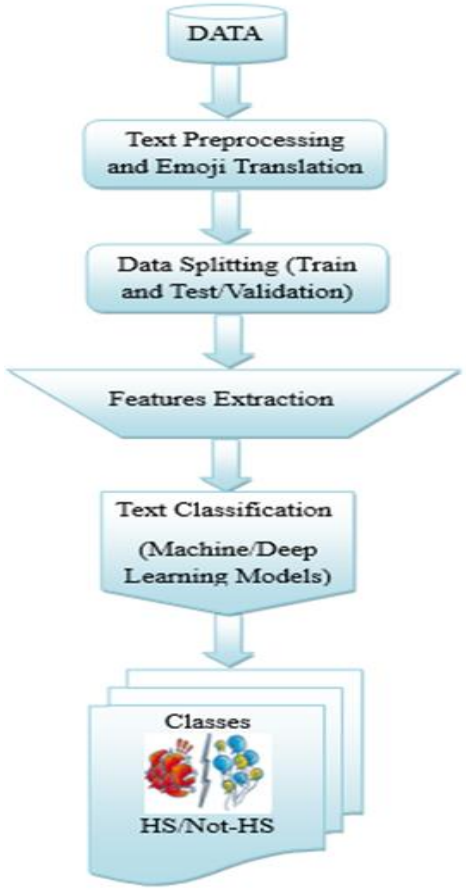

Fig. 1. The architecture of the proposed approach

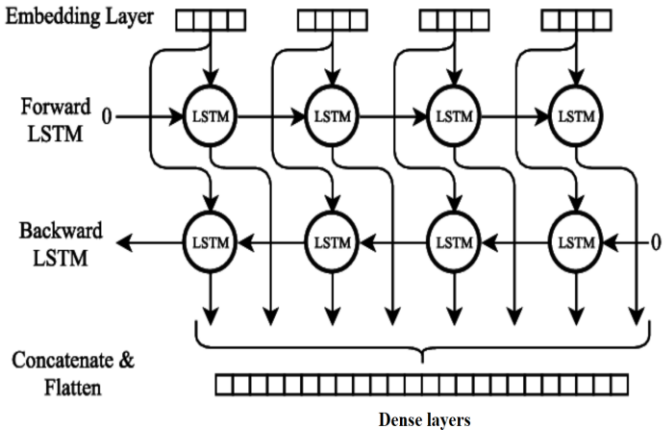

Fig. 2. The architecture of Bi-LSTM model

IV. RESULTS AND DISCUSSION

This study employed a set of four publicly available labeled datasets that focused on hate and offensive speech in social media. These datasets included the Arabic Levantine Hate Speech Detection (ALHS) [26], Abusive Language Detection on Arabic Social Media (ALDA) [27], Open-Source Arabic Corpora and Corpora Processing Tools (OSACT4) [7], and the Corpus of Offensive Language in Arabic (COLA) [28]. The data discription is displayed in Tables I and II. The tests were conducted on four datasets, and the accuracy of each model was compared across all datasets. The analysis focused on identifying the best-performing models, with accuracy as the criterion for evaluating model performance.

Table III presents the accuracy of each model's text classification task across the datasets without using the translation function. Table IV shows the accuracy of the different models employed for text classification using the function of Emoji-To-Text. These findings indicate the effectiveness of each model in reliably detecting instances of hate/offensive speech from the four datasets.

The Bi-LSTM model exhibited superior accuracy across all datasets, achieving a rate of 97.56% for the ALHS dataset, 95.87% for the ALDA dataset, 98.4% for the OSACT4 dataset, and 89.08% for the COLA dataset. Upon comparing the findings of Tables III and IV, it is evident that employing emojis by not deleting them from the data and translating them to text significantly enhances the classification results. This highlights the significant impact that the Emoji-To-Text function have on the outcomes of hate speech and offensive language detection in social media. However, it is essential to acknowledge that while the traditional machine learning model has demonstrated commendable performance in the field of text classification, there are unique methods that consistently outperform others for each dataset. The linear Support Vector Classifier (SVC) had the highest accuracy rate of 82.91% in the ALHS dataset. The Stochastic Gradient Descent (SGD) model exhibited the highest level of accuracy, at 94.2%, when applied to the ALDA dataset. Similarly, the SGD model demonstrated the highest level of accuracy and achieved a rate of 97.8% when applied to the OSACT4 dataset. When applied to the COLA dataset, the XGBoost algorithm's classification accuracy was 80.12%. Compared to traditional machine learning models, the deep learning model demonstrated improved performance, achieving higher levels of accuracy across all datasets. However, despite the dataset being in Arabic, the five implemented algorithms produced promising results. The pre-processing function for Arabic natural language processing (NLP) is crucial in improving and optimizing deep learning and machine learning models. This enhancement pertains primarily to text classification, specifically within detecting hate speech inside social media posts.

The results we collected exhibited significant progress in this field. Moreover, employing emojis instead of removing them can enhance text categorization and increase the identification of hate speech in Arabic content on social networks.

TABLE I. THE DATASETS USED FOR EXPERIMENTAL ANALYSIS

Dataset	Source	Size	Classes
ALHS	Twitter	5846	3 (normal, abusive, hate)
ALDA	Aljazeera comments	32000	3 (obscene, offensive, clean)
OSACT4	Twitter	8000	Offensive (NOT-OFF, OFF) Hate Speech (NOT-HS, HS)
COLA	YouTube comments	15050	2 (Offensive or Not)

TABLE II. TWEETS CLASSIFICATION EXAMPLES

Tweet	Pre-processed Text	Label	Class
@DimaSadek يا حمارة يا صهيونية يا عاهرة يا 💩 هؤلاء بفضلهم انت لست في خضن تكفيري الا اذا كنت تحبين الغير	يا حمار يا صهيوني يا عاهرة يا كومة من البراز فضل انت خاض تكفير الا اذ كان تحبين غير	1	Hate Speech
يا صباح الخير يا صباح الحب يا صباح الربيع<LF>💛😎<LF>URL عيش_الآن#<F	يا صباح خير يا حب يا ربيع وجه مبتسم شكل قلب أصفر نظار شمسي عيش آن	0	Not Hate Speech

TABLE III. CLASSIFICATION RESULTS REMOVING EMOJIS FROM TEXT

Dataset Model	ALHS	ALDA	OSACT 4	COLA
Linear SVC	81.86%	82.05%	84.65%	77.92%
LR	81.34%	82.94%	84.28%	65.23%
RF	79.86%	83.13%	83.15%	77.03%
SGD	74.43%	83.15%	85.71%	75.18%
XGB	80.17%	83.05%	89.29%	74.05%
Bi-LSTM	**90,5%**	**82,34%**	**96,88%**	**76,18 %**

TABLE IV. CLASSIFICATION RESULTS WITH EMOJIS TRANSLATION

Dataset Model	ALHS	ALDA	OSACT4	COLA
Linear SVC	82.91%	82.83%	85.93%	78.26%
LR	81.54%	86.29%	85.07%	69.7%
RF	82.14%	77.52%	84.2%	78.97%
SGD	75.9%	94.20%	97.8%	80.08%
XGB	79.83%	93.5%	97.4%	80.12%
Bi-LSTM	**97,56%**	**95,87%**	**98,40%**	**89,08%**

V. CONCLUSION

The issue of hate speech and offensive language on social media platforms has become a prominent concern in Arabic-speaking countries, leading to increased levels of discrimination and potential harm towards vulnerable individuals, particularly young people. This paper outlines our research findings about the advancement of Arabic pre-processing tools, with a specific emphasis on translating emojis into Arabic text. The text classification was performed using a hybrid approach that integrated deep learning and machine learning models, utilizing various strategies for extracting features. This technique aimed to ascertain whether a provided text may be classified as hateful/offensive or fall within the civility domain. The model was assessed using a set of four public datasets, and the results showed potential, recognizing the inherent difficulties presented by the complex structure and nuances of the Arabic language. Our research demonstrates that incorporating emoji translation into text, pre-processing and utilizing it for hate speech classification with the machine and deep learning models can significantly enhance model performance and improve accuracy results.

Our future study will improve the effectiveness of our approach by including different deep learning models, such as Convolutional Neural Networks (CNN), Gated Recurrent Units (GRU), and the AraBERT model.

ACKNOWLEDGMENT

The authors have nothing to acknowledge.

REFERENCES

[1] Brown A. What is hate speech? Part 1: The Myth of Hate. *Law and Philos* 2017; 36: 419–468.

[2] Poletto F, Basile V, Sanguinetti M, et al. Resources and benchmark corpora for hate speech detection: a systematic review. *Lang Resources & Evaluation* 2021; 55: 477–523.

[3] Al-Dossari AA-H and H. Detection of Hate Speech in Social Networks: A Survey on Multilingual Corpus. *Computer Science & Information Technology (CS & IT)* 2019; 9: 83.

[4] Uysal AK, Gunal S. The impact of pre-processing on text classification. *Information Processing & Management* 2014; 50: 104–112.

[5] AlOtaibi S, Khan MB. Sentiment Analysis Challenges of Informal Arabic Language. *International Journal of Advanced Computer Science and Applications (ijacsa)*; 8. Epub ahead of print 28 2017. DOI: 10.14569/IJACSA.2017.080237

[6] Salloum SA, AlHamad AQ, Al-Emran M, et al. A Survey of Arabic Text Mining. In: Shaalan K, Hassanien AE, Tolba F (eds) *Intelligent Natural Language Processing: Trends and Applications*. Cham: Springer International Publishing, pp. 417–431.

[7] Mubarak H, Darwish K, Magdy W, et al. Overview of OSACT4 Arabic Offensive Language Detection Shared Task. In: *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*. Marseille, France: European Language Resource Association, pp. 48–52.

[8] Albadi N, Kurdi M, Mishra S. Are they Our Brothers? Analysis and Detection of Religious Hate Speech in the Arabic Twittersphere. In: *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 2018, pp. 69–76.

[9] Soliman AB, Eissa K, El-Beltagy SR. AraVec: A set of Arabic Word Embedding Models for use in Arabic NLP. *Procedia Computer Science* 2017; 117: 256–265.

[10] Aluru SS, Mathew B, Saha P, et al. Deep Learning Models for Multilingual Hate Speech Detection. Epub ahead of print 9 December 2020. DOI: 10.48550/arXiv.2004.06465.

[11] Duquenne P-A, Gong H, Schwenk H. Multimodal and Multilingual Embeddings for Large-Scale Speech Mining. In: *Advances in Neural Information Processing Systems*. Curran Associates, Inc., pp. 15748–15761.

[12] Aldjanabi W, Dahou A, Al-qaness MAA, et al. Arabic Offensive and Hate Speech Detection Using a Cross-Corpora Multi-Task Learning Model. *Informatics* 2021; 8: 69.

[13] AlKhamissi B, Diab M. Meta AI at Arabic Hate Speech 2022: MultiTask Learning with Self-Correction for Hate Speech Classification. Epub ahead of print 16 May 2022. DOI: 10.48550/arXiv.2205.07960.

[14] Bjerva J. One Model to Rule them all: Multitask and Multilingual Modelling for Lexical Analysis. Epub ahead of print 3 November 2017. DOI: 10.48550/arXiv.1711.01100.

[15] Shapiro A, Khalafallah A, Torki M. AlexU-AIC at Arabic Hate Speech 2022: Contrast to Classify. In: *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*. Marseille, France: European Language Resources Association, pp. 200–208.

[16] [16] Althobaiti MJ. BERT-based Approach to Arabic Hate Speech and Offensive Language Detection in Twitter: Exploiting Emojis and Sentiment Analysis. International Journal of Advanced Computer Science and Applications (IJACSA); 13. Epub ahead of print 40/31 2022. DOI: 10.14569/IJACSA.2022.01305109.

[17] [17] Rex R. Pre-processing Techniques for Text Mining, https://www.academia.edu/35015140/Pre-processing_Techniques_for_Text_Mining (accessed 14 February 2023).

[18] Sarang P. Natural Language Understanding. In: Sarang P (ed) *Artificial Neural Networks with TensorFlow 2: ANN Architecture Machine Learning Projects*. Berkeley, CA: Apress, pp. 405–469.

[19] Sunagar P, Kanavalli A, Shetty ND. Feature Extraction And Selection Techniques For Text Classification: A Survey. *International Journal of Advanced Research in Engineering and Technology (IJARET*. Epub ahead of print December 2020. DOI: 10.34218/IJARET.11.12.2020.268.

[20] Pradnya K, Manisha M. A Survey on Feature Selection Techniques and Classification Algorithms for Efficient Text Classification. *IJSR* 2016; 5: 1267–1275.

[21] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. *Information Processing & Management* 1988; 24: 513–523.

[22] Zhang Y, Jin R, Zhou Z-H. Understanding bag-of-words model: a statistical framework. *Int J Mach Learn & Cyber* 2010; 1: 43–52.

[23] Li S, Gong B. Word embedding and text classification based on deep learning methods. *MATEC Web Conf* 2021; 336: 06022.

[24] Liu H, Cocea M. Traditional Machine Learning. In: Liu H, Cocea M (eds) *Granular Computing Based Machine Learning: A Big Data Processing Approach*. Cham: Springer International Publishing, pp. 11–22.

[25] Kowsher Md, Tahabilder A, Islam Sanjid MdZ, et al. LSTM-ANN & BiLSTM-ANN: Hybrid deep learning models for enhanced classification accuracy. *Procedia Computer Science* 2021; 193: 131–140.

[26] Mulki H, Haddad H, Bechikh Ali C, et al. L-HSAB: A Levantine Twitter Dataset for Hate Speech and Abusive Language. In: *Proceedings of the Third Workshop on Abusive Language Online*. Florence, Italy: Association for Computational Linguistics, pp. 111–118.

[27] Mubarak H, Darwish K, Magdy W. Abusive Language Detection on Arabic Social Media. In: *Proceedings of*

the First Workshop on Abusive Language Online. Vancouver, BC, Canada: Association for Computational Linguistics, pp. 52–56.

[28] Alakrot A, Murray L, Nikolov NS. Dataset Construction for the Detection of Anti-Social Behaviour in Online Communication in Arabic. *Procedia Computer Science* 2018; 142: 174–181.