Social Network Analysis and Mining (2024) 14:111
https://doi.org/10.1007/s13278-024-01258-1

REVIEW PAPER

A comprehensive review on Arabic offensive language and hate speech detection on social media: methods, challenges and solutions

Mahmoud Mohamed Abdelsamie[1] · **Shahira Shaaban Azab**[1] · **Hesham A. Hefny**[1]

Received: 1 October 2023 / Revised: 11 December 2023 / Accepted: 6 April 2024

Abstract
In recent years, social media has witnessed an exponential growth in promoting healthy relationships and communication between family, friends, and acquaintances, but it isn't without its flaws. It is clear that sometimes social media freedom can create an unattractive online environment. Hate speech and offensive language are frequently spread on social media platforms. Thus, they encompass different negative effects on our society. Therefore, detecting hate speech and offensive language has become the theme of one of the major research trends. Although the Arabic language occupies a distinct position among the languages on social media networks such as Twitter and Facebook, the ability to identify Arabic hate speech and offensive language is still developing due to the variety and complexity of Arabic dialects and forms. In this paper, we present an in-depth review focused on studies published between 2019 and September 2023 related to Arabic offensive language and hate speech detection. To conclude, we highlighted the most significant methods, Arabic datasets, taxonomy analysis, and challenges. Moreover, this review provides a foundation of knowledge that can help the researchers design and implement reliable and more accurate solutions.

Keywords Arabic offensive language · Arabic hate speech · Arabic dialects · Social media · Deep learning (DL) · Machine learning (ML) · Taxonomy · Natural language processing (NLP)

1 Introduction

Disclaimer: due to the nature of this kind of study, some examples of offensive or hate speech may be included in this survey. These examples are solely for the purpose of understanding this issue and do not represent the views or opinions of the survey creators or any of their affiliated organizations. We do not condone or support offensive or hate speech of any kind. This work is an attempt to help fight such speech.

Social media networks have revolutionized the way we communicate and interact with each other. Through these networks (Shannaq et al. 2022), people from all over the world can connect and communicate instantly. Moreover, they can feel emboldened to freely (ElZayady et al. 2023; Mansur et al. 2023; Makram 2022) share and express their thoughts, views, and opinions in ways that may not be on a personal level (Azzi and Zribi 2022). Although offensive language and hate speech are unfortunate, they have become very common on social media platforms such as Facebook and Twitter.

In common language, hate speech refers to the term used to describe offensive statements in everyday discourse. Hate speech (Ruwandika and Weerasinghe 2018) can also be defined as the use of language to disparage or incite hatred towards a person or group based on their religion, race, gender, or social standing. Excessive use of social media has led to the spread of this kind of speech. Thus, it impacts negatively on mental health and may lead to real-world consequences such as hate crimes, discrimination, and intimidation. This can affect individuals and communities' well-being and social cohesion, as mentioned in (Shannaq et al. 2022; Althobaiti 2022). Therefore, finding a solution for detecting hate speech has become crucial for countries, companies, and academic institutions (Elzayady et al. 2023). In addition, numerous studies on hate

✉ Shahira Shaaban Azab
Shahiraazazy@cu.edu.eg

Mahmoud Mohamed Abdelsamie
mahmoudak.official@gmail.com

Hesham A. Hefny
hehefny@cu.edu.eg

[1] FGSSR, Department of Computer Science, Cairo University, Cairo, Egypt

Published online: 30 May 2024

speech detection have been published, with a greater focus on the English language. In contrast, investigations into detecting Arabic hate speech are still emerging (Abuzayed 2020; Elzayady et al. 2022). Recently, due to the great interest in detecting online hate speech, we found a set of papers published to find appropriate solutions in an automated way for detecting hate speech in Arabic on social media platforms using different approaches and methods.

In the scope of our survey, we have concentrated on studies published in the last five years (2019–2023) pertaining to Arabic offensive language and hate speech detection. However, it is crucial to acknowledge that investigations predating 2019 have made substantial contributions to our comprehension of the distinctive challenges, solutions and the trends in this period regarding Arabic offensive language and hate speech detection. For instance, the authors in (Alakrot et al. 2018a, b) presented a comprehensive approach for detecting abusive language on Arabic social media using a large dataset of YouTube comments in Arabic to train a support vector machine classifier, exploring combinations of word-level features, N-gram features, and various preprocessing techniques achieving superior results. Another approach for detecting abusive language on Arabic social media, specifically in dialectal Arabic, was presented in a study by (Mubarak et al. 2017) The approach utilized two datasets: the first comprised 1100 manually labeled dialectal tweets, and the second included 32k comments flagged as inappropriate by moderators of prominent Arabic newswires. The authors introduced a statistical approach centered on a list of offensive words, achieving better outcomes. Thus, the insights gleaned from earlier research have laid a foundational understanding, providing valuable steps that continue to inform contemporary studies in this evolving field.

Therefore, this review focuses on the most recent studies on the detection of hate speech, offensive language, and abusive texts in Arabic. Our goal is to help researchers in the natural language processing (NLP) field understand the extent of the problem, evaluate the effectiveness of existing models, and develop customized solutions to mitigate the negative impacts of Arabic hate speech on social media. So, we presented this comprehensive survey, including the earlier studies, Arabic datasets, various machine learning (ML) and deep learning (DL) models, hybrid solutions, and data preparation processes: Arabic language preprocessing steps and feature extraction methods. The existing challenges with methods and the Arabic language are discussed. Moreover, we highlighted the challenges for future trends in this field.

1.1 Methodology

This section presents the procedures followed in this review, such as the search strategy, the keywords, inclusion and exclusion criteria, data extraction, and data synthesis.

The main objective of this study is to investigate the state of the art of the latest techniques in NLP to automatically detect Arabic hate speech and offensive language on different social media platforms. This survey covers the following research questions:

- *Q1* What is your understanding of offensive language and hate speech in the Arabic language?
- *Q2* What are the most promising NLP techniques, common preprocessing, and feature extraction methods for Arabic hate speech detection, and how can these techniques be optimized for Arabic datasets?
- *Q3* What are the most available Arabic datasets and how are the datasets annotated? Which social media platforms are the most frequently used?
- *Q4* What are the specific linguistic and socio-cultural features of Arabic language that make it challenging for offensive language and hate speech detection using NLP techniques?
- *Q5* What are the future directions for research in Arabic hate speech detection using NLP, and what are the key challenges and opportunities for advancing this field?

1.1.1 Search strategy

The primary objective of this review is to investigate the current scientific literature from 2019 to September 2023 that concerns Arabic offensive language and hate speech detection on social media platforms. The study aims to analyze and synthesize recent works conducted on social media platforms for detecting offensive Arabic language and hate speech in order to provide an all-inclusive summary of advancements made in this area. Therefore, we formulated a search query to find the most relevant papers on the subject of interest as follows: firstly, we established the most frequently used keywords, such as offensive language, hate speech, Arabic, Arabic offensive, Arabic hate, abusive language, classification, and detection. Second, these terms were used in multiple combinations using the Boolean operators (AND) and (OR) to form the search query.

The databases used in our search process are IEEE, Springer, Science Direct, ACL Anthology, ACM DL, IJECE Journal, IJACSA Journal, SCITEPRESS, Taylor & Francis, Emerald, applied science, Revue d'Intelligence Artificielle Journal, IOS Press, and Scopus. These databases have been carefully selected based on their abundant scientific competence in several high-impact research papers, or at least the databases that are indexed in Scopus provide fair coverage of the reviewed literature.

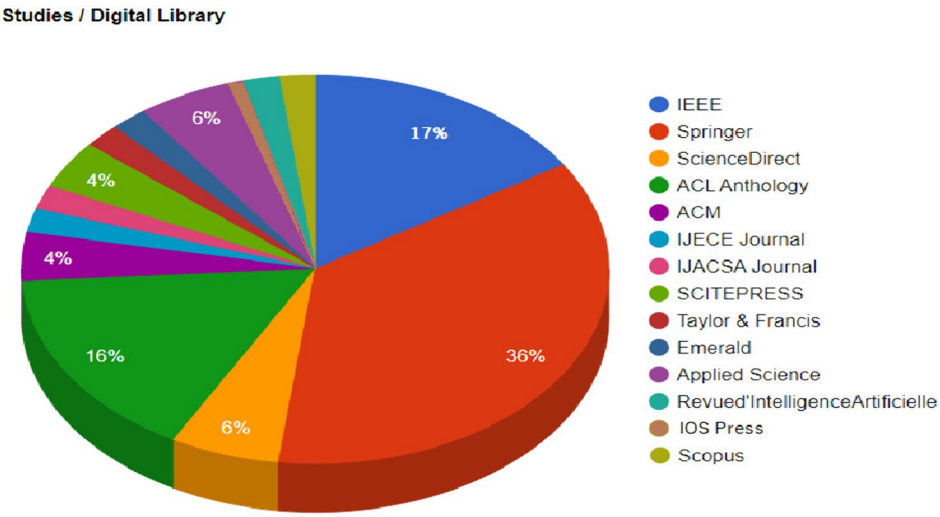

Fig. 1 The number of studies per digital library

1.1.2 Inclusion and exclusion criteria

The inclusion and exclusion criteria were used in the selected studies to identify which studies fulfilled the target of this review. The inclusion criteria involved papers that were published from 2019 to September 2023. Our main focus was only on studies about offensive language and hate speech related to the Arabic language and its challenges, whether these studies are experimental, comparative, reviews, or survey articles. While the exclusion criteria are as follows: we excluded all papers related to the detection of offensive and hate speech in other languages, such as English, Turkish, Indian, etc. Also, any publications before 2019 were excluded.

After deep analysis, we have included 54 studies in this review from variant databases published in the last five years, as shown in Fig. 1 and Fig. 2 respectively.

This survey was conducted to provide a background on Arabic offensive language and hate speech detection on social media by answering the questions mentioned above. The rest of this paper is organized as follows: the above section provides a brief introduction to the main topic. Section 2 presents a theoretical background. Then preprocessing steps and feature extraction methods will be presented in Section 3. Section 4 will go through NLP, ML, and DL techniques for detecting offensive Arabic language and hate speech. Thereafter, Section 5 presents the datasets used in previous experiments. Section 6 will go through the work related to Arabic offensive language and hate speech detection. Then a discussion about challenges and future research directions will be presented in Section 7. Finally, we concluded the work in this paper.

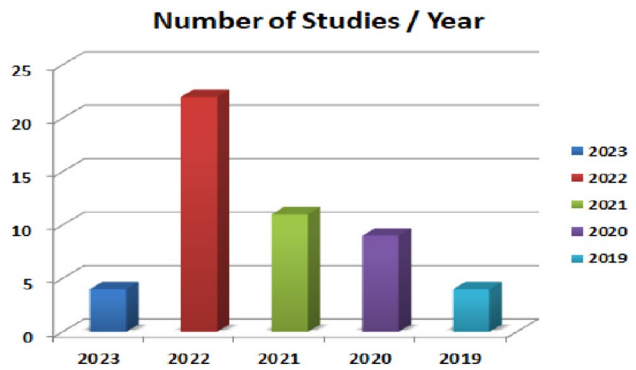

Fig. 2 The number of studies published per year

2 Background

This section introduces the Arabic language and its significance, as well as defining Arabic hate speech and offensive language. Understanding the uniqueness of the Arabic language and cultural nuances is crucial for effectively detecting and addressing offensive language and hate speech within the Arab-speaking communities. By acknowledging the importance of detecting and combating such harmful speech, we aim to contribute to a safer and more inclusive online environment for Arabic speakers.

2.1 Arabic language

Arabic is a unique language. It is also the original language of the Quran and the Hadith[1] (Referring to reports of statements or actions of the Prophet Muhammad, or of his tacit approval or criticism of something said or done in his presence). The Arabic language, with its profound historical and cultural significance, has distinct characteristics that shape

[1] https://en.wikipedia.org/wiki/Hadith

English Meaning	Arabic Word	Position
Moon	قمر	Beginning of the word
Article	مـقـال	Middle of the word
Sunrise	الشروق	End of the word

Fig. 3 Shapes of Arabic letters based on the location of the letter

Fig. 4 Different meanings of words have the same shape

its linguistic landscape. It comprises 28 letters, follows a right-to-left writing system, and incorporates gender-specific forms for various parts of speech (Rahma et al. 2023; Husain and Uzuner 2022a, b). For example, the word "Qaseera/ قصيرة" refers to a short female, and the word "Qaseer/ قصير " refers to a short male. Moreover, the limited presence of vowels (أ/alef, waaw/و, and ي/yaa) adds another layer of intricacy (Azzi and Zribi 2021). An additional characteristic of the Arabic language is the variability in the appearance of each letter, contingent upon its position within a word. To illustrate, the letter "ق/qaf" can manifest in various forms, such as "ـقـ / قـ / ق, " depending on whether it is positioned at the word's outset, in the middle, or at the end. Refer to Fig. 3 for a visual representation. Diacritics, commonly referred to as Tashkil or Harakat in Arabic. These diacritics play a crucial role in conveying the precise meaning of an Arabic word. Interestingly, they facilitate disambiguation, as words with distinct meanings may share the same visual form. For example, the Arabic word "شَمال" means the north cardinal directions, and "شِمال" carries a dual meaning, referring not only to the left direction but also encompassing a connotation of something negative or offensive in language in Arabic. Refer to Fig. 4 for a visual representation. Similarly, singular, dual, and plural forms contribute to the language's expressive depth. On the other hand, the Arabic language consists of mixed dialects (Alsafari et al. 2020a, b), such as Gulf Arabic, Egyptian Arabic, and Levantine Arabic. From the aforementioned characteristics, the Arabic language presents several challenges in the context of natural language processing (NLP), stemming from its complex morphology and the use of dialects with rich cultural and historical roots. Although, the Arabic language has witnessed a substantial increase in its prevalence on various digital spaces, including but not limited to social networks. Moreover, it holds the fourth position among the most frequently utilized languages on the web (Khezzar et al. 2023). Unfortunately, there has been a surge in offensive language and hate speech on Arabic social media platforms in recent years (Shannaq et al. 2022; Mohaouchane et al. 2019). However, in response to these challenges, researchers have leveraged advanced technologies, including natural language processing, machine learning, and deep learning techniques in their studies. The findings from these studies emphasize that hate speech and offensive language in Arabic have evolved into a pressing concern, underscoring the need for further investigation and the development of effective mitigation strategies (ElZayady et al. 2023), (Althobaiti 2022).

2.2 Offensive language

Abusive or offensive language definition is a very complex task and a debatable issue (Husain and Uzuner 2021). Offensive language on social media refers to any language used that is intended to harm, insult, degrade, or discriminate against an individual or group of individuals based on their race, gender, sexual orientation, religion, nationality, or disability. It can take many forms (Alshalan and Al-Khalifa 2020) including hate speech, cyberbullying, trolling, and harassment. For instance, a YouTube comment like " صوتك عامل زي الحمارالله يلعنك ", which means: "May God curse you; your voice is like a donkey's voice". As mentioned in (Azzi and Zribi 2021), offensive language can be defined as any content that contains some form of abusive behavior, exhibiting actions with the intention of harming others, causing hurt, and making others angry. Also, (Azzi and Zribi 2022) provides some offensive language classes, namely, racism, sexism, xenophobia, violence, hate, pornography, religious hatred, and LGBTQ hate. The definition of offensive language depends on people's social and political backgrounds. Regarding the types of offensive language, (Azzi and Zribi 2021) provides the main types of offensive language on social media as follows: discriminative content includes any sort of prejudice against a person showing different physical characteristics, belongings, or preferences, while violent content is the use of any term threatening or promoting an intentioned act of violence. Adult content includes pornography, texts illustrating sexual behavior and more importantly children sexual abuse. Vulnerable categories of people like children or youth are particularly vulnerable to the psychological threat of adult-oriented content on social media.

To the best of our knowledge, detecting offensive language on social media is a complex task due to the sheer volume of data, new words continuously emerging (Mubarak

and Darwish 2019), the use of slang or highly contextualized language, and the rapidly changing nature of language use on social media platforms. Research has been conducted on the detection of offensive language on social media using natural language processing (NLP) techniques, including machine learning, deep learning techniques, and sentiment analysis.

2.3 Hate speech

The definition of hate speech has always been a topic of discussion (Boulouard et al. 2022a, b). According to (ElZayady et al. 2023; Alhejaili et al. 2022; Awane et al. 2021; Husain and Uzuner 2021; AlKhamissi 2022), hate speech is any form of public expression that promotes, incites, or justifies hatred, discrimination, or hostility against one person or a group of people based on their identity. For instance, a tweet like "قليل الأدب و أنا لو منك كنت ضربته قلمين علي وشه دا عيل", which means: "This is a poorly mannered family, and if I were you, I would slap him on his face". Hate speech in (Guellil et al. 2020) was defined as any communication that disparages or defames a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic, and it was classified into four categories: gender-based hate speech, religious hate speech, racial hate speech, and disability hate speech. The study (Faris et al. 2020) defines hate speech as the use of offensive language to spread hatred and discrimination based on race, sex, religion, or disability.

Finally, hate speech is complex and ambiguous because it is not just word identification (Haddad 2020). It can occur in different linguistic styles and through different acts, such as insulting, abusing, provocation, and aggression (Omar et al. 2020).

2.4 Importance of offensive language and hate speech detection

Arabic offensive language detection and hate speech detection on social media would be crucial for several reasons. First, with the increasing number of Arabic speakers on social media (Boulouard et al. 2022a, b), it is essential to have effective tools to detect offensive language and hate speech produced in Arabic. Second, social media has been used as a platform for hate speech and offensive language due to the freedom of expression on such platforms (Badri et al. 2022). The spread of misinformation, propaganda, and biased narratives has led to social unrest and violence in some countries. By detecting and removing such content in Arabic, social media platforms can promote a safe and inclusive online environment. Third, automated detection systems that can perform real-time analysis of large volumes of social media data in Arabic can help governments and authorities detect and prevent hate crimes, radicalization, and other forms of extremist behavior. In conclusion, Arabic offensive language detection and hate speech detection on social media are critically important for promoting peace, harmony, and inclusivity in society. This review can enhance our understanding of the challenges of detecting and addressing offensive language and hate speech on social media and help develop effective algorithms and tools for mitigating such content.

3 Preprocessing and feature extraction methods

Hate speech and abusive language are prevalent on social media platforms, and controlling such language is essential to promoting a safer and more inclusive online environment. In recent years, researchers have started to develop algorithms and models to detect hate speech and abusive language in Arabic and its dialects. Preprocessing steps and feature extraction methods play a critical role in the accuracy of these algorithms. Preprocessing steps usually involve segmentation, normalization, and cleaning techniques. Feature extraction methods used for Arabic language hate speech and abusive language detection include lexical, syntactic, and semantic features. This section aims to provide an overview of the preprocessing steps and feature extraction methods used for Arabic language and dialect hate speech and abusive language detection on social platforms.

3.1 Preprocessing steps

In the literature presented, researchers have employed various preprocessing steps to improve the accuracy of Arabic offensive language detection and hate speech detection methods on social media. Some of the most commonly used preprocessing steps include:

3.1.1 Stop words removal

Stop words are frequently occurring words that do not carry much meaning. Researchers remove these words from the text before running any analysis (Alshalan and Al-Khalifa 2020; Husain 2020; Alotaibi and Abul Hasanat 2020; Abdel-Hamid et al. 2022). In addition, the authors in (Albadi et al. 2019) presented that they didn't remove any negation words since these are usually informative in sentiment analysis tasks.

3.1.2 Noise removal

Researchers remove various forms of noise such as URLs, Emojis, digits, punctuation marks, non-Arabic words,

repeated characters, mentions, HTML tags, and other symbols such as , emails, dates, and diacritics. Diacritics are short vowels and characters above and beneath letters, such as fatha, damma, kasra, etc. (Shannaq et al. (2022); Elzayady et al. 2023a, b); Makram 2022; Azzi and Zribi 2022; Althobaiti 2022; Berrimi et al. 2020; Alshalan and Al-Khalifa 2020; Haddad 2020; Omar et al. 2020; Husain 2020; Mubarak 2020; Alakrot et al. 2021; AbdelHamid et al. 2022; Badri et al. 2022; Alsafari et al. 2020a, b; Mostafa 2022; Alzubi 2022; Boulouard et al. 2022a, b; Khezzar et al. 2023). In addition, the authors in (Elzayady et al. 2022) raised the removal of empty lines to obtain cleaner text.

3.1.3 Tokenization

Researchers split the text into small units, such as words or phrases, to facilitate analysis. This operation therefore makes it possible to segment a text document into word tokens (Badri et al. 2022).

3.1.4 Stemming and lemmatization

Stemming and lemmatization are used to reduce words to their base forms and reduce the number of unique words in the dataset (Elzayady et al. 2023a, b; Boulouard et al. 2022a, b).

3.1.5 Emoji and emoticon conversion

It means changing emoji and emoticons into Arabic textual labels that explain the content of them such as 😊, is replaced by (سعيد) which means 'happy' (Elzayady et al. (2023a, b; Husain and Uzuner 2022a, b; Alshalan and Al-Khalifa 2020; El-Alami et al. 2022), and (Alzubi 2022).

3.1.6 Normalization

(Shannaq et al. (2022); Husain and Uzuner 2022a, b; AlFarah et al. 2022) The normalization of Arabic characters, such as changing the letters (ئ, آ, أ) to (ا), and (ي) to (ى). Also, (Al-Hassan and Al-Dossari 2021) included the removal of the Arabic dash that is used to expand the word (e.g., الت ـــعلم) to (التعلم) which means 'learning'.

3.2 Feature extraction methods

Feature extraction is the process of transforming raw data into features that can be used for model training. Different feature extraction methods have been used to identify the presence of offensive language and hate speech in Arabic social media texts. The most commonly used feature extraction methods include:

3.2.1 Bag of words (BOW)

This method involves counting the frequency of each word in the text and then treating the counts as features.

3.2.2 TF-IDF

This method assigns a weight to each word based on its frequency in the document and its frequency across all documents.

3.2.3 N-grams

This technique involves extracting a sequence of n words from the text and treating them as features, where n can be any positive integer.

3.2.4 Word embedding (WE)

This method involves representing words in a vector space, such that words with similar meanings are closer together. We can also say that it helps in capturing the underlying semantic relationships between words.

3.2.5 Linguistic-based features (part of speech tagging (POS))

This technique involves identifying the grammatical structure of the text and using it to extract meaningful features. It involves labeling each word in the text with its corresponding part of speech, such as noun, verb, adjective, etc., and extracting features based on the frequency of hate speech keywords in each part of speech category.

Finally, Table 1 demonstrates the different feature extraction and word representation methods used in Arabic offensive language and hate speech detection.

4 Taxonomy: NLP, ML and DL models FOR Arabic offensive and hate speech detection

Natural language processing (NLP) is an advanced computational approach that deals with the analysis, understanding, performing natural-language commands, and generation of human language (Mansur et al. 2023). Over the past few years, NLP has gained significant attention from researchers and practitioners due to its promising applications in several fields, including but not limited to text classification,

Table 1 Feature extraction and word representation methods

Method	Reference
TF-IDF	Elzayady et al. (2023a, b, 2022); Althobaiti (2022); Al-Hassan and Al-Dossari (2021); Alhejaili et al. (2022); Haddad (2020); Boulouard et al. (2022a, b); Shannag et al. (2022); Husain (2020); AbdelHamid et al. (2022); AlFarah et al. (2022); Alzubi (2022); Aljuhani et al. (2022); Khezzar et al. (2023); Khairy et al. (2023)
Aravec	Shannaq et al. (2022); Azzi and Zribi (2022); Mohaouchane et al. (2019); Husain and Uzuner (2022a, b); Faris et al. (2020); Haddad (2020); Husain (2020); Mubarak (2021); AbdelHamid et al. (2022); Badri et al. (2022); Albadi et al. (2019); Alsafari et al. (2020a, b); Aljuhani et al. (2022)
Skip-Gram (SG)	Shannaq et al. (2022); Azzi and Zribi (2022); Elzayady et al. (2023a, b); Mohaouchane et al. (2019); Duwairi et al. (2021); Guellil et al. (2020); Faris et al. (2020); Haddad (2020); Mubarak et al. (2021); Alsafari and Sadaoui (2021a, b); AbdelHamid et al. (2022); Alsafari et al. (2020a, b); Alsafari and Sadaoui (2021a, b)
CBOW	Shannaq et al. (2022); Azzi and Zribi (2022; Elzayady et al. (2023a, b); Mohaouchane et al. (2019); Duwairi et al. (2021); Alshalan and Al-Khalifa (2020); Guellil et al. (2020); Haddad (2020); Albadi et al. (2019); Alsafari et al. (2020a, b); Anezi (2022); Aljuhani et al. (2022)
Word2Vec	(Azzi and Zribi (2022); Alshalan and Al-Khalifa (2020); Guellil et al. (2020); Faris et al. (2020); Haddad (2020); Alsafari and Sadaoui (2021a, ba, b, 2021a, ba, b); Anezi (2022); Aljuhani et al. (2022)
n-grams	Shannaq et al. (2022); Husain and Uzuner (2022a, b); Alshalan and Al-Khalifa (2020); Mubarak and Darwish (2019); Alsafari et al. (2020a, b); Alsafari and Sadaoui (2021a, b)
FastText	Guellil et al. (2020; Mubarak and Darwish (2019); Mubarak (2021); Badri et al. (2022); Alsafari et al. (2020a, b)
AraBert WE	Alsafari et al. (2020a, b); Mubarak (2021); Alsafari and Sadaoui (2021a, b); Alsafari and Sadaoui (2021a, b)
AraVec2.0	Elzayady et al.(2023a, b, 2022)
MUSE	Duwairi et al. (2021); Alzubi (2022)
MARBERT	Makram (2022); Elzayady et al. (2023a, b)
Mazajak WE	Mubarak (2021); Alzubi (2022)
GloVe	Shannaq et al. (2022); Anezi (2022)
biLM	El-Alami et al. (2022)
ELMo	El-Alami et al. (2022)
AraBERTv0.2-Twitter large	Alzubi (2022)
Emoji score	Alzubi (2022)
BERTbase-multilingual	Mubarak (2021)
DistilBert	Alsafari and Sadaoui (2021a, b)
Blend Embeddings	Aljuhani et al. (2022)
FastText-SkipGram	Alsafari et al. (2020a, b)
MBert WE	Alsafari et al. (2020a, b)
Part-Of-Speech Tagger	Alakrot et al. (2021)
AraVec3.0	Shannag et al. (2022)
Bert	Abbes et al. (2023)
Count of positive and negative terms, based on polarity lexicon	Mubarak (2021)

sentiment analysis, and speech recognition. Text classification can be useful for automatically identifying offensive language by assigning labels to new unseen texts (Husain and Uzuner 2021). To the best of our knowledge, one of the most pressing challenges that NLP has recently faced is the rise of offensive language and hate speech on social media platforms. Arabic, as a language with a rich history and a broad user-base, has been heavily affected by this challenge. Therefore, in this section, we aim to provide a comprehensive taxonomy analysis of various methods used in this domain, including machine learning, deep learning, transformer-based methods, and ensemble approaches.

Machine learning methods have been widely used in hate speech detection tasks. These methods have shown promising results in identifying hate speech, but they may struggle to capture complex semantic relationships and dependencies in Arabic text. Table 2 provides a summary of the most common ML methods used in the selected studies.

To overcome this limitation, deep learning techniques have gained popularity due to their ability to capture intricate patterns in text data. Deep learning models, such as Convolutional Neural Networks (CNN), recurrent neural networks (RNN). Table 3 provides a summary of the most common DL methods used in the selected studies.

Table 2 A summary table of the most ML methods used in the selected studies

ML techniques	Reference
Logistic regression (LR)	Shannaq et al. (2022); Elzayady et al. (2023a, b); Makram (2022); Althobaiti (2022); Alhejaili et al. (2022); Husain and Uzuner (2022a, b); Alshalan and Al-Khalifa (2020); Guellil et al. (2020); Haddad (2020); Omar et al. (2020); Boulouard et al. (2022a, b); Mubarak (2021); Alakrot et al. (2021); Badri et al. (2022); AlFarah et al. (2022); Albadi et al. (2019); Alsafari et al. (2020a, b); Anezi (2022); Aljuhani et al. (2022); Khezzar et al. (2023); Khairy et al. (2023); Muaad et al. (2023)
Support vector classifier (SVM) or (SVC)	Shannaq et al. (2022); Elzayady et al. (2023a, b); Azzi and Zribi (2022); Althobaiti (2022); Al-Hassan and Al-Dossari (2021); Alhejaili et al. (2022); Husain and Uzuner (2022a, b); Alshalan and Al-Khalifa (2020); Haddad (2020); Omar et al. (2020); Boulouard et al. (2022a, b); Shannag et al. (2022); Mubarak (2021); AlFarah et al. (2022); Albadi et al. (2019); Alsafari et al. (2020a, b); Aljuhani et al. (2022); Khezzar et al. (2023); Khairy et al. (2023); Muaad et al. (2023)
Random forest (RF)	Shannaq et al. (2022); Elzayady et al. (2023a, b); Makram (2022); Alhejaili et al. (2022); Husain and Uzuner (2022a, b); Guellil et al. (2020); Boulouard et al. (2022a, b); Mubarak (2021); Badri et al. (2022); Alsafari et al. (2020a, b); Anezi (2022); Khezzar et al. (2023); Khairy et al. (2023); Muaad et al. (2023)
Decision tree (DT)	Shannaq et al. (2022); Elzayady et al. (2023a, b); Alhejaili et al. (2022); Omar et al. (2020); Mubarak (2021); Alakrot et al. (2021); AlFarah et al. (2022); Anezi (2022); Khezzar et al. (2023); Muaad et al. (2023)
Naive Bayes (NB)	Shannaq et al. (2022); Boulouard et al. (2022a, b); AlFarah et al. (2022); Alsafari et al. (2020a, b); Anezi (2022); Muaad et al. (2023)
Linear support vector machine (LinearSVC)	Guellil et al. (2020); Alakrot et al. (2021); Khezzar et al. (2023)
Extreme Gradient Boosting (XGBoost)	Shannaq et al. (2022); Elzayady et al. (2023a, b); AbdelHamid et al. (2022)
Multi-layer perceptron (MLP)	Guellil et al. (2020); Anezi (2022)
Gaussian naïve Bayes (GNB)	Alhejaili et al. (2022); Mubarak (2021); Guellil et al. (2020)
Stochastic gradient descent (SGD)	Guellil et al. (2020); Omar et al. (2020); Khezzar et al. (2023)
KNearestNeighbor (KNN)	Shannaq et al. (2022); Alhejaili et al. (2022); Khezzar et al. (2023); Khairy et al., (2023)
MultinomialNB	Omar et al. (2020); Khezzar et al. (2023)
BernoulliNB	Omar et al. (2020); Khezzar et al. (2023); Muaad et al. (2023)
Ridge	Haddad (2020); Omar et al. (2020)
Perceptron	Omar et al. (2020); Mubarak (2021)
AdaBoost	Alhejaili et al. (2022); Mubarak (2021)
extra trees	Elzayady et al. (2023a, b); Alakrot et al. (2021); Muaad et al. (2023)
Gradient boosting	Elzayady et al. (2023a, b); Mubarak (2021)
Nu-support vector classification (NuSVC)	Omar et al. (2020)
Complement NB	Omar et al. (2020)
Nearest centroid	Omar et al. (2020)
CatBoost	AbdelHamid et al. (2022)
Passive-aggressive classifier (PAC)	Elzayady et al. (2022)

On other hand, transformer-based methods such as, BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pretrained Transformer) have achieved remarkable success in various natural language processing tasks, including hate speech detection. Table 4 provides a summary of the most Transformer-based and transfer learning methods used in the selected studies.

Additionally, ensemble methods have been proposed to capitalize on the strengths and weakness of different models and enhance hate speech detection performance further. Table 5 provides a summary of the most common ensemble models used in the selected studies.

Finally, an extensive taxonomy analysis of machine learning, deep learning, transformer-based, and ensemble methods for offensive language and Arabic hate speech detection is illustrated in Fig. 5.

5 Datasets

The datasets used in Arabic offensive language and hate speech detection play a crucial role in determining the effectiveness and accuracy of the techniques used. The quality and quantity of the data directly impact the performance of these techniques. Consequently, it is essential

Table 3 A summary table of the most DL methods used in the selected studies

DL techniques	Reference
CNN	Azzi and Zribi (2022); Mohaouchane et al. (2019); Duwairi et al. (2021); Alshalan and Al-Khalifa (2020); Faris et al. (2020); Haddad (2020); Omar et al. (2020); Alotaibi and Abul Hasanat (2020); El-Alami et al. (2022); Alsafari et al. (2020a, b); Khezzar et al. (2023)
LSTM	Elzayady et al. (2023a, b); Al-Hassan and Al-Dossari (2021); Husain and Uzuner (2022a, b); Guellil et al. (2020); Faris et al. (2020); Boulouard et al. (2022a, b); Husain (2020); El-Alami et al. (2022); Alsafari et al. (2020a, b); Boulouard et al. (2022a, b)
BiLSTM bidirectional LSTM	Elzayady et al. (2023a, b); Azzi and Zribi (2022); Mohaouchane et al. (2019); Guellil et al. (2020); Husain (2020); El-Alami et al. (2022); Alsafari and Sadaoui (2021a, b); Aljuhani et al. (2022)
The gated recurrent unit (GRU)	Elzayady et al. (2023a, b); Al-Hassan and Al-Dossari (2021); Alshalan and Al-Khalifa (2020); Husain (2020); Albadi et al. (2019); El-Alami et al. (2022); Alsafari et al. (2020a, b)
RNN	Husain and Uzuner (2022a, b); Faris et al. (2020); Omar et al. (2020); Husain (2020)
Bidirectional gated recurrent unit with attention (BI-GRU)	Azzi and Zribi (2022); Elzayady et al. (2022); Haddad (2020); Husain (2020)
EL LSTM, ESoA, ELSoA (Soft attention mechanism)	Berrimi et al. (2020)
Bi-LSTM with attention mechanism	Mohaouchane et al. (2019); Abbes et al. (2023)
CNN_ATT,	Haddad (2020)
Bi-GRU_ATT	Haddad (2020)
DRNN-2	Anezi (2022)
DRNN-1	Anezi (2022)

Table 4 A summary table of the most Transformer-based and Transfer Learning methods used in the selected studies

Techniques	Reference
AraBERT	Elzayady et al. (2023a, b); Husain and Uzuner (2022a, b); Duwairi et al. (2021); Husain and Uzuner (2022a, b); Mubarak (2021); AbdelHamid et al. (2022); El-Alami et al. (2022); Alsafari et al. (2020a, b); Mostafa (2022); Alzubi (2022); De Paula (2022); Boulouard et al. (2022a, b); Khezzar et al. (2023); Muaad et al. (2023); M. Abbes et al. (2023); Mohamed et al. (2023)
MBERT	Duwairi et al. (2021); El-Alami et al. (2022); Alsafari et al. (2020a, b), De Paula (2022)
BERT	Azzi and Zribi (2022); Awane et al. (2021); Alshalan and Al-Khalifa (2020); Mubarak (2021)
QARiB	Duwairi et al. (2021); Mostafa (2022)
ArabicBERT	Husain and Uzuner (2022a, b); AbdelHamid et al. (2022)
MARBERT	Elzayady et al. (2023a, b); Mostafa (2022); Mohamed et al. (2023)
XLM-Roberta	Duwairi et al. (2021); De Paula (2022)
AraElectra	De Paula (2022)
Albert-Arabic	De Paula (2022)
AraGPT2	De Paula (2022)
MARBERTV2	Mostafa (2022); Ahmed et al. (2022); Mohamed et al. (2023)
GigaBERT	AbdelHamid et al. (2022)
AraULMFiT	El-Alami et al. (2022)
BERT base-multilingual	Mubarak (2021); Ahmed et al. (2022)
BERTEN	Boulouard et al. (2022a, b)
mBERTAR	Boulouard et al. (2022a, b)
mBERTEN	Boulouard et al. (2022a, b)
bert-large-arabertv02-twitter	Ahmed et al. (2022)
Bert-base-arabic-camelbert-mix	Ahmed et al. (2022); Al-Dabet et al. (2023)
Araelectra-base-discriminator	Ahmed et al. (2022)
Camelbert-DA, Camelbert-CA, Camelbert-MSA	Al-Dabet et al. (2023)

Table 5 A summary table of the most ensembles of models used

Techniques	Reference
CNN + GRU	Elzayady et al. (2023a, b); Al-Hassan and Al-Dossari (2021); Alshalan and Al-Khalifa (2020); Badri et al. (2022); El-Alami et al. (2022)
CNN-LSTM	Elzayady et al. (2023a, b); Mohaouchane et al. (2019); Al-Hassan and Al-Dossari (2021); Duwairi et al. (2021)
BiLSTM-CNN	Elzayady et al. (2023a, b); Duwairi et al. (2021); El-Alami et al. (2022)
CNN + AraBert	Alsafari et al. (2020a, b); Alsafari and Sadaoui (2021a, b)
BiLSTM + AraBert	Alsafari et al. (2020a, b); Alsafari and Sadaoui (2021a, b)
CNN + DistilBert	Alsafari and Sadaoui (2021a, b)
BiLSTM + DistilBert	Alsafari and Sadaoui (2021a, b)
CNN + SG	Alsafari and Sadaoui (2021a, b)
BiLSTM + SG	Alsafari and Sadaoui (2021a, b)
CNN + Bert	Alsafari and Sadaoui (2021a, b)
Emoji-Score,AraBERT,Char + word + MUSE + Emoji	Alzubi (2022)
LightGBM + MARBERT + MARBERTV2	Mostafa (2022)
AraBERT-B-T + MARBERT + QARiB	Mostafa (2022)
MARBERTV2 + MARBERT + QARiB	Mostafa (2022)
Majority vote and Highest sum	De Paula (2022)
AraHS model	AlKhamissi (2022)
GA-XGBoost	Shannaq et al. (2022)
GA-SVM	Shannaq et al. (2022)
Bagging (Random forst)	Khairy et al. (2023); Muaad et al. (2023)
Boosting (Adaboost)	Khairy et al. (2023); Muaad et al. (2023)

to use high-quality datasets that can accurately represent the different types of offensive language. The datasets building process involves three stages (Omar et al. 2020): data collection, data filtering, and data annotation. Figure 6 depicts the dataset building process. In this section, to present a clear overview, we have provided a comprehensive table (Table 6) outlining the datasets employed, offering crucial details such as their names, sizes, sources, and characteristics. These datasets were representing a diverse range of offensive language and Arabic hate speech instances, allowing for a more thorough examination of the problem at hand.

Furthermore, we introduced the dataset from availability/non-availability perspective. Figure 7 shows the percentage of dataset availability. On other hand, this survey revealed that a large majority of the Arabic hate speech datasets are imbalanced in nature. This means that the datasets contain a disproportionate amount of data representing certain types of hate speech, while other types are underrepresented. By analyzing the datasets used in this review, researchers can identify common features and patterns that could be leveraged to improve the accuracy and efficiency of hate speech detection algorithms. Moreover, by comparing the results of different studies and analyzing the underlying datasets, researchers can determine the most effective approaches and identify areas for future improvement. Also, in Table 7 we showed the most frequent datasets used for Arabic offensive language and hate speech detection in recent studies.

6 Literature review

This section highlights a brief summary of the earlier studies related to the domain of our survey and how they contribute to the existing body of knowledge on Arabic offensive language detection on social media. First of all, it should be mentioned that the Arabic language is one of the most widely spoken languages globally, and social media platforms are widely used by Arabic-speaking communities (Azzi and Zribi 2022; Berrimi et al. 2020; Husain and Uzuner 2022a, b; Mohaouchane et al. 2019; Al-Hassan and Al-Dossari 2021). In the research conducted by (Elzayady et al. (2023, 2022); Abuzayed 2020; Husain and Uzuner 2022a, b), and to the best of our knowledge, studies done in Arabic compared to other languages to find an optimum solution for automatically detecting offensive and hate speech are still few. Recently, the researchers paid

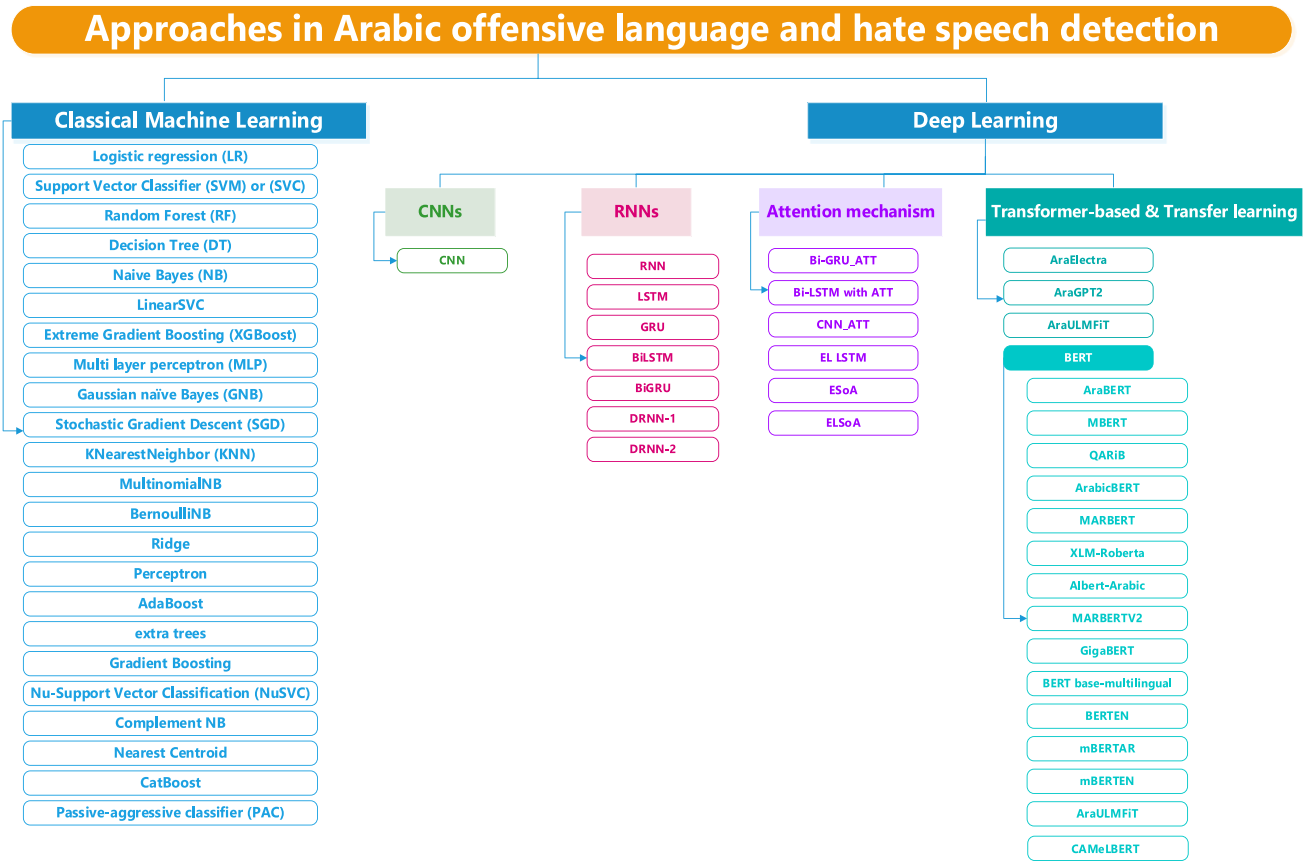

Fig. 5 Taxonomy of the approaches in Arabic offensive language and hate speech detection studies

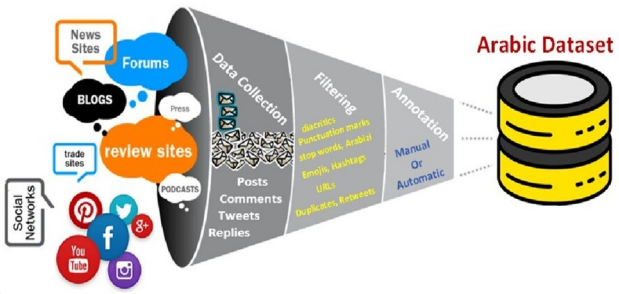

Fig. 6 The datasets building process

attention to Arabic natural language processing (ANLP) and its challenges in developing automatic solutions for Arabic offensive language detection on social media.

Researchers used a variety of approaches to detect and classify offensive Arabic languages in these competitions. For instance, some authors examined ML methods such as NB, KNN, SVM, RF, XGBoost, DT, and LR (Shannaq et al. 2022; EL-Zayady et al. 2023a, b; Azzi and Zribi 2022; Makram 2022; Althobaiti 2022; Alhejaili et al. 2022). Others applied a fine tuning of deep bidirectional transformers for Arabic, such as AraBERT and MARBERT (Althobaiti 2022; Elzayady et al. 2023; Husain and Uzuner 2022a, b). However, (Elzayady et al. 2023a, b; Azzi and Zribi 2022; Mohaouchane et al. 2019; Al-Hassan and Al-Dossari 2021; Alsafari et al. 2020a, b; Duwairi et al. 2021) trained various deep neural network models.

This review of Arabic offensive language and hate speech detection does not exceed fifty-four studies, as mentioned above. In addition, a brief summary of the studies, contributions, techniques, and superior results is presented in Table 8.

Several attempts are conducted in the literature to detect Arabic offensive language using a variety of datasets collected from different social media platforms. For instance, the authors in (Shannaq et al. 2022) proposed an intelligent prediction system to detect offensive language in Arabic tweets. For this purpose, they tested the proposed approach on an Arabic Cyber Bullying Corpus (ArCybC), which contains 4505 tweets collected from different domains on Twitter: gaming, sports, news, and celebrities, by fine-tuning the pre-trained word embedding models using seven ML classifiers, namely NB, KNN, SVM, RF, XGBoost, DT, and LR. They found that the XGBoost and SVM

Table 6 Arabic offensive language and hate speech datasets in the selected studies

Reference			Platform	#Records	Annotation	Labels	Data distribution			Availability	Download links
Arabic cyber bullying corpus (ArCybC) Shannaq et al. (2022)			Twitter	4505	Manual	Cyberbullying (CB), Non-cyberbullying (Non-CB), Offensive (Off), Non-Offensive (Non-off)	Category		# Tweets	YES	https://data.mendeley.com/v1/datasets/z2dfgrzx47/draft?a=12a9ff5d-6c5c-4b2e-8990-7d044d7c12e2
							appearance		905		
							intellectual		769		
							Racial		2349		
							Sexual		482		
Private dataset (Azzi and Zribi 2022)			Twitter and YouTube	6000	Manual	Racism, Sexism, Xenophobia, Violence, Hate, Pornography, Religious hatred, and LGBTQa hate or normal	1914 out of the 6000 lines (31%) were labeled as Normal while the rest is marked as abusive			NO	–
Four datasets (Berrimi et al. 2020)	AJComments Classification-CF (Mubarak et al. 2017)		Aljazeera.net	32k	Crowd-Flower	Obscene, offensive, or clean	2% obscene, 79% offensive, and 19% clean			YES	http://alt.qcri.org/~hmubarak/offensive/AJCommentsClassification-CF.xlsx
	AraHate Albadi et al. (2018, 2019)		Twitter	6000	Crowd-Flower	Hate or not hate	Category	# Tweets	Category	YES	https://github.com/nuhaalbadi/Arabic_hatespeech
							Hate	2526	Hate		
							Non-hate	3474	Non-hate		
	offenseval2020-arabic (OSACT4)		Twitter	10,000	Manual	Off or not off	Category	# Tweets	Category	YES	https://sites.google.com/site/offensevalsharedtask/multilingual
							Not off	8009	Not off		
							Off	1991	Off		
	Combination L-HSAB, T-HSAB	L-HSAB (Mulki et al. 2019)		5846	Manual	Abusive, hate or normal	Category	# Tweets			https://github.com/Hala-Mulki/L-HSAB-First-Arabic-Levantine-HateSpeech-Dataset
							Abusive	1728			
							Hate	468			
							Normal	3650			
		T-HSAB (Haddad et al. 2019)	Facebook & YouTube	6024	Manual	Abusive, hate or normal	Category	# Rec		YES	https://github.com/Hala-Mulki/T-HSAB-A-Tunisian-Hate-Speech-and-Abusive-Dataset
							Abusive	1126			
							Hate	1077			
							Normal	3820			

Table 6 (continued)

Reference		Platform	#Records	Annotation	Labels	Data distribution		Availability	Download links
ArabicCommentsFromYouTube (Mohaouchane et al. 2019; Boulouard et al. 2022a, b; Alakrot et al. 2021)		YouTube	15,050	Manual	offensive or not offensive	Category Offensive Not offensive	# Comm 9233 5817	YES	https://onedrive.live.com/?authkey=%21ACDXj%5FZNcZPqzy0&id=6EF6951FBF8217F9%21105&cid=6EF6951FBF8217F9
private Twitter dataset (Alhejaili et al. 2022)		Twitter	5408	Manual	Hate or not hate	Category Hate Non-hate	# Tweets 1019 4387	NO	–
Four datasets (Husain and Uzuner 2022a, b)	Egyptian Tweets (Waseem and Chung 2017)	Twitter	1100	Manual	Offensive or not offensive	Category Offensive Not offensive	# Tweets 647 453	YES	http://alt.qcri.org/~hmubarak/offensive/TweetClassification-Summary.xlsx
	L-HSAB (Mulki et al. 2019)	Twitter	5846	Manual	Offensive (where hate and abusive classes are treated as offensive or not offensive	Category Offensive Not offensive	# Tweets 2196 3650	YES	https://github.com/Hala-Mulki/L-HSAB-First-Arabic-Levantine-HateSpeech-Dataset
	T-HSAB (Haddad et al. 2019)	Facebook & YouTube	6024	Manual	Offensive (where hate and abusive classes are treated as offensive or not offensive	Category Offensive Not offensive	# Rec 3820 2204	YES	https://github.com/Hala-Mulki/T-HSAB-A-Tunisian-Hate-Speech-and-Abusive-Dataset
	(OSACT4) (Mubarak 2020)	Twitter	10,000	Manual	Offensive or not offensive	Category Offensive Not offensive	# Tweets 1991 8009	YES	https://sites.google.com/site/offensevalsharedtask/multilingual

Table 6 (continued)

Reference	Platform	#Records	Annotation	Labels	Data distribution		Availability	Download links
ArabicHateSpeechDataset Alsafari et al.(2020a, b)	Twitter	5340	Manual	Binary classification (Clean(C) vs Offensive/ Hate(OH))	Category	# Tweets	YES	https://github.com/sbalsefri/ArabicHateSpeechDataset
					Clean	3480		
					Offensive/Hate	1860		
				3-way classification (Clean(C) vs Offensive(O) vs Hate(H))	Category	# Tweets		
					Clean	3480		
					Offensive	437		
					Hate	1423		
				6-way classification Clean(C) vs Offensive(O) vs GenderHate(GH) vs ReligiousHate(RH) vs NationalityHate (NH) vs EthnicityHate(EH)	Category	# Tweets		
					Clean	3480		
					Offensive	437		
					GenderHate	352		
					ReligiousHate	321		
					NationalityHate	368		
					EthnicityHate	382		
private dataset (Elzayady et al. 2022)	Facebook, Twitter, YouTube, and Instagram	20,000	Manual	hateful or non-hateful	Category	# Tweets	NO	–
					Hateful	10,000		
					Non-hateful	10,000		
Private dataset (Al-Hassan and Al-Dossari 2021)	Twitter	11k	Manual	none, religious, racial, sexism or general hate	Category	# Tweets	NO	–
					none	71%		
					religious	6%		
					racial	5%		
					sexism	6%		
					general hate	12%		

Table 6 (continued)

Reference	Platform	#Records	Annotation	Labels	Data distribution		Availability	Download links
ArabicCommentsFromYouTube (Boulouard et al. 2022a, b)	YouTube	15,050	Manual	after preprocessing 11,268 YouTube comments with 4748 hateful and 6520 non-hateful comments			YES	https://onedrive.live.com/?authkey=%21ACDXj%5FZNcZPqzy0&id=6EF6951FBF8217F9%21105&cid=6EF6951FBF8217F9
ArHS dataset and combination of ArHS + L-HSAB + OSACT4 datasets (Duwairi et al. 2021)	Twitter	ArHS: 9833	Manual	Misogyny, racism, religious discrimination, abusive, normal	Category	# Tweets	No	–
					Misogyny	925		
					Racism	2447		
					Religious discrimination	1410		
					Abusive	855		
					Normal	4196		
		Combined: 23,678	Manual		Category	# Tweets	YES	L-HSAB + OSACT4 are available (mentioned above)
					Misogyny	995		
					Racism	3013		
					Religious discrimination	1561		
					Abusive	3855		
					Normal	14,265		
OSACT2022 shared task (OSACT5) (sub-task A and sub-task B) (Makram 2022; AlKhamissi 2022; Mostafa 2022; De Paula 2022)	Twitter	13k	Manual	Sub-task A: off, not off Sub-task B: HS, not HS Sub-task C: not HS, HS1 (Race), HS2 (Religion), HS3 (Ideology), HS4 (Disability), HS5 (Social Class), and HS6 (Gender) Not VLG, VLG Not VIO, VIO	only sub-task A and sub-task B labels used:35% are offensive, and 11% are hate speech		YES	https://github.com/kirollos-hany/OSACT2022-source-code
Arapersonality dataset + OSACT4 EL-Zayady et al. (2023, 2023)	Twitter	~294,400	Manual	Arapersonality labels: (O C E A N) openness, conscientiousness, extraversion, agreeableness, and neuroticism	Arapersonality dataset consists of 92 users tweets each user has an average around 3200 tweets		YES	https://www.kaggle.com/datasets/e825c59935cdf1dc138408b4a27b055044321415304569cb9f0aaded163d4c38

Table 6 (continued)

Reference	Platform	#Records	Annotation	Labels	Data distribution		Availability	Download links
OSACT2022 shared task (OSACT5) (sub-task A and sub-task B) (Althobaiti 2022), (Alzubi 2022)	Twitter	13k	Manual	Sub-task A: off, not off Sub-task B: HS, not HS Sub-task C: not HS, HS1 (Race), HS2 (Religion), HS3 (Ideology), HS4 (Disability), HS5 (Social Class), and HS6 (Gender) Not VLG, VLG Not VIO, VIO	Category Off Not off HS Not HS HS1 HS2 HS3 HS4 HS5 HS6 Not HS	# Tweets 4463 8235 1339 11,359 288 31 158 1 82 508 9089	YES	https://github.com/kirollos-hany/OSACT2022-source-code
Seed corpus of Combined Arabi-cHateSpeechDataset (HS1) andOSACT (HS2) + private large unlabeled corpus Alsafari and Sadaoui (2021a, b)	Twitter	combined (13,140)	Manual	clean, offensive / hateful	HS1: 3480 clean 1860 offensive/hateful HS2: offensive 1915 Vulgar 225 Hate speech 506 Clean 8085		HS1 and HS2 are available	https://github.com/sbalsefri/ArabicHateSpeechDataset https://sites.google.com/site/offensevalsharedtask/multilingual
		Private (5million)			Private: unlabeled		NO	–
Private dataset (Alotaibi and Abul Hasanat 2020)	Twitter	Not clarified	Manual	racist, non-racist	Not clarified		NO	–

Table 6 (continued)

Reference	Platform	#Records	Annotation	Labels	Data distribution		Availability	Download links
Private Levantine dataset (AbdelHamid et al. 2022)	Twitter	17,554	Manual	normal, hate	Category	# Tweets	NO	–
					Normal	16,683		
					Hate	861		
					Appearance	905		
					Intellectual	769		
					Racial	2349		
					Sexual	482		
Tun-EL Badri et al. 2022)	Facebook, Twitter and YouTube	23,033	Manual	Normal, hate, and abusive	Category	# Rec	YES	https://github.com/NabilBADRI/Multidialect-Project
					Normal	12,353		
					hate	6830		
					Abusive	3850		
Private cyberbullying dataset (AlFarah et al. 2022)	Twitter, YouTube	24,560	Manual	cyberbullying, non-cyberbullying	Category	# Tweets	NO	–
					cyberbullying	12,280		
					non-cyberbullying	12,280		
					Offensive	9233		
					Not offensive	5817		
Private dataset (Omar et al. 2020)	Facebook, Twitter, YouTube, and Instagram	20,000	Manual	hateful or non-hateful	Category	# Tweets	NO	–
					Hateful	10,000		
					Non-hateful	10,000		
Egyptian Tweets dataset + private dataset (Mubarak and Darwish 2019)	Twitter	1100	Manual	Offensive or not offensive	Not clarified		YES	http://alt.qcri.org/~hmubarak/offensive/TweetClassification-Summary.xlsx
		36.6 M	Automatically tagged				NO	–

Table 6 (continued)

Reference	Platform	#Records	Annotation	Labels	Data distribution		Availability	Download links
AraHate (Albadi et al. 2019; Berrimi et al. 2020)	Twitter	6000	Crowd-Flower	Hate or not hate	Category	# Tweets	YES	https://github.com/nuhaalbadi/Arabic_hatespeech
					Hate	2526		
					Non-hate	3474		
Private dataset (Faris et al. 2020)	Twitter	3696	Manual	Hate, normal and neutral	Category	# Tweets	NO	–
					Hate	843		
					Normal	791		
					Clean	2062		
SemEval-2020 Task (English dataset) SemEval-2020 Task (Arabic dataset) (El-Alami et al. 2022)	Twitter	English 5994 Arabic 7800	Manual	Offensive, not offensive	Category	# Tweets	YES	https://sites.google.com/site/offensevalsharedtask/solid
					Offensive	3193		
					Not offensive	2991		
				3-way classification (Clean(C) vs offensive(O) vs Hate(H))	Category	# Tweets		https://sites.google.com/site/offensevalsharedtask/multilingua
					Clean	3480		
					Offensive	437		
					Hate	1423		
				6-way classification Clean(C) vs offensive(O) vs GenderHate(GH) vs ReligiousHate(RH) vs Nationality Hate(NH) vs Ethnicity Hate(EH)	Category	# Tweets		
					Clean	3480		
					Offensive	437		
					GenderHate	352		
					ReligiousHate	321		
					NationalityHate	368		
					EthnicityHate	382		
Combined datasets (L-HSAB, Multi-Platform Arabic News Comment MPOLD, ArabicCommentsFromYouTube) (Awane et al. 2021)	Twitter, Facebook, YouTube	38,654	Manual	Hate, neutral	Category	# Rec	YES	https://drive.google.com/file/d/1MCXY5eyI7myKyQQ2ZPpI1RHPZR7Emd2e/view?usp=sharing
					Hate	28,898		
					Neutral	9756		
(OSACT4) (Mubarak 2021)	Twitter	10,000	Manual	Offensive, vulgar, hate speech, clean	Category	# Rec	YES	https://alt.qcri.org/resources/OSACT2020-sharedTask-CodaLab-Train-Dev-Test.zip
					offensive	1915		
					Vulgar	225		
					Hate speech	506		
					Clean	8085		

Table 6 (continued)

Reference	Platform	#Records	Annotation	Labels	Data distribution		Availability	Download links
(OSACT4) (Berrimi et al. 2020; Husain and Uzuner 2022a, b; Husain 2020)	Twitter	10,000	Manual	Offensive or not offensive	Category Offensive Not offensive	# Tweets 1991 8009	YES	https://sites.google.com/site/offensevalsharedtask/multilingual
OSACT4, L-HSAB (Husain and Uzuner 2022a, b)	Twitter	OSACT4: 10,000	Manual	Offensive or not offensive	19% of the entire dataset is offensive tweets, and 5% contains hate speech (468 tweets), (1728 tweets), or (3650 tweets)		YES	https://sites.google.com/site/offensevalsharedtask/multilingual
		L-HSAB: 5648		Hate speech, abusive language, normal	Category Hate Abusive Normal	# Tweets 468 1728 3650	YES	https://GitHub.com/Hala-Mulki/L-HSAB-First-Arabic-Levantine-HateSpeech-Dataset
Private YouTube dataset (Guellil et al. 2020)	YouTube	3384	Manual	Hateful, non-hateful	Category Hateful Non-hateful	# Records 919 2465	NO	–
GHSD, RHSD (Alshalan and Al-Khalifa 2020)	Twitter	GHSD: 9316	Manual	Hate, non hate	Category Hate Non hate	# Tweets 2539 2539	YES	https://github.com/raghadsh/Arabic-Hate-speech
		RHSD: 600			Category Hate Non hate	# Tweets 250 350		
Private dataset (Anezi 2022)	Facebook, Twitter, and others	4203	Manual	Religion, racist, gender equality, insulting/bullying, violent, normal positive, normal negative	Category Religion Racist Gender equality Insulting/bullying Violent Normal positive Normal negative	# Rec 407 394 410 404 788 900 900	NO	–
		Private (5million)						
Private dataset (Aljuhani et al. 2022)	Twitter	~30k tweets	Manual	Offensive or clean	Category Offensive Clean	# Tweets 14,659 15,231	NO	–

Table 6 (continued)

Reference	Platform	#Records	Annotation	Labels	Data distribution	Availability	Download links
(OSACT4) (Haddad 2020)	Twitter	10,000	Manual	Off/HS, not_off/not_HS	only 5% of tweets are labeled as hate speech while 19% of the tweets are labeled as offensive and the other 81%	YES	https://sites.google.com/site/offensevalsharedtask/multilingual
arHateDataset (Khezzar et al. 2023)	Twitter	34,107	Manual	Hate, normal	32% are hate tweets and the remaining 68% are normal tweets	YES	https://github.com/ramzi-kh/arHateDetector
Combined dataset (Khairy et al. 2023)	Facebook & Twitter	12,000	Manual	Cyberbullying, non-cyberbullying	6000 instances of cyberbullying and 6000 instances of non-cyberbullying	YES	https://github.com/omammar167/Arabic-Abusive-Datasets
Fine-grained hate speech Detection on Arabic Twitter (Muaad et al. 2023)	Twitter	17,655	Manual	Hate, non-hate	Category: # Tweets Hate: 542 Non-hate: 17,113	YES	https://sites.google.com/view/arabichate2022/home
MC_TunNS (Abbes et al. 2023)	Facebook & Twitter	20,000	Manual	Abusive, hate, normal, racism, sexism, gender, religion	T-HSAB: abusive, hate, normal 6 k TNHS: abusive, hate, normal 12 k The new dataset: racism, sexism, gender, religion, normal 2 k	NO	T-HSAB is only available: https://github.com/Hala-Mulki/T-HSAB-A-Tunisian-Hate-Speech-and-Abusive-Dataset
Egyptian-Arabic dialect (Ahmed et al. 2022)	Publicly available Arabic sentiment, offensive language and hate speech datasets Social media posts, tweets and comments A conducted Google survey	~8000	Manual	Neutral, offensive, sexism, religious discrimination, and racism	Category: # Records Neutral: 2135 Offensive: 1531 Sexism: 1502 Religious discrimination: 1500 Racism: 1501	NO	–

Table 6 (continued)

Reference	Platform	#Records	Annotation	Labels	Data distribution		Availability	Download links
OSACT 2020 (Mohamed et al. 2023)	Twitter	~10,000	Manual	Hateful and non-hateful	Category	# tweets	YES	https://sites.google.com/site/offensevalsharedtask/multilingual
					Hateful	506		
					Non-hateful	9497		
Combined Dataset (OSACT 2022 + OSACT4 + Multi-Platform dataset + ArCybC) (Al-Dabet et al. 2023)	Facebook, Twitter and YouTube	31,203	Manual	Offensive and not-offensive	Category	# Records	YES	https://sites.google.com/site/offensevalsharedtask/multilingual https://github.com/shammur/ArabicOffensive-Multi-Platform-SocialMedia-Comment-Dataset https://github.com/kirollos-hany/OSACT2022-source-code https://data.mendeley.com/v1/datasets/z2dfgrzx47/draft?a=12a9ff5d-6c5c-4b2e-8990-7d044d7c12e2
					Offensive	9016		
					Not-offensive	21,297		

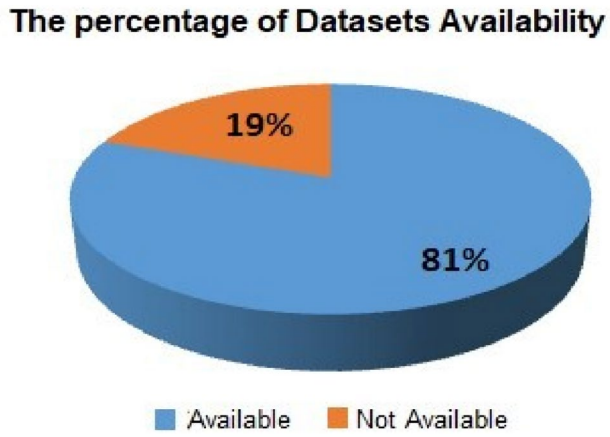

Fig. 7 The percentage of datasets availability

algorithms gave excellent results. Therefore, they applied a hybrid approach to these two classifiers using a genetic algorithm (GA), namely GA-SVM and GA-XGBoost, to reduce the time and cost and mitigate the challenges of optimizing their hyperparameters.

The SVM algorithm with the Aravec SkipGram word embedding model achieved superior results in terms of accuracy (88.2%) and F1-score rate (87.8%).

Similarly, the authors in (Shannag et al. 2022) presented the development and evaluation of a multi-dialect and annotated Arabic cyberbullying corpus (ArCybC) for detecting and analyzing cyberbullying in Arabic. They highlighted the lack of annotated Arabic cyberbullying data as a hindrance to the development of effective detection models. To address this, they introduced machine learning models and experimented with techniques such as support vector machine (SVM), random forest (RF), XGBoost, decision tree (DT), and logistic regression (LR) using both TF-IDF and Aravec word embedding. The authors used the same corpus in (Shannaq et al. 2022), and the results of the experiments reveal that the SVM model with word embedding performed the best, achieving an accuracy rate of 86.3% and an F1-score rate of 85%.

In another cyberbullying study (AlFarah et al. 2022), the authors focused on the detection in the Arabic language using machine learning techniques of cyberbullying. They identified the challenges of working with an imbalanced dataset, where the number of cyberbullying instances is significantly lower than the number of non-bullying instances, and proposed the use of sampling techniques such as SMOTE to overcome this issue. The authors used a dataset of 24,560 Arabic tweets and comments collected from Twitter and YouTube and oversampled the minority class to balance the data. They also compared the performance of various machine learning algorithms and found that Naïve Bayes achieved the highest AUC at 89%. The proposed approach shows promise in effectively detecting cyberbullying in Arabic tweets, despite the imbalanced nature of the dataset.

Moreover, the surveys (Khairy et al. 2021; ALBayari et al. 2021) reviewed cyberbullying classification methods for Arabic, classified into three categories: deep learning-based, machine learning-based, and hybrid. These reviews also highlighted the challenges posed by the Arabic language for natural language processing tasks as well as the growing interest in developing machine learning and deep learning models for detection. Contextual features such as sentiment analysis and user profiling were found to be more effective in capturing the nuances of the Arabic language. Results show that SVM and CNN are the most used algorithms, but the quality of datasets and features has a significant impact on performance.

Table 7 The taxonomy of the most frequent Arabic offensive language and hate speech detection datasets

Dataset	Reference
OSACT4	Elzayady et al. (2023a, b); Berrimi et al. (2020); Husain and Uzuner (2022a, b); Duwairi et al. (2021); Mubarak (2020); Husain (2020); Mubarak (2021); Haddad (2020); Mohamed et al. (2023); Al-Dabet et al. (2023)
OSACT5	Makram (2022); Althobaiti (2022); AlKhamissi (2022); Mostafa (2022); Alzubi (2022); De Paula (2022); Al-Dabet et al. (2023)
ArabicCommentsFromYouTube	Mohaouchane et al. (2019); Awane et al. (2021); Boulouard et al. (2022a, b); Alakrot et al. (2021); Boulouard et al. (2022a, b)
ArabicHateSpeechDataset	Alsafari et al. (2020a, b); Alsafari and Sadaoui (2021a, b)
L-HSAB	Berrimi et al. (2020); Husain and Uzuner (2022a, b); Mulki et al. (2019); Husain and Uzuner (2022a, b)
T-HSAB	Berrimi et al. (2020); Husain and Uzuner (2022a, b); Haddad et al. (2019)
AraHate	Berrimi et al. (2020); Albadi et al. (2018); Albadi et al. (2019)
ArCybC	Shannaq et al. (2022); Al-Dabet et al. (2023)
Arapersonality	Elzayady et al. (2023a, b)
Tun-EL	Badri et al. (2022)

Table 8 Summary of the included studies, contribution, techniques and superior results

Reference	Contribution	Algorithms	Dataset(s)	Results	
Shannaq et al. (2022)	An intelligent prediction system to detect the offensive language in Arabic tweets	Seven ML algorithms (NB, KNN, SVM, RF, XGBoost, DT, and LR).Fine-tuned pre-trained word embedding models (AraVec Unigram, AraVec N-gram, and GloVe) GA-XGBoost, GA-SVM	ArCybC	GA-SVM with Aravec SkipGram modelhas the highest accuracy 88.2% and F1-score rate of 87.8%	
Azzi and Zribi (2022)	comparing the performance of deep learning on the detection of eight specific subtasks of abusive language in Arabic social platforms	SVM, CNN, BiLSTM, BiGRU, BERT	Private dataset	BERT model achieved the best results precision (90%), micro-averaged F1- Score (79%)	
Berrimi et al. (2020)	a novel deep learning model based on the attention mechanism for smooth and accurate learning and classification to filter out offensive and abusive Arabic content on social media posts and comments	EL LSTM, ESoA and ELSoA	AJCommentsClassification-CF AraHate Offenseval2020-arabic Combination of L-HSAB and T-HSAB	ELSoA model has achieved the highest accuracy 97.47%	
Husain and Uzuner (2022a, b)	A transfer learning approach across different Arabic dialects for offensive language detection using BERT model	BERT (Pre-Trained AraBERT Model)	Egyptian Tweets L-HSAB T-HSAB OSACT	The Egyptian and Tunisian dialects gained better performance than Levantine in terms of accuracy 0.86% and F1 0.85%	
Alsafari et al. (2020a, b)	Single and ensemble deep learning classifiers for Arabic hate speech detection	Single learner classifiers CNN + AraBert, BiLSTM + AraBert Ensemble learner classifiers CNNs Average, BILSTMs Average	ArabicHateSpeechDataset Two-Class, Three-Class, Six-Class	Two-Class	F-Macro 88.90%
				Three-Class	F-Macro 79.21%
				Six-Class	F-Macro 76.97%
				Two-Class	F-Macro 88.88%
				Three-Class	F-Macro 78.84%
				Six-Class	F-Macro 76.40%
				Two-Class	F-Macro 91.12%
				Three-Class	F-Macro 84.01%
				Six-Class	F-Macro 76.08%
				Two-Class	F-Macro 90.75%
				Three-Class	F-Macro 79.91%
				Six-Class	F-Macro 80.23%

Table 8 (continued)

Reference	Contribution	Algorithms	Dataset(s)	Results		
Mohaouchane et al. (2019)	Evaluation and comparison for four different deep learning models for detecting offensive texts on social media platforms	Convolutional Neural Network (CNN). Bidirectional Long Short-Term Memory (Bi-LSTM).Bi-LSTM with attention mechanism. Combined CNN and LSTM	ArabicComments FromYou-Tube	F1-Score CNN, 84.05 Combined CNN-LSTM, 83.65		
Alhejaili et al. (2022)	an automatic way to detect hate speech in Arabic tweets during COVID-19 pandemic using machine learning models	Support vector machine (SVM), random forest (RF), logistic regression (LR), decision tree (DT), AdaBoost, k-nearest neighbors (KNN), Gaussian naïve Bayes (GNB)	Twitter dataset of 5408 tweets	The LR model achieved the highest accuracy up to 90.8% with unigram, and the AdaBoost model achieved the highest in precision at 90.8% with trigram		
Elzayady et al. (2022)	Two efficient models (Online Machine Learning and Deep Learning Models) for Improving the detection of Arabic hate speech	Passive-aggressive classifier (PAC) Bidirectional Gated Recurrent Unit with Attention (BI-GRU)	Private dataset	BI-GRU outperformed PAC BI-GRU: accuracy 99.1%, F1-score 99.1% PAC: accuracy 98.4%		
Duwairi et al. (2021)	The ability of different deep learning models to automatically detect hateful content on social media such as misogyny, racism, and religious discrimination along with abusive language and normal language	CNN, CNN-LSTM, BiLSTM-CNN	ArHS dataset, combined dataset (ArHS + L-HSAB + OSACT4)	Dataset	Class	Results
				ArHS	Binary	CNN acc. of 81%
					Ternary	both CNN, BiLSTM-CNN acc. of 74%
					Multiclass	CNN-LSTM, BiLSTM-CNN models acc. of 73%
				Combined	Binary	BiLSTM-CNN acc. of 73%
					Ternary	BiLSTM-CNN acc. of 67%
					Multiclass	CNN-LSTM, BiLSTM-CNN acc. of 65%

Table 8 (continued)

Reference	Contribution	Algorithms	Dataset(s)	Results
Makram (2022)	A hybrid machine learning approach consists of two classic machine learning classifiers (Logistic Regression, Random Forest) based on the Arabic pre-trained Bert language model MARBERT for feature extraction of the Arabic tweets of the first and second tasks of OSACT2022 for offensive and hate speech detection	Logistic regression, random forest	OSACT2022 shared task (OSACT5)	The results for the hate offensive the Logistic Regression model with accuracy, precision, recall, and f1-score of 80%, 78%, 78%, and 78%, respectively The results for the hate speech tweet detection task were 89%, 72%, 80%, and 76%
Al-Hassan and Al-Dossari (2021)	Comparing the performance of four deep learning models (LTSM model, Ensemble model of LTSM and layer of CNN, GRU model, ensemble model of GRU and a layer of CNN) with baseline model (SVM)	LTSM, CNN + LTSM, GRU, CNN + GRU and baseline SVM	Private dataset of 11 K tweets	CNN + LTSM enhanced the overall performance of detection with 72% precision, 75% recall and 73% F1 score

Table 8 (continued)

Reference	Contribution	Algorithms	Dataset(s)	Results
Althobaiti (2022)	New automatic method for detecting offensive language and fine-grained hate speech from Arabic tweets using BERT model with various levels of preprocessing, including cleaning, appending sentiments as additional textual features, and replacing emojis with their corresponding textual descriptions and comparing its performance with two other ML models (SVM, LR)	SVM + word n-grams + (TF-IDF), LR + word n-grams + (TF-IDF), AraBERT, QARiB, mBERT, XLM-RoBERTa, new proposed BERT-based model with some suggested preprocessing levels	OSACT2022 shared task (OSACT5)	The BERT models outperformed the SVM and LR classifiers The new proposed model achieved the best performance on all tasks: Offensive language detection with 84.3% F1-score, Hate speech detection with 81.8% F1-score, Fine-grained hate speech recognition (e.g., race, religion, social class, etc.) with 45.1% F1-score
Elzayady et al. (2023a, b)	A novel approach for detecting hate speech on Arabic social media based on personality traits for hate speech detection in Arabic social media	RF, extra trees, DT, SVM, gradient boosting, XGBoost, and (LR) LSTM, (BI-LSTM), (GRU) CNN-LSTM, CNN-BILSTM, and CNN-GRU) AraBERT based model	Arapersonality + OSACT4	The proposed AraBERT model achieved a macro-F1 score of 82.3%
Elzayady et al. (2023a, b)	A novel method for enriching the MARBERT model that incorporates static word embedding (AraVec 2.0) and personality trait features for Arabic hate speech detection	MARBERT with hybrid features	Arapersonality + OSACT4	Macro-F1 score of 86.4%

Table 8 (continued)

Reference	Contribution	Algorithms	Dataset(s)	Results
Alsafari and Sadaoui (2021a, b)	New approach for Arabic hate speech detection called semi-supervised self-learning (SSSL), which combines supervised and unsupervised learning to improve the accuracy of hate speech detection	CNN + SG, BiLSTM + SG, CNN + Bert, CNN + Bert	Seed corpus of Combined ArabicHateSpeechDataset (HS1) and OSACT (HS2) + private large unlabeled corpus	CNN + SG, F1-Score of 88.59%
Alakrot et al. (2021)	Machine learning approach to detecting offensive language in Arabic online communication	Logistic regression (LR), support vector machines (LinearSVC) models with L1 regularization and selecting features based on their regularized weights, feature ranking with recursive feature elimination using logistic regression (RFE), decision tree classifier (ExtraTreesClassifier), singular-value decomposition (SVD)	ArabicCommentsFromYouTube	The approche RFE∪LR-L1 achieved superior results with reasonable overall accuracy of 0.84, and precision, recall and F1-score of 0.89, 0.76 and 0.81, respectively
Alotaibi and Abul Hasanat (2020)	New model for detecting racism in Arabic tweets using deep learning and text mining techniques	CNN	Private dataset	Only the authors indicated that their proposed model outperformed the statistical machine learning models
AbdelHamid et al. (2022)	A dataset for Levantine Arabic hate speech detection in OSNs, and utilize various models and algorithms to detect hate speech, including a hybrid approach that combines both deep learning and traditional machine learning techniques	RandomForest, support vector machines, XGB, catBoost, MLP and three deep learning classifiers AraBERT, ArabicBERT, GigaBERT	Private Levantine dataset	The best model is GigaBERT with 0.81 macro F1-score

Table 8 (continued)

Reference	Contribution	Algorithms	Dataset(s)	Results
Shannag et al. (2022)	Design, creation, and evaluation of a multi-dialect and annotated Arabic Cyberbullying Corpus (ArCybC) and a rigorous comparison of different machine learning approaches was applied	Support vector machine (SVM), RF, XGBoost, DT, and LR (using both TF-IDF and Aravec word embedding)	ArCybC	SVM with word embedding achieved superior accuracy rate of 86.3% and an F1-score rate of 85%
Badri et al. (2022)	Building a new Arabic hate speech and abusive language dataset (Tun-EL) covering three dialects (Tunisian, Egyptian, and Lebanese). Proposing a new approach for inappropriate content detection. Our approach is based on a combination of AraVec and fastText word embedding as input features and a Convolutional Neural Network-Bidirectional Gated Recurrent Unit (CNN-BiGRU) model	Baseline machine learning models, LR and RF.CNN-BiGRU + FastText + Aravec	Tun-EL	CNN-BiGRU + FT + Ara model was able to classify 88% of the hateful content and 76% of the abusive content
AlFarah et al. (2022)	Investigating Arabic cyberbullying detection using five machine learning techniques using real Arabic messages from Twitter and YouTube	logistic regression, decision tree, random forest, naïve Bayes (NB) and SVM	private cyberbullying dataset	The highest AUC achieved is 89% using naïve Bayes where SVM and logistic regression achieve 88%

Table 8 (continued)

Reference	Contribution	Algorithms	Dataset(s)	Results
Boulouard et al. (2022a, b)	Investigating hate speech detection in Arabic social media using five machine learning techniques using Arabic comments from YouTube	Logistic regression, random forest, naïve Bayes (NB) support vector machines, long short-term memory (LSTM)	ArabicCommentsFromYouTube	Superior results for LSTM with F1-Score 0.82%, followed by SVM with 0.76%
Omar et al. (2020)	Constructing a standard Arabic dataset for hate speech and abusive detection in online social networks. Comparing the performance of machine learning and deep learning models for hate speech detection in OSNs	ML: MultinomialNB, complement NB, BernoulliNB, SVC, NuSVC, LinearSVC, LogisticRegression, decision tree, SGD, Ridge, perceptron, and nearest centroid DL: CNN, RNN	Private dataset	Recurrent neural network (RNN) outperformed all other classifiers with an accuracy of 98.7%
Mubarak and Darwish (2019)	A method for building a large training corpus for detecting offensive language based on a seed word list of offensive words. Comparing a character n-gram deep learning model to build a robust offensive language classifier for Arabic tweets with SVM classifier	Word-list (baseline), FastText, and SVM	Egyptian tweets dataset + private dataset	FastText classifier led to identical precision, recall, and F1 measure of 90%

Table 8 (continued)

Reference	Contribution	Algorithms	Dataset(s)	Results
Albadi et al. (2019)	Create the first public Arabic dataset of tweets annotated for religious hate speech detection Create three public Arabic lexicons of terms related to religion along with hate scores Investigating the effect of combining GRU neural networks with handcrafted features for religious hatred detection on Arabic Twitter space	AraHate-PMI, AraHate-Chi, AraHate-BNS, logistic regression, SVM, GRU + word embeddings, GRU with word embeddings + handcrafted features	AraHate	GRU + word embeddings + handcrafted features achieved superior results for detecting religious hatred in Arabic in terms of Recall of 0.84%
El-Alami et al. (2022)	A multilingual offensive language detection method based on transfer learning from transformer fine-tuning model	SVM-Khi2, CNN, LSTM, GRU, BiLSTM, CNN-BiLSTM, Multilingual Elmo, mBERT, CNN-GRU, BiLSTM, AraULMFiT, AraBERT	SemEval-2020 Task (English dataset) SemEval-2020 Task (Arabic dataset)	Arabic BERT (AraBERT) achieves over 93% and 91% in terms of F1-score and accuracy, respectively
Alsafari et al. (2020a, b)	An extensive empirical analysis by evaluating a variety of feature selection methods within a supervised classification framework, including machine and deep learning methods	Unigram, Word-ngrams, Char-ngrams, Word/Char-ngrams, WE (fastext), WE (aravec-cbow) Support Vector Machines (SVM), Naive Bayes, Logistic Regression, and Random Forest, Fastext, AraVec-cbow, AraVec-skipgram, mBert CNN, LSTM and GRU	ArabicHateSpeechDataset Two-class, Three-class, Six-class	SVM outperformed naive Bayes and Logistic Regression across three tasks and all feature extraction methods. CNN + mBert model outperformed all other learned models across the three prediction tasks, with 87.05% for the 2-class task, 78.99% for the 3-class task, and 75.51% for the 6-class task
Alzubi (2022)	An ensemble based approach to detecting offensive tweets	Char-tfidf, Word-tfidf, Emoji, Muse, Emoji-Score, AraBERT, Mazajak, Char + word + MUSE, Char + word + MUSE + Emoji, Ensemble of Boldface Models (Emoji-Score, AraBERT, Char + word + MUSE + Emoji)	OSACT5	Macro F1: 0.85

Table 8 (continued)

Reference	Contribution	Algorithms	Dataset(s)	Results
Faris et al. (2020)	Investigating the detection of hate speech in the Arabic language using word embedding and deep learning techniques	word2vec(cbow), word2vec(cbow), Aravec(N_grams&cbow), Aravec(N_grams&cbow), Aravec(N_grams&SG), Aravec(N_grams & SG) CNN, RNN, LSTM	Private dataset	The recurrent convolutional networks was very good and competent and achieves high accuracy and F1-score of 71.688%
Mostafa (2022)	Deep learning ensemble learning system consisting of three different deep learning models for offensive text detection. Trying to overcome the loss function convention for data-imbalanced Arabic offensive text detection by experimenting with alternative loss functions rather than using the traditional weighted cross-entropy loss	MARBERT (Without emojis), AraBERT-Large-Twitter, QARiB, AraBERT-Base-Twitter, MARBERT, MARBERTV2, LightGBM(QARiB Embeddings), Ensemble(LightGBM + MARBERT + MARBERTV2), Ensemble(AraBERT-BT + MARBERT + QARiB), Ensemble(MARBERTV2 + MARBERT + QARiB)	OSACT5	Ensemble(MARBERTV2 + MARBERT + QARiB) of Macro F1: 87.044
Awane et al. (2021)	Investigation of BERT model for detection of hate speech in the Arab electronic press and social networks. Building a large corpus which is combination of three hate speech datasets, made up of texts in classical Arabic, Levantine, and North African dialect	BERT	Combined dataset of three hate speech datasets	F1-score of 89%
De Paula (2022)	A combination of transformer-based models and ensemble techniques to improve the classification performance	AraBert, AraElectra, Albert-Arabic, AraGPT2, mBert, and XLM-Roberta. Two ensemble methods: majority vote and highest sum	OSACT5	In terms of F1-Macro, the the AraBert model in Task-A achieved 0.827, Highest sum ensemble achieved 0.792 in Task-B, and finally in Task-C, AraBert achieved 0.423

Table 8 (continued)

Reference	Contribution	Algorithms	Dataset(s)	Results
AlKhamissi (2022)	An ensemble of several trained models each of which uses a different set of hyper-parameters for Arabic hate speech detection	AraHS	OSACT5	F1-Macro: 82.7% on Hate Speech Detection (HSD) subtasks
Mubarak (2021)	Building the largest Arabic offensive language dataset, performed analysis to identify peculiarities, and experimented with SOTA classification techniques to detect offensive language. It is not biased by topic or dialect	Lexical features: SVM Pre-trained static embeddings: fastText/SVM, AraVec/SVM, Mazajak/SVM Embeddings trained: fastText/fastText Contextualized embeddings: BERT base-multilingual, AraBERT Mazajak embeddings: decision tree, random forest, Gaussian NB perceptron, AdaBoost, gradient boosting, logistic regression, SVM	OSACT4	AraBERT, achieving an F1 score of 83.2
Haddad (2020)	Investigating attention-based deep neural networks for Arabic offensive language detection, and comparing the results with State-Of-Art ML classifiers	DL classifiers CNN, Bi-GRU, CNN_ATT, Bi-GRU_ATT Baseline ML classifiers Bow_LR, Tf-idf_LR, Bow_Ridge, Bow_SVM, Tf-idf_Ridge, Tf-idf_SVM	OSACT4	(BiGRU_ATT) achieved 0.859 F1 score for the task of offensive language detection, and 0.75 F1 score for the task of hate speech detection
Husain et al. (2020)	Deep learning approach for Arabic offensive language detection and comparing the results with LR classifiers	TF-IDF feature, AraVec word embeddings RNN, a GRU, a Bi-GRU, an LSTM, and a Bi-LSTM, Logistic regression (LR)	OSACT4	The Bi-directional Gated Recurrent Unit (Bi-GRU) based model, reports a macro-F1 score of 0.83
Husain and Uzuner (2022a, b)	Investigating the impact of preprocessing on Arabic offensive language classification using traditional ML classifiers and deep learning classifiers	SVM, LR, Random Forest, Bagging RNN (BOW), LSTM (BOW) RNN (AraVec), LSTM (AraVec) AraBERT, Arabic-BERT	OSACT4, L-HSAB	The AraBert model Achieved best results in all tasks with maximum Micro-Average F1 of 0.82

Table 8 (continued)

Reference	Contribution	Algorithms	Dataset(s)	Results
Alshalan and Al-Khalifa (2020)	A novel approach for automatic hate speech detection in the Saudi Twittersphere using deep learning techniques creating a new public dataset. Comparing the performance of three neural network models, and evaluating BERT for the Arabic hate speech detection task	SVM (char n-grams), LR (char n-grams) CNN, GRU,CNN + GRU, BERT	GHSD, RHSD	CNN successfully outperformed other models, with an F1-score of 0.79 and AUROC of 0.89
Boulouard et al. (2022a, b)	A transfer learning solution to detect hateful and offensive speech on Arabic websites and social media platforms	BERTEN, AraBERT, LinearSVC, mBERTAR, LSTM, mBERTEN	ArabicCommentsFromYouTube	BERTEN provided the best results with 98% accuracy, closely followed by AraBERT with 96% accuracy
Anezi (2022)	Building a new hate speech dataset in Arabic with seven different classes and modified machine learning algorithms to correctly classify and detect hate speech. The system has various advantages and uses, such as being used in Arabic-speaking countries to ensure peace and tolerance, saving lives, being used in smart cities, being expanded to include other Arabic dialects, and being used by western countries for Arabic-speaking populations	DT, MLP, NB, LR, RF DRNN-2 DRNN-1	Private dataset	A recognition rate of 99.73% was achieved for binary classification, 95.38% for the three classes of Arabic comments, and 84.14% for the seven classes of Arabic comments

Table 8 (continued)

Reference	Contribution	Algorithms	Dataset(s)	Results
Alsafari and Sadaoui (2021a, b)	A semi-supervised classification approach with self-training to leverage the abundant social media content and develop a robust offensive and hate speech classifier	SVM + WCNG, CNN + W2VSG, BiLSTM + W2VSG, CNN + AraBert, BiLSTM + AraBert, CNN + DistilBert, BiLSTM + DistilBert	Seed corpus of Combined ArabicHateSpeechDataset (HS1) and OSACT (HS2) + private large unlabeled corpus	Self-training approach outperformed the baseline model The CNN + W2VSG achieved F1-Score 89.51
Guellil et al. (2020)	An approach for hate speech detection against politicians in the Arabic community on social media (e.g. YouTube)	(Word2vec & fastText), GNB, LR, RF, SGD LSVC, CNN, MLP, LSTM, and Bi-LSTM	Private YouTube dataset	LSVC, BiLSTM and MLP achieved an accuracy up to 91%, when it is associated with the SG model
O. Aljuhani et al. (2022)	A new dataset for Arabic tweets containing both offensive and non-offensive language for detecting offensive language on Twitter. Proposing a new approach for constructing and labeling the dataset, as well as building domain-specific word embedding for the Arabic offensive language domain A deep learning-based approach that combines bidirectional long short-term memory (BiLSTM) with domain-agnostic and domain-specific word embedding to detect offensive Arabic language	SVM and LR with word-n-gram and character n-gram features. BiLSTM used three embedding models (AraVec, ArOffW2V, blending model of AraVec, and ArOffW2V)	Private dataset	BiLSTM with Blend Embeddings achieved superior results in terms of F1-Score of 0.93

Table 8 (continued)

Reference	Contribution	Algorithms	Dataset(s)	Results
Khairy et al. (2023)	Creating a new Arabic balanced dataset to be used in the offensive detection process, comparing the performance of single classifiers has been improved using ensemble machine learning	LR, KNN, Linear SVC, Bagging (Random Forest), boosting (Adaboost),voting ensemble	Combined dataset	Voting, F1-Score 98%
Muaad et al. (2023)	Transfer learning approach model compared with various machine learning (ML) and deep learning (DL) models	Passive-aggressive (PA), (LR), (RF), (DT), (K-NN (Linear SVC), (SVC), (NB), (BNB),extra tree classifier (ET), Ensemble bagging classifier, ensemble AdaBoost, ensemble gradient boosting classifier (GB), and (Arab-BiER	Fine-grained hate speech detection on Arabic Twitter	AraBERT, F1-Score 79%
Abbes et al. (2023)	A deep-learning solution to find hateful and offensive speech on Arabic social media sites like Facebook Collecting a new dataset, the first publicly available Tunisian dataset for hate speech from Facebook	Bi-LSTM + attention AraBERT	MC_TunNS	AraBert model provided the best results with 97.84% accuracy, followed by BI-LSTM with the attention mechanism at 93.17% accuracy
Ahmed et al. (2022)	Fine-tuning different Arabic pre-trained transformer models for Egyptian-Arabic dialect offensive language and hate speech detection and classification Collecting new Egyptian-Arabic dialect custom dataset	bert-large-arabertv02-twitter, Bert-base-arabic-camelbert-mix, MARBERTv2, Araelectra-base-discriminator, and Bert-base-multilingual-uncased	Egyptian-Arabic dialect	Achieving an average accuracy of about 96% across all fine-tuned transformer models

Table 8 (continued)

Reference	Contribution	Algorithms	Dataset(s)	Results
Mohamed et al. (2023)	A comprehensive evaluation of oversampling techniques and their impact on the performance of deep learning models in imbalanced datasets Applying a focal loss function that penalizes wrong predictions on the minority class in an attempt to balance the data Employing three different transformer models: MARBERT v2, MARBERT v1, and ARBERT For fine-tuning those models, they applied a deep learning architecture named QRNN that merged the sequential approach of recurrent neural networks (RNNs) with the parallel processing method of convolutional neural networks (CNNs) for handling input tokens Finally, they applied the majority vote ensemble approach using the pre-trained models fine-tuned with QRNN	MARBERT v2, MARBERT v1, and ARBERT	OSACT 2020	The proposed ensemble model achieved a macro-F1 score of 91.6%

Table 8 (continued)

Reference	Contribution	Algorithms	Dataset(s)	Results
Al-Dabet et al. (2023)	A proposed approach employs versions of the CAMeLBERT model and is validated using a mixture of four benchmark Twitter Arabic datasets annotated for hate speech detection task,	CAMeL-BERT-Da CAMeL-BERT-Ca CAMeL-BERT- Msa CAMeL-BERT-Mix	Combined dataset (OSACT 2022 + OSACT4 + multi-platform dataset + ArCybC)	CAMeLBERT mix version achieved superior results with 87.15% accuracy and 83.6% F1 score

In another study (Azzi and Zribi 2022), the authors aimed to investigate various state-of-the-art models for detecting abusive language in Arabic social media. They conducted their experiments to detect eight specific subtasks of abusive language in Arabic social platforms, namely racism, sexism, xenophobia, violence, hate, pornography, religious hatred, and LGBTQa hate, using CNN, BiLSTM, and BiGRU deep neural networks with pre-trained Arabic word embeddings (AraVec) and also pre-trained Arabic word embeddings and a BERT model comparing the results with an ML-based algorithm (SVM). They compiled a dataset from two famous platforms, which are Twitter and YouTube. The dataset consists of 6000 records. They performed manual annotation for it; 1914 out of the 6000 lines (31%) were labelled as normal, while the rest were marked as abusive. The result shows that CNN, BiLSTM, BiGRU, and BERT have outperformed the base ML classifier SVM, and the BERT model achieved the best results in terms of precision (90%) and micro-averaged F1-Score (79%).

Unlike, a more specific dataset was presented in (Alsafari et al. 2020a, b) for hate and offensive speech, containing 5340 records collected from Twitter. It was written in the most common Arabic languages: the Gulf Arabic dialect, spoken by the Arabian Peninsula countries, and modern standard Arabic, understandable by all Arabic speakers. This corpus has been divided among two-class, three-class, and six-class labelling datasets. The authors proposed single and ensemble artificial neural network (ANN) architectures, CNN and BiLSTM that are trained with different word embedding techniques, non-contextual: Fasttext-SkipGram, and contextual: multilingual Bert (MBERT) and AraBert. The challenge was a six-class classification setting where the goal was not only to detect the existence of hate but also the type of hate.

The experiments showed that for single learners, CNN + AraBERT is the best single classifier on each classification task, and the two contextualized word embeddings, AraBert and MBert, outperformed the non-contextualized FastText with both ANN models. For an ensemble of learners, the ensemble models perform better than the single models across all performance metrics. The CNN-Average improves the performance across two-class and three-class labels, but for the challenge of six-class classifications, the average-based BiLSTMs ensemble model obtained an F1-Macro of 80.23%, which outperformed the average-based CNN ensemble. For future work, they aim to try semi-supervised classification.

Reference (Berrimi et al. 2020) worked on a novel deep learning model based on the attention mechanism for smooth and accurate learning and classification to filter out offensive and abusive Arabic content on social media posts and comments. They used four available Arabic datasets from some previous studies (Mubarak et al. 2017; Albadi et al.

2018; Mulki et al. 2019; Haddad et al. 2019) related to inappropriateness in the Arabic language. These datasets are collected from different platforms as follows: the first one was obtained from comments deleted from Aljazira.com. It contains 32k comments that were annotated using CrowdFlower; these comments were labelled as obscene, offensive, or clean. The second, AraHate dataset, consists of 6000 tweets collected from Twitter and labelled as "hate" or "not hate". Third, they used the Subtask 'A' dataset shared within the 4th Workshop on open-source Arabic Corpora and Processing Tools (OSACT4). It contains 10,000 tweets that were manually annotated and labelled as OFF or NOT OFF. Finally, they combined two datasets, namely, L-HSAB (5846 tweets) and T-HSAB (6024 records collected from Facebook and YouTube), to obtain a larger dataset of different Arabic dialects of abusive and hate speech. The authors proposed a soft attention mechanism to detect different types of inappropriate speech by applying three models, namely, EL LSTM, ESoA, and ELSoA, to each dataset. The results indicated that the ELSoA model has achieved superior accuracy of 97.47%.

Also, F. Husain et al. (Husain and Uzuner 2022a, b) worked on the same dialectal datasets (L-HSAB, T-HSAB, and OSACT4) used in (Berrimi et al. 2020), except the authors used the Egyptian Tweets dataset instead of Aljazira.com. The Egyptian Tweets dataset consists of 1100 records collected from Twitter and labelled as offensive or not offensive. They proposed a transfer learning approach across different Arabic dialects for offensive language detection using the BERT model. They built on the pre-trained AraBERT model using the above dialectal datasets for fine-tuning and evaluating the model to see the effect of different Arabic dialects on offensive language detection. The experiments indicated that the Egyptian and Tunisian dialects gained better performance than Levantine in terms of accuracy of 0.86% and $F1$ rate of 0.85%.

On the other hand, the study (Mohaouchane et al. 2019) aimed to fill this gap in Arabic offensive language detection on social media. The authors proposed four different neural network models, namely: convolutional neural network (CNN), bidirectional long short-term memory (Bi-LSTM), attentional Bi-LSTM, and a combined CNN-LSTM model for detecting offensive texts on social media platforms. They used an available dataset of 15,050 records of Arabic YouTube comments taken from popular, controversial YouTube videos about Arab celebrities. The dataset was manually annotated, and the data was labelled either offensive or not offensive. The dataset was imbalanced, so the authors used the random oversampling technique to balance its classes and obtain accurate classification results. The combined CNN-LSTM network achieved the best recall rate of 83.46%, while it was clear that the CNN model achieved the best accuracy and precision rates of 87.84% and 86.10%, respectively.

Likewise, the authors in (Alhejaili et al. 2022) built a dataset during the COVID-19 pandemic period from January 31 to March 6, 2021, to provide an automatic way to detect hate speech in Arabic tweets during this pandemic using a variety of machine learning classifiers. The dataset was collected and preprocessed from Twitter and consists of 5408 tweets, which were then annotated as hate or not hate. They used TF-IDF for feature extraction and trained the dataset in three types: unigram, bigram, and trigram. The authors used a set of machine learning classifiers, namely support vector machine (SVM), random forest (RF), logistic regression (DT), decision tree, AdaBoost, k-nearest neighbours (KNN), and Gaussian naïve Bayes (GNB), to classify the content into hate or not hate. The seven classifiers did well, but the classifier LR achieved the highest performance in accuracy (Acc) of 90.8% with unigram. Otherwise, the AdaBoost model achieved the highest precision (P) at 90.8% with trigram. In the future, they aim to use deep learning models for Arabic hate speech detection during COVID-19 and compare the results with the above machine learning models results.

Elzayady et al. 2022 proposed two effective models using online supervised machine learning and deep neural networks, namely, passive-aggressive classifiers (PAC) and bidirectional gated recurrent units with attention (BI-GRU), to improve Arabic hate speech identification. The authors used the first Arabic hate speech multi-platform dataset. It was collected from four social media networks that contributed comments: Twitter, YouTube, Facebook, and Instagram. The dataset is well-balanced and consists of 20,000 posts, tweets, and comments, of which 10,000 are hateful and the other 10,000 are non-hateful. A variety of preprocessing steps for data preparation have been conducted. They used both term frequency-inverse document (TF-IDF) and pre-trained AraVec2.0 word embeddings as feature extraction techniques for text representation. The experiments were done and tested in Google Colab Pro by using NumPy, Pandas, Re, Alphabet Detector, Sklearn, and Keras packages. The results were assessed in terms of accuracy, precision, recall, and F1 score values. It was clear that the BI-GRU model outperformed PAC, where Bi-GRU with an attention layer provided an accuracy of 99.1% and PAC achieved 98.4%.

Moreover, Duwairi et al. 2021 proposed a deep learning framework for automatic detection of hate speech within Arabic tweets. The framework was developed using a hybrid approach of recurrent and convolutional neural networks, namely: CNN, LSTM-CNN, and BiLSTM-CNN, along with pre-processing techniques such as word-level (SG, CBOW) and sentence-level (pre-trained MUSE) embedding to represent and classify text data. The authors evaluated their

model on a large dataset of 23,678 Arabic tweets, which was compiled from three datasets: the Arabic Hate Speech (ArHS) dataset, the Levantine Hate Speech and Abusive (L-HSAB) dataset, and the 4th workshop on Open-Source Arabic Corpora and Processing Tools (OSACT4) shared task dataset, and compared its performance with other existing methods, demonstrating the effectiveness of the proposed framework in accurately detecting hate speech. The study highlights the potential of deep learning approaches for hate speech detection in languages other than English. The results showed that the SG-BiLSTM-CNN and SG-CNN were the best-performing models with the multi-class classification using the ArHS dataset.

In addition, the study by (Makram 2022) introduces machine learning and transformer-based models as a hybrid model for detecting offensive and hateful Arabic speech. The model consists of multiple classifiers, such as logistic regression and random forest; each specialized in detecting a specific type of offensive language. The authors trained the model on a dataset of Arabic social media posts using the Arabic pre-trained Bert language model MARBERT for feature extraction of the Arabic tweets in the dataset provided by the OSACT2022 shared task. The results were divided among hate and offensive classes, where the best results achieved for the offensive tweet detection task were achieved by the logistic regression model with accuracy, precision, recall, and f1-score of 80%, 78%, 78%, and 78%, respectively, while the results for the hate speech tweet detection task were 89%, 72%, 80%, and 76%. The authors also discussed the limitations and future directions for improving the model's performance. They also plan to investigate different machine learning classifiers such as SVM and Naive Bayes for the binary classification tasks using different representation models in the hope of achieving higher scores.

Also, Al-Hassan et al. (Al-Hassan and Al-Dossari 2021) presented a method for detecting hate speech in Arabic-language tweets using deep learning techniques. The authors collected a dataset of Arabic tweets consisting of 11k tweets and manually annotated them as five distinct classes: none, religious, racial, sexism, or general hate. The authors used the SVM model with TF-IDF word representation as a baseline for several deep learning models, namely, LTSM, CNN + LTSM, GRU, and CNN + GRU, to classify new tweets. The results showed that the proposed models achieved high accuracy in detecting hate speech in Arabic tweets, suggesting that these techniques can be useful for identifying and addressing hate speech in online Arabic-language communities. Overall, the ensemble model of CNN + LTSM obtained superior performance with 72% precision, 75% recall, and 73% F1 score.

Another study (Althobaiti 2022) proposed a new approach using the BERT model for hate speech and offensive language detection in Arabic tweets. The approach utilizes both emojis and sentiment analysis as appending features along with the textual content of the tweets in order to improve the accuracy of the detection. The authors compared their model with two conventional machine learning classifiers, support vector machine (SVM) and Logistic Regression (LR). They used the largest and most recently released dataset (OSACT 2022 shared task) for offensive language and hate speech detection in Arabic, which contains 12,698 tweets. The dataset was defined for three tasks: offensive language, hate speech, and fine-grained hate speech, which focus on specific types of hate speech. Various levels of preprocessing were done for data preparation, such as cleaning (CLN), appending sentiments (SA) as additional textual features, and replacing emojis (EmoTxt) with their corresponding textual descriptions. As a result, there are five versions of the dataset: original tweets, CLN, CLN + SA, CLN + EmoTxt, and CLN + SA + EmoTxt. They trained SVM and LR on these datasets' versions using word n-grams and TF-IDF, and they built five BERT models for each task as follows: AraBERT, QARiB, mBERT, XLM-RoBERTa, and their proposed model with its suggested preprocessing levels. The results of the experiments demonstrate that the proposed approach achieves high performance in detecting offensive language, hate speech, and fine-grained hate speech in Arabic tweets, with an F1-Score of 84.3%, 81.8%, and 45.1% for each task respectively.

Another work by Elzayady et al. (2023a, b) proposed a hybrid approach for hate speech detection in Arabic social media by combining machine learning algorithms and personality trait analysis. They collected a dataset of social media posts (the Arapersonality dataset and the OSACT dataset) and extracted linguistic features using natural language processing (NLP) techniques. They investigated the implementation of both machine learning models: RF, extra trees, DT, SVM, gradient boosting, XGBoost, and logistic regression (LR) and deep learning models: recurrent neural networks (RNNs) and CNN, namely, LSTM, bidirectional long short-term memory (BI-LSTM), a gated recurrent unit (GRU), and hybrids of CNN and RNN models (CNN-LSTM, CNN-BILSTM, and CNN-GRU). Then, they analyzed the personality traits of the authors and used them as additional features in their proposed AraBERT model. The proposed approach achieved promising results with a macro-$F1$ score of 82.3% compared to other state-of-the-art methods.

Similarly, in another study by (Elzayady et al. 2023a, b), the authors continued their work in (Elzayady et al. 2023a, b) using the same datasets, proposing a novel method for enriching the MARBERT model with hybrid features that incorporate static word embedding (AraVec 2.0) and personality trait features for Arabic hate speech detection. They implemented their experiments by fine-tuning the MARBERT model with hybrid features using the convolutional neural network (CNN) to be utilized for classification. The

results showed that they achieved outstanding outcomes for Arabic hate speech challenges, greatly surpassing previous studies, where the proposed model achieved a high-performance score in terms of macro-F1 score of 86.4% compared with the traditional MARBERT. In the future, the authors will need to extend their proposed methodology to include multi-personality trait features rather than binary ones and investigate sampling methods in greater depth to address the issue of imbalanced data. They will also try to improve their proposed model for future goals in Arabic hate speech classification using multi-task learning approaches.

Another trend is semi-supervised learning, which is a hybrid of supervised and unsupervised learning, combining labelled and unlabeled data to understand how it can change learning behaviour. It is of great interest in machine learning and data mining, as it can use readily available unlabeled data to improve supervised learning tasks. Several attempts were made to analyze the effectiveness of several semi-supervised learning approaches. For instance, (Alsafari and Sadaoui 2021a, b) proposed a new approach for Arabic hate speech detection called semi-supervised self-learning (SSSL). The authors used two datasets, a smaller seed dataset of labelled data and a larger unlabeled corpus of data, to train the model that can detect hate speech in Arabic text. The experiments for the SSSL framework consisted of three primary phases: training several pairs of deep learning classifiers with non-contextualized or contextualized word embedding models, labelling the unlabeled dataset using the optimal classifier artificially, and fine-tuning the baseline classifier. The results showed that the CNN + SG achieved superior performance in terms of an $F1$-Score of 88.59%.

Similarly, in (Alsafari and Sadaoui 2021a, b), the authors presented a semi-supervised self-training framework to detect hate and offensive speech on social media. The authors used the same datasets in (Alsafari and Sadaoui 2021a, b) to train the model. They conducted six groups of experiments to validate the SSST approach and selected the best classifier by assessing several text vectorization algorithms and machine learning algorithms. The results of the experiments showed that the self-training approach outperformed the baseline model, achieving higher accuracy, precision, and recall. The authors also found that ensemble-based selection of confident pseudo-labelled data achieved comparable results to classical self-training. Finally, the CNN + W2VSG achieved an $F1$-Score of 89.51.

In a further work (Alakrot et al. 2021), the authors introduced a novel approach to identifying offensive language in Arabic online communication using machine learning algorithms. Their dataset of 15,050 labelled YouTube comments served as a unique resource for future research on anti-social behaviour in Arabic online communities. The authors applied a series of text preprocessing, feature extraction, and feature selection techniques to represent the data, including word n-grams, character n-grams, and part-of-speech tags. Various classifiers, such as naive Bayes, support vector machines, and random forest, are trained to detect offensive language, and their performance is evaluated using precision, recall, and $F1$-score metrics. Additionally, the authors examined different methods for feature selection, including logistic regression, support vector machines with $L1$ regularization, and feature ranking with recursive feature elimination. The superior results from the RFE $\cup$ LR-L1 method demonstrate the efficacy of the machine learning approach in effectively detecting offensive language in Arabic online communication.

The issue of racism on social media platforms has become increasingly prevalent, and with it comes potential harm to individuals and society. To address this problem, the authors in (Alotaibi and Abul Hasanat 2020) propose a model for detecting racism in Arabic tweets using deep learning and text mining techniques. This automated tool applies convolutional neural networks (CNN) to classify Arabic tweets as either racist or non-racist, utilizing a Twitter dataset that contains both types of tweets. Preprocessing of the data involves cleaning and tokenizing the tweets, converting them to vectors, and feeding them into models for training and testing. The results demonstrate the effectiveness of deep learning and text mining techniques in detecting racism on Twitter, surpassing the performance of statistical machine learning models. Such models are crucial in mitigating the impact of harmful content on individuals and society and therefore represent a significant contribution to the field of social media analysis.

The Levantine Arabic dialect is very close to standard Arabic. The study presented in (AbdelHamid et al. 2022) is concerned with the detection and classification of hate speech in Arabic tweets from the Levant region. The authors highlighted the harmful effects of hate speech on individuals and society and argued for the need for automated and accurate hate speech detection methods. The authors utilized a variety of models and algorithms for detecting hate speech, including deep learning and traditional machine learning techniques. A hybrid approach was adopted that combined word embedding and TF-IDF features for traditional classification models and BERT models for deep learning models. The dataset used in the study was collected from Twitter using specific keywords related to hate speech from the Levant region and was manually annotated into two classes. The experiment results demonstrate that the concatenation of word embedding and TF-IDF can improve classification performance and that deep learning classifiers show superior performance compared to traditional ones. The best model, using GigaBERT, achieved an AUC-ROC curve of 94.6% and a macro $F1$-score of 0.81, outperforming other models.

Also, the work (Awane et al. 2021) focused on the detection of hate speech in the Arab electronic press and social networks. The authors proposed the BERT Large model in Arabic, which was pre-trained on various dialects. They used a combination of three hate speech datasets to analyze 38,654 entries made up of texts in classical Arabic, Levantine, and North African dialects. The proposed model was evaluated using precision metrics, recall, and $F1$-Score, reaching an accuracy of 83% and an $F1$-Score of 89%.

Furthermore, (Badri et al. 2022) presented a new approach for detecting inappropriate content in Arabic hate speech and abusive language by using multi-dialecticism. The authors built a large dataset called Tun-EL, which covers three Arabic dialects, and proposed a CNN-BiGRU model with fastText and AraVec word embeddings to classify the content. The experimental results showed that the deep learning model outperformed traditional machine learning models, achieving 88% classification accuracy for hateful content and 76% classification accuracy for abusive content. However, the model's performance varied depending on the dialect used. Therefore, the authors suggested enlarging the dataset and fine-tuning the hyperparameters to improve the model's accuracy.

However, authors in (Faris et al. 2020) discussed a study on detecting hate speech in Arabic using word embeddings and deep learning techniques. They highlighted the challenges of detecting hate speech in Arabic and presented a novel approach that uses pre-trained word embeddings and deep neural networks. The approach was evaluated using several deep learning models, including convolutional neural networks (CNN), recurrent neural networks (RNN), and long short-term memory (LSTM). The dataset used for the study was collected from Twitter and consisted of 3696 tweets that were manually annotated and labelled as hate, normal, and neutral. The experiments showed that the AraVec word embedding approach with the recurrent convolutional networks was competent and achieved a high accuracy and $F1$-score of 71.688% compared to existing methods, demonstrating its effectiveness in identifying hate speech in Arabic.

In a recent review article (Azzi and Zribi 2021), the author discussed the use of machine learning and deep learning techniques for detecting abusive messages in Arabic social media. The authors introduced the problem of detecting abusive messages and explained why it is important. They then provide an overview of machine learning and deep learning methods and techniques, along with their taxonomy. The authors also discussed common datasets used for training and testing models for detecting abusive messages. Finally, the paper concluded with a summary of the research and discussed future challenges in this area of research. The results suggested that deep learning models perform better than traditional machine learning models for detecting abusive messages in Arabic social media.

Also, the survey by (Husain and Uzuner 2021) provided a structured overview of previous approaches, including core techniques, tools, resources, methods, and main features used for offensive language detection in the Arabic language. The paper also discussed the limitations and gaps of the previous studies. It concluded that there is still a need for more research in this area and that there are several challenges that need to be addressed, such as data scarcity, dialectal variation, and context dependence. As for the best methods and algorithms used for offensive language detection in Arabic, the paper mentions several approaches, such as supervised learning, unsupervised learning, rule-based methods, and deep learning. However, it does not provide a definitive answer as to which method is best, as each approach has its own advantages and disadvantages depending on the specific use case.

In (Mubarak and Darwish 2019), the focus was on developing a classifier for offensive Arabic language in tweets. Offensive language has become a major concern on social media platforms such as Twitter, prompting the need for a reliable and robust classifier. The main objective of this research was to build a large word list of offensive words and create a classifier that outperformed using a word list. The authors used a seed list of offensive words to tag a large number of tweets, which enabled them to discover other offensive words by contrasting those tweets with random ones. They employed word-list, fastText, and SVM classifiers and used an existing dataset of 1100 Arabic tweets with offensive language. To train the fastText classifier, they utilized 36.6 million automatically tagged tweets and compared the fastText setup to another SVM classifier with promising results. The results of this study showed that the FastText classifier achieved a high level of precision, recall, and $F1$ of 90%.

The study (Omar et al. 2020) conducted a comprehensive comparison of traditional machine learning and deep learning algorithms for identifying Arabic hate speech on social media platforms. The authors collected a diverse dataset of 20,000 posts, tweets, and comments from multiple social network platforms and manually annotated them as hate or non-hate speech. They trained twelve machine learning algorithms and two deep learning classifiers, CNN and RNN, on the dataset to determine which approach yielded better results. The study found that the RNN model in deep learning achieved the highest accuracy score of 98.70%, while Complement NB in machine learning had the best performance, achieving an accuracy score of 97.59%. The authors concluded that deep learning algorithms are more effective in detecting Arabic hate speech in online social networks and outperform traditional machine learning approaches.

Boulouard et al. (Boulouard et al. 2022a, b) addressed the issue of hate speech in Arabic social media using machine learning techniques. The authors highlighted the

negative impact that hate speech can have on society and identified the need for effective tools to prevent and identify such speech online. They trained and evaluated several machine learning algorithms, including logistic regression, Naïve Bayes, random forests, support vector machines, and long short-term memory, on a dataset of 15,050 comments from YouTube channels known for publishing controversial videos. The authors used TF-IDF for feature extraction and found that LSTM had the best performance in terms of *F*1-Score, with SVM following closely behind. The authors conclude that machine learning algorithms show promise in detecting hate speech in Arabic social media but suggest that fine-tuning is necessary in the preprocessing step and that additional feature extraction may improve performance. Overall, this study demonstrates the potential of machine learning to combat hate speech.

The study (Albadi et al. 2019) investigated the effectiveness of combining handcrafted features and gated recurrent unit (GRU) neural networks for detecting religious hatred on Arabic Twitter. The authors emphasized the importance of addressing issues related to hate speech, specifically religious hate speech. They used an available dataset for evaluating the proposed approach, which was an automatically annotated dataset of Arabic tweets containing religious hatred. The dataset consists of 6,000 Arabic tweets collected from Twitter. They also created three public Arabic lexicons of terms related to religion along with hate scores using three well-known feature selection methods to generate these lexicons: pointwise mutual information (PMI), chi-square, and bi-normal separation (BNS). They employed three different approaches to detect religious hate speech: a lexicon-based approach, N-gram-based approach, and GRU + word embeddings. The proposed approach is a hybrid approach that combines GRU neural networks with handcrafted features to detect religious hatred in Arabic Twitter achieved superior results for detecting religious hatred in Arabic in terms of recall (0.84%).However, the authors in (El-Alami et al. 2022) proposed a multilingual offensive language detection method using transfer learning from transformer fine-tuning models like BERT, mBERT, and AraBERT to improve accuracy across different languages. The authors evaluated their model on a bilingual dataset from SOLID and compared BERT models to various neural models such as CNN, RNN, and bidirectional RNN. They conduct three experiments using joint-multilingual, joint-translated monolingual, and translation methods to evaluate the performance of different models. The results show that BERT outperforms other models in terms of accuracy and *F*1 value, where the translation-based method in conjunction with Arabic BERT (AraBERT) achieves over 93% and 91% in terms of *F*1 score and accuracy, respectively.

The study (Alsafari et al. 2020a, b) examined the detection of hate and offensive speech on Arabic social media platforms. The authors highlighted the need to detect such content, as it can have negative effects. However, the complexity of the language and lack of resources make this a unique challenge. The authors used several algorithms and methods, including SVM, naive Bayes, logistic regression, deep neural networks, and various feature extraction methods. They created an Arabic hate/offensive corpus consisting of 5340 manually annotated tweets. The results showed that SVM outperformed other models, and the CNN + mBert model performed the best across all prediction tasks. Additionally, word embedding is efficient with deep learning models and less effective with machine learning models.

Another study (Alzubi 2022) proposed an approach to detect hate speech on social media platforms, which is a critical social issue with severe consequences. The approach is specifically designed for Arabic, which has a complex structure and relies heavily on context. They used a dataset consisting of 12,698 annotated tweets and focused on offensive speech detection. The approach includes three main steps: augmentation, pre-processing, and passing data through an ensemble. The ensemble includes models such as AraBERTv0.2-Twitter-large, Mazajak Pre-trained Embeddings, Character + Word Level N-gram TF-IDF Embeddings, MUSE, and Emoji Score. The results showed that AraBERT outperformed all other models with F1-macro of 0.85%.

In (Mostafa 2022), the GOF (gradient over-fitting) team for the Arabic hate speech detection shared task at the 5th Workshop on Open-Source Arabic Corpora and Processing Tools aimed to improve the performance of imbalanced text detection models in Arabic. They used a dataset of 13,000 Arabic tweets labelled as 35% offensive and 11% hate speech. The team experimented with five different loss functions, including weighted cross-entropy, focal loss, and Tversky loss, and proposed pre-trained models such as QARiB, MARBERT, and AraBERT. The team also proposed a deep learning ensemble approach that achieved superior results with a macro F1 score of 87.044.

The authors in (De Paula 2022) discussed the approach taken by researchers for the Arabic Hate Speech 2022 Shared Task to detect offensive language and hate speech in Arabic social media comments. The team used transformer-based models such as AraBert, AraElectra, Albert-Arabic, AraGPT2, mBert, and XLM-Roberta and ensemble techniques like majority vote and highest sum to improve classification performance. They used the OSACT5 dataset, which contained around 13k tweets, with only 35% annotated as offensive and 11% as hate speech, while tweets marked as vulgar and violent only accounted for 1.5% and 0.7%, respectively. The team achieved impressive results in both offensive language and hate speech detection subtasks. The AraBert model achieved the highest *F*1-Macro scores in Tasks A and C, while the highest sum ensemble achieved the best results in Task B.

Similarly, the paper (AlKhamissi 2022) was about the authors' approach to the Arabic hate speech detection (AHSD) task, which is part of the Meta AI competition in 2022. The approach involved using multi-task learning (MTL) with a self-correction mechanism to enhance the classification of hate speech in Arabic text. The dataset used, OSACT5, consists of around 13k tweets, 35% of which are annotated as offensive and only 11% as hate speech. The proposed approach is the AraHS model, which outperformed the QARiB baseline models. MARBERTv2, pretrained with 1B multi-dialectal Arabic (DA) tweets and passed to 3 task-specific classification heads, is used as the core model. The final AraHS model is an ensemble of several trained models, each using different hyperparameters. Self-consistency correction is used to correct errors in one classification head. The results show that the AraHS model is more effective in detecting hate speech and offensive language by utilizing the self-consistency correction mechanism. The authors also conducted a detailed error analysis to identify the strengths and weaknesses of their approach and provide insights for future improvements.

The article (Mubarak 2021) analyzed the use of offensive language in Arabic tweets and evaluated machine learning models' effectiveness in identifying such language. The authors developed a method to construct an unbiased dataset and produced the most extensive Arabic dataset to date. The dataset involved 10,000 tweets manually annotated with special tags for vulgarity and hate speech. The authors employed various state-of-the-art representations and classifiers, including static and contextualized embeddings, transformer-based and SVM classifiers, and other classification techniques like AdaBoost and logistic regression. The study's results indicated that AraBERT was the most successful model, attaining an $F1$ score of 83.2.

The authors in (Haddad 2020) focused on identifying offensive language in Arabic text using deep neural networks with attention. They also utilized the OffensEval2020 dataset, which contains Arabic tweets labelled as offensive or non-offensive. They applied different methods to balance out the dataset and improve model performance. The proposed models, including CNN, Bi-GRU, CNN_ATT, and Bi-GRU_ATT, were tested alongside baseline machine learning classifiers. Their results indicated that the attention-based models performed better, with BiGRU_ATT achieving the highest F1 score of 0.859 for the offensive language detection task and 0.75 for hate speech detection.

In (Husain 2020), the SalamNET deep learning model was developed to detect offensive language in Arabic texts for SemEval-2020 Task 12. The authors tested various deep learning architectures, including a baseline LR-based model, using the Scikit-learn and Keras libraries of Python. The dataset used was the Arabic OffensEval 2020 dataset, which consisted of 10,000 tweets labelled as either offensive or not offensive. However, the dataset had a highly imbalanced distribution of offensive and non-offensive tweets, with only 1900 tweets labelled as offensive. The SalamNET Bi-directional Gated Recurrent Unit (Bi-GRU)-based model achieved a macro-$F1$ score of 0.83.

The study (Husain and Uzuner 2022a, b) examined six preprocessing techniques that impact the automatic detection of offensive Arabic language. The techniques included different forms of normalization, conversion of selected words to their hypernyms, hashtag segmentation, and cleaning. The study used various traditional and ensemble machine learning classifiers and artificial neural network classifiers. It analyzed two datasets: one that contains multiple dialects, a highly imbalanced dataset, and the other focused on the Levantine dialect. Both datasets were manually annotated. The research showed significant variations in preprocessing effects on each classifier, with AraBert achieving the best results.

The authors in (Alshalan and Al-Khalifa 2020) presented a novel approach for automatic hate speech detection in the Saudi Twittersphere using deep learning techniques. They also discussed the negative impact of hate speech in the Arab world and the challenges of detecting it due to the complexity of the Arabic language and the lack of labelled datasets. They proposed a deep learning model based on four models that were trained on a large Arabic Twitter dataset collected from Saudi Arabia. This dataset was developed using 9316 tweets classified as hateful, abusive, or normal that covered different types of hate speech to test their models. After performing several preprocessing steps and binary classification, the model's performance was evaluated using different metrics. The results showed that CNN outperformed other models, with an $F1$-score of 0.79 and an AUROC of 0.89.

Another study (Boulouard et al. 2022a, b) discussed the use of transfer learning to detect hateful and offensive speech in Arabic social media. The authors emphasized the negative consequences of hate speech on individuals and communities and compared the performance of different BERT-based models trained using a dataset of Arabic social media posts collected from YouTube. They preprocessed the dataset by removing missing values, leaving 11,268 YouTube comments with 42% hateful and 58% non-hateful comments. They use a pre-trained language model (BERT) to extract features from the text and a binary classification model to determine whether a given message is hateful or not, including features related to sentiment and emotion. The authors trained different BERT-based models, evaluated their performance using precision, recall, and F1 scores, and found BERT-EN provided accuracy of 98%.

Furthermore, the authors in (Anezi 2022) discussed the issue of hate speech on social media and the need for effective detection mechanisms. The study used a unique dataset of 4203 Arabic comments from various sources and

manually labelled them into different categories. The authors conducted experiments using a deep recurrent neural network model (DRNN-2), along with another model (DRNN-1) for binary classification. The models were evaluated using different performance metrics and compared with traditional ML classifiers. The authors found that their proposed approach provides a valuable contribution to hate speech detection research and could have potential applications in combating hate speech on social media platforms.

In another study (Guellil et al. 2020), the authors aimed to develop a supervised learning approach for detecting hate speech against politicians in Arabic social media. Two datasets, one unbalanced and the other balanced, were constructed from YouTube comments and manually annotated. The authors used various preprocessing techniques and experimented with different feature extraction methods, including bag-of-words, word embeddings, and character n-grams. The proposed approach included classical and deep learning algorithms like GNB, LR, RF, SGD Classifier, and LSVC, as well as CNN, MLP, LSTM, and Bi-LSTM. The performance of the LSVC, BiLSTM, and MLP models was the best, with an accuracy rate of up to 91% when associated with the SG model.

In (Aljuhani et al. 2022), the authors proposed a new method to detect offensive Arabic language in microblogs using deep learning and domain-specific word embeddings. They aimed to address the increasing prevalence of online hate speech on Arabic social media platforms. They built a new large multi-domain and multi-dialect Arabic dataset of offensive language, consisting of almost 30k tweets, and manually annotated it. The proposed approach uses the bidirectional long-short-term memory (BiLSTM) model and two domain-specific word embeddings (Word2Vec and FastText) to classify tweets as offensive or not. The results show that the BiLSTM model with Blend Embeddings achieved superior performance with an F1-Score of 0.93. Overall, the study demonstrates the effectiveness of using domain-specific word embeddings and deep learning for detecting offensive language in Arabic microblogs.

The article (Khezzar et al. 2023) detailed the development of arHateDetector, a web application designed to detect hate speech in both standard and dialectal Arabic tweets. The authors explained that the diversity of the Arabic language and the lack of research on hate speech in dialectal Arabic make the task challenging. To address this challenge, the authors integrated and compiled multiple online public datasets into the arHateDataset, which consists of 34,107 tweets. The system used machine learning models such as linear SVC, random forest, and logistic regression in addition to deep learning models like convolutional neural networks (CNNs) and AraBERT, achieving high accuracy in detecting hate speech in Arabic tweets. The linear SVC model achieved the highest accuracy of 89%, but the AraBERT model is the best overall, with the highest accuracy result of 93%. The authors concluded that arHateDetector can be a valuable tool for identifying and removing instances of hate speech in Arabic.

Khairy et al. (2023) conducted a study to automate the detection of offensive language or cyberbullying. They created a new Arabic offensive balanced dataset of 12,000 records using two available datasets which were collected for Facebook & Twitter. They examined the effectiveness of several single and ensemble machine learning algorithms (Linear SVC, Logistic Regression, and *K* Neighbors) and three ensemble machine learning approaches (Bagging-Random Forest, Voting, and Boosting-Adaboost). The authors found that the impact of the ensemble machine learning methodology is better than that of the single learner machine learning. They also discussed that the reliance on machine learning algorithms is one of the major weakness to detect offensive language and cyberbullying. Finally, they found that voting is the best performing trained ensemble machine learning classifier, outperforming the best single learner classifier (65.1%, 76.2%, and 98%).

The authors in (Muaad et al. 2023) proposed an Arabic hate speech detection (AHSD) model which composed of preprocessing, feature extraction, detection, and classification to identify hate speech on the Arabic benchmark dataset. They conducted various experimental setup with standalone ML, ensemble learning, and transfer learning models namely as follows: passive-aggressive (PA), logistic regression (LR), random forest (RF), decision tree (DT), K-nearest neighbors (K-NN), linear support vector classifier (Linear SVC), support vector classifier (SVC), naïve Bayes classifier (NB), Bernoulli naïve Bayes classifier (BNB), extra tree classifier (ET), ensemble bagging classifier, ensemble AdaBoost, ensemble gradient boosting classifier (GB), and Arabic bidirectional encoder representations (Arab-BiER) as a Transformer. The final results showed that proposed AraBERT model outperformed the others and got very good results in terms of accuracy, precision, recall, and f-score which were equal to 84%, 79%, 80%, and 79% for hate speech binary classification. The authors finally suggested as a future work designing a new model to cover imbalanced datasets using transfer learning techniques with Zero-shot learning and deep active learning could enhance the performance of the proposed model.

The authors of (Abbes et al. 2023) proposed a solution, employing Bi-LSTM with an attention mechanism and integrating BERT to find hateful and offensive speech on Arabic social media sites like Facebook. Extending their contribution, they introduced the multi-class Tunisian hate speech (MC_TunNS) dataset, providing a comprehensive benchmark with six labeled classes. The dataset consist of 20,000 comments sourced from Facebook. They performed their experiments by

integrate the araBERT-based contextual embedding with Bi-LSTM and attention mechanism; we also trained the araBert language model for hate speech detection. The findings indicated that AraBert exhibited superior performance, achieving the highest accuracy of 97.84%. Following closely, BI-LSTM with the attention mechanism demonstrated commendable results with an accuracy of 93.17%, showcasing enhanced proficiency in classifying Tunisian comments.

In a complementary study by (Ahmed et al. 2022), the emphasis lies on the detection of hate speech and offensive language. The authors undertake fine-tuning of Arabic pre-trained transformer models, specifically attuned to the Egyptian-Arabic dialect. Their efforts are grounded in a tailored dataset comprising 8000 text samples, meticulously labeled into five distinct classes: neutral, offensive, sexism, religious discrimination, and racism. They selected distinct Arabic pre-trained transformer models, namely bert-large-arabertv02-twitter, Bert-base-arabic-camelbert-mix, MARBERTv2, Araelectra-base-discriminator, and Bert-base-multilingual-uncased for conducting their experiments. Finally, they achieved an average accuracy of about 96% across all fine-tuned transformer models.

In a different study, an offensive speech detection model, leveraging the CAMeLBERT transformer different versions, is introduced by (Al-Dabet et al. 2023). The model's effectiveness is validated across four benchmark Arabic Twitter datasets. The combined dataset consists of 31,203 records. Notably, the proposed CAMeLBERT model, specifically the (CAMeLBERT-Mix) version, outperformed other variants and models, demonstrating superior performance through its utilization of diverse Arabic language forms with 87.15% accuracy and 83.6% *F*1 score.

Concluding our exploration of related work, (Mohamed et al. 2023) made notable experiments in tackling the class imbalance challenge within the context of hate speech and offensive language detection. Their approach involved the strategic incorporation of data augmentation techniques, leveraging oversampling methods, and introducing a focal loss function alongside traditional loss functions. They used the dataset provided by the shared task of (OSACT) in LREC 2020, which consists of 10k tweets, labelled as: hateful and non-hateful. The experiments involved the utilization of three distinct transformer models: MARBERT v2, MARBERT v1, and ARBERT. To refine the performance of these models, they implemented the QRNN deep learning architecture. In the final stage, they adopted a majority vote ensemble approach, combining the outcomes of the pre-trained models fine-tuned with QRNN. The proposed ensemble model demonstrated superior performance compared to the comparative models evaluated in this study, achieving a Macro-*F*1 score of 91.6%. This underscores the significance of incorporating diverse loss functions and oversampling techniques to enhance model performance on imbalanced datasets.

7 Challenges and future work

7.1 Challenges

Arabic offensive language and hate speech detection on social media pose significant challenges in terms of language complexity, cultural sensitivities, and limited available and well balanced resources. At the moment, identifying offensive language and hate speech on social media is a challenging task (Elzayady et al. 2022). There is a scarcity of research studies that focus solely on this aspect, and those that are available indicate the need for tailored NLP techniques. The vast majority of offensive language or hate speech detection studies have focused on English, not Arabic (Alsafari et al. 2020a, b). Additionally, Arabic's rich and complex morphology (Elzayady et al. 2023a, b), (Elzayady et al. 2022), (Duwairi et al. 2021), (Mubarak and Darwish 2019), and syntax require special preprocessing techniques and attention to common linguistic nuances that can be used to spread offensive language and hate speech. Another challenge is that Arabic includes a huge number of dialects, which may negatively affect the annotators' effectiveness, especially if they are native speakers of only one of the dialects. This drawback has been found in most previous studies, mainly when annotators were chosen via crowdsourcing (Alsafari et al. 2020a, b). Furthermore, to effectively identify hate speech in postings, it is necessary to support multi-dialect languages and use large datasets (Khezzar et al. 2023). Furthermore, certain Arabic words may undergo semantic shifts across different dialects, altering their meaning and offensiveness levels. For example, the word "شيخة" it means "**a dancer**" in the Moroccan dialect, while in other dialects like Egyptian and Gulf, it doesn't have any negative meaning, but rather indicates a high status. In addition, with multi-labelled datasets, the task becomes more challenging due to the labels' correlation (Azzi and Zribi 2022). The similarities between the different dialects mean that annotators had difficulty labelling some tweets as being in a specific dialect (Badri et al. 2022). Moreover, imbalanced datasets are a common drawback in several studies, such as (Makram 2022) and (Badri et al. 2022). For the dataset imbalance problem, the authors presented some methods for handling class imbalance using re-sampling methods, including ROS, SMOTE, and ADASYN, and different loss functions (Mansur et al. 2023), (Badri et al. 2022). While other researchers have attempted to increase the number of samples of a rare species (Husain and Uzuner 2021), another issue is that social media posts lack uniformity and grammar standards, making language models difficult to build (Berrimi et al. 2020). Many of the tweets are not in MSA; finding a good stemmer could

be challenging (Berrimi et al. 2020), and the lack of clear policies hinders automated hate speech detection (Duwairi et al. 2021). For instance, it is challenging to extract representative characteristics from tweets due to their short length and syntactic and grammatical errors. Additionally, if the dataset is limited, it is impossible to train the model using just the dataset since static word embedding may not contain all of the dataset's vocabularies (Shannaq et al. 2022). The most popular social media network for data collection is Twitter, but Facebook and YouTube are also widely used. Due to Facebook's tight data usage regulations, accessing data is more challenging. Although Twitter is a wonderful resource, details like the limit on the length of a tweet might condense information (Azzi and Zribi 2021). In (Farghaly and Shaalan 2009), the authors presented the difficulty of the Arabic script to read due to its lack of dedicated letters, changes in the form of the letter depending on its place in the word, and the absence of capitalization and punctuation. To manage this problem, NLP systems normalize the input text, but this increases the probability of ambiguity. Moreover, homographs and internal word structure ambiguities are two of the most common ambiguities. Homograph ambiguity, internal word structure ambiguity, syntactic ambiguity, constituent boundary ambiguity, anaphoric ambiguity, and features of Arabic contribute to ambiguity. Normalizing dialects and misspelt words also presents a significant challenge (Mohaouchane et al. 2019), (Alzubi 2022).

Finally, to the best of our knowledge, to overcome these challenges, the following steps can be taken:

1. Developing language models specific to Arabic, including Arabic dialects and colloquial language, to facilitate the detection of offensive Arabic language and hate speech.
2. Encouraging the use of NLP techniques and large pre-trained language models to improve the automation and efficiency of Arabic offensive language and hate speech detection.
3. Implementing a multilingual approach to Arabic offensive language and hate speech detection, including linguistic rules in different contexts.

7.2 Future trends

In recent years, there has been a growing interest in developing effective algorithms to detect offensive language and hate speech on social media platforms. While a significant amount of research has been conducted in this area, studies focusing on the Arabic language have been relatively limited. However, there are some promising developments in this emerging field. Between 2019 and April 2023, a number of studies investigating the use of machine learning and deep learning techniques to detect offensive Arabic language and hate speech are growing. These studies highlighted various future directions for the detection of abusive and hate speech in Arabic. These directions include experimenting with BERT models and its recent variants like AraBERT (Azzi and Zribi 2022), investigating the use of pre-trained Arabic embeddings (Berrimi et al. 2020), exploring semi-supervised classification techniques (Berrimi et al. 2020), detecting other forms of offensive content such as video or audio containing offensive speech (Elzayady et al. 2022), (Mohaouchane et al. 2019), assessing the effects of various contextualized word embedding techniques (e.g., BERT, GPT, GPT-2, and Elmo) on hate speech models (Elzayady et al. 2022), expanding the dataset to cover different dialects and cultures (Husain and Uzuner 2022a, b), (Omar et al. 2020), (AbdelHamid et al. 2022), (Badri et al. 2022), and using powerful GPUs for deep learning models (Al-Hassan and Al-Dossari 2021). To the best of our knowledge, it is important to investigate self-learning, zero-shot and few-shot learning using different pre-trained large language models for lablelling Arabic datasets; this is because datasets annotation is more expensive and may be biased by different annotators. The future work also emphasizes the use of active learning techniques (AbdelHamid et al. 2022), the incorporation of socio-cultural context, and building a balanced Arabic dataset (Khairy et al. 2021). Additionally, future research should focus on the continuous development and modification of machine learning and deep learning techniques for better accuracy in classifying hate speech in Arabic. Overall, as AI technology develops, there is hope that hate speech and offensive language can be detected more accurately and efficiently.

8 Conclusion

Offensive language on social media platforms is a growing concern, especially in the Arab world. However, there have been several effective solutions that have been developed to detect and remove such content. The impact of offensive language on social media platforms is significant. It can lead to cyberbullying, hate speech, and even violence in extreme cases. Therefore, it is critical that all social media platforms continue to invest in solutions that can help mitigate this problem. This review has shed light on the issue of offensive Arabic language and hate speech on social media. The aim of this study is to determine the effectiveness of current tools and methods utilized for detecting and moderating offensive language in Arabic. Moreover, the prevalence of offensive language on social media requires effective tools and methods for detecting

and moderating harmful content, particularly in Arabic, where the language is known for its complexity and high variability among its dialects. Machine learning and deep learning techniques have proven to be effective in detecting offensive language and hate speech; however, there are still challenges to overcome, such as the complexity of the Arabic language, the lack of standardization in the datasets, and cultural nuances. As such, more research is needed to develop and refine models that can accurately detect offensive language and hate speech in Arabic. This study serves as a stepping stone for researchers to conduct further investigation towards the advancement of offensive language and hate speech detection techniques in Arabic. By addressing the challenges and improving the detection methods, we can work with the community and leverage technology towards creating safer online environments for everyone, irrespective of their race, gender, religion, or nationality. Ultimately, this study can provide insights and recommendations for the development of robust and accurate tools to combat offensive language and hate speech on social media platforms.

Author contributions This Review was a collaborative effort by the authors, each of whom made substantial contributions to its conception, design, and implementation. Firstly, author "Mahmoud Abdel Samie" played a pivotal role in the acquisition of data for the study. They meticulously collected a wide range of Arabic offensive language and hate speech detection studies, which fulfilled the target of this review. Furthermore, inclusion and exclusion Criteria tasks. Moreover, He summarized the selected studies and made analysis for Arabic datasets, Techniques and challenges which will serve as a stepping stone for researchers to conduct further investigation towards the advancement of offensive language and hate speech detection techniques in Arabic. Overall, he highlighted the most significant methods, Arabic datasets, taxonomy analysis, and challenges and future trends in this field. Author "Shahira Shaaban Azab" is the corresponding author which is responsible for the overall contributions and communication related to the research. He plays a substantial role by leading the research efforts, conceptualizing the paper's structure, and drawing upon a wide range of relevant literature to provide a comprehensive overview of the topic. He also ensure that the methodologies and findings of various studies are accurately summarized and integrated. Additionally, he critically analyzed the existing research, identify gaps, and propose future research directions. Overall, His input ensures the accuracy and credibility of the paper. Lastly, Author "Hesham A. Hefny" contribution lies in the expertise and knowledge in the field which enabled us to provide valuable insights and critical analysis to ensure the paper's quality. He extensively research and study existing literature, contributing to the identification of relevant studies and concepts. His extensive background in the subject matter allows us to identify gaps in the existing research, suggest novel methodologies, and propose future research directions. Moreover, His input ensures the accuracy and credibility of the paper, as they meticulously review the content, verify the sources, and make significant revisions to improve its overall quality.

Declarations

Competing interests The authors declare no competing interests.

References

Abbes M, Kechaou Z, Alimi AM (2023) Deep learning approach for Tunisian hate speech detection on Facebook. In 2023 IEEE symposium on computers and communications (ISCC), Gammarth, Tunisia, p 739–744. https://doi.org/10.1109/ISCC58397.2023.10217909

AbdelHamid M, Jafar A, Rahal Y (2022) Levantine hate speech detection in twitter. Soc Netw Anal Min. https://doi.org/10.1007/s13278-022-00950-4

Abuzayed A (2020) Quick and simple approach for detecting hate speech in Arabic tweets. In: ACL anthology. https://aclanthology.org/2020.osact-1.18/

Ahmed I, Abbas M, Hatem R, Ihab A, Fahkr MW (2022) Fine-tuning Arabic pre-trained transformer models for Egyptian-Arabic dialect offensive language and hate speech detection and classification. In: 2022 20th international conference on language engineering (ESOLEC). Cairo, Egypt, p. 170–174. https://doi.org/10.1109/ESOLEC54569.2022.10009167

Alakrot A, Murray L, Nikolov NS (2018a) Dataset construction for the detection of anti-social behaviour in online communication in Arabic. Procedia Comput Sci 142:174–181. https://doi.org/10.1016/j.procs.2018.10.473

Alakrot A, Murray L, Nikolov NS (2018) Towards accurate detection of offensive language in online communication in Arabic. In 4th international conference on Arabic computational linguistics (ACLING), p 315–320. https://doi.org/10.1016/j.procs.2018.10.491

Alakrot A, Fraifer M, Nikolov NS (2021) Machine learning approach to detection of offensive language in online communication in Arabic. In: 2021 IEEE 1st international Maghreb meeting of the Conference on sciences and techniques of Automatic Control and Computer Engineering MI-STA. https://doi.org/10.1109/mi-sta52233.2021.9464402

Albadi N, Kurdi M, Mishra S (2019) Investigating the effect of combining GRU neural networks with handcrafted features for religious hatred detection on Arabic Twitter space. Soc Netw Anal Min. https://doi.org/10.1007/s13278-019-0587-5

Albadi N, Kurdi M, Mishra S (2018) 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). https://doi.org/10.1109/asonam.2018.8508247

ALBayari R, Abdullah S, Salloum SA (2021) Cyberbullying classification methods for Arabic: a systematic review. In: Proceedings of the international conference on artificial intelligence and computer vision (AICV2021), p 375–385. https://doi.org/10.1007/978-3-030-76346-6_35

Al-Dabet S, ElMassry A, Alomar B, Alshamsi A (2023) Transformer-based Arabic offensive speech detection. In: 2023 international conference on emerging smart computing and informatics (ESCI), Pune, India, p. 1–6. https://doi.org/10.1109/ESCI56872.2023.10100134

AlFarah ME, Kamel I, Al Aghbari Z, Mouheb D (2022) Arabic cyberbullying detection from imbalanced dataset using machine learning. In: Soft computing and its engineering applications, p 397–409. https://doi.org/10.1007/978-3-031-05767-0_31

Al-Hassan A, Al-Dossari H (2021) Detection of hate speech in Arabic tweets using deep learning. Multimed Syst 28:1963–1974. https://doi.org/10.1007/s00530-020-00742-w

Alhejaili R, Alsaeedi A, Yafooz WMS (2022) Detecting hate speech in Arabic tweets during COVID-19 Using machine learning approaches. In: Proceedings of third doctoral symposium on computational intelligence, p 467–475. https://doi.org/10.1007/978-981-19-3148-2_39

Aljuhani KO, Alyoubi KH, Alotaibi FS (2022) Detecting Arabic offensive language in microblogs using domain-specific word embeddings and deep learning. Tehnički Glasnik 16:394–400. https://doi.org/10.31803/tg-20220305120018

AlKhamissi B (2022) Meta AI at Arabic hate speech 2022: MultiTask learning with self-correction for hate speech classification. In: arXiv.org. https://doi.org/10.48550/arXiv.2205.07960

Alotaibi A, Abul Hasanat MH (2020) Racism detection in Twitter Using deep learning and text mining techniques for the Arabic language. In: 2020 first international conference of smart systems and emerging technologies (SMARTTECH). https://doi.org/10.1109/smart-tech49988.2020.00047

Alsafari S, Sadaoui S (2021b) Semi-supervised self-training of hate and offensive speech from social media. Appl Artif Intell 35:1621–1645. https://doi.org/10.1080/08839514.2021.1988443

Alsafari S, Sadaoui S, Mouhoub M (2020b) Hate and offensive speech detection on Arabic social media. Online Soc Netw Media 19:100096. https://doi.org/10.1016/j.osnem.2020.100096

Alsafari S, Sadaoui S (2021) Semi-supervised self-learning for Arabic hate speech detection. In: 2021 IEEE international conference on systems, man, and cybernetics (SMC). https://doi.org/10.1109/smc52423.2021.9659134

Alsafari S, Sadaoui S, Mouhoub M (2020) Deep learning ensembles for hate speech detection. In: 2020 IEEE 32nd international conference on tools with artificial intelligence (ICTAI). https://doi.org/10.1109/ictai50040.2020.00087

Alshalan R, Al-Khalifa H (2020) A deep learning approach for automatic hate speech detection in the Saudi Twittersphere. Appl Sci 10:8614. https://doi.org/10.3390/app10238614

Althobaiti MJ (2022) BERT-based approach to Arabic hate speech and offensive language detection in Twitter: exploiting emojis and sentiment analysis. Int J Adv Comp Sci Appl. https://doi.org/10.14569/ijacsa.2022.01305109

Alzubi S (2022) aiXplain at Arabic hate speech 2022: an ensemble based approach to detecting offensive tweets. In: ACL Anthology. https://aclanthology.org/2022.osact-1.28/

Anezi FYA (2022) Arabic hate speech detection using deep recurrent neural networks. Appl Sci 12:6010. https://doi.org/10.3390/app12126010

Awane W, Ben Lahmar EH, El Falaki A (2021) Hate speech in the Arab electronic press and social networks. Revue D'intelligence Artificielle 35:457–465. https://doi.org/10.18280/ria.350603

Azzi S, Zribi C (2022) Comparing deep learning models for multi-label classification of Arabic abusive texts in social media. In: Proceedings of the 17th international conference on software technologies. https://doi.org/10.5220/0011141700003266

Azzi SA, Zribi CBO (2021) From machine learning to deep learning for detecting abusive messages in Arabic social media: survey and challenges. Adv Intell Syst Comput. https://doi.org/10.1007/978-3-030-71187-0_38

Badri N, Kboubi F, Habacha Chaibi A (2022) Towards automatic detection of inappropriate content in multi-dialectic Arabic text. Adv Comput Collect Intell. https://doi.org/10.1007/978-3-031-16210-7_7

Berrimi M, Moussaoui A, Oussalah M, Saidi M (2020) Attention-based networks for analyzing inappropriate speech in Arabic text. In: 2020 4th international symposium on informatics and its applications (ISIA). https://doi.org/10.1109/isia51297.2020.9416539

Boulouard Z, Ouaissa M, Ouaissa M (2022a) machine learning for hate speech detection in Arabic social media. Eai/springer Innov Commun Comput. https://doi.org/10.1007/978-3-030-77185-0_10

Boulouard Z, Ouaissa M, Ouaissa M et al (2022b) Detecting hateful and offensive speech in Arabic social media using transfer learning. Appl Sci 12:12823. https://doi.org/10.3390/app122412823

Duwairi R, Hayajneh A, Quwaider M (2021) A deep learning framework for automatic detection of hate speech embedded in Arabic tweets. Arab J Sci Eng 46:4001–4014. https://doi.org/10.1007/s13369-021-05383-3

El-Alami F, Ouatik El Alaoui S, En Nahnahi N (2022) A multilingual offensive language detection method based on transfer learning from transformer fine-tuning model. J King Saud Univ Comput Inf Sci 34:6048–6056. https://doi.org/10.1016/j.jksuci.2021.07.013

Elzayady H, Mohamed MS, Badran K et al (2023) Arabic hate speech identification by enriching MARBERT model with hybrid features. Intell Sustain Syst. https://doi.org/10.1007/978-981-19-7663-6_53

ElZayady H, Mohamed MS, Badran K, Salama G (2023) A hybrid approach based on personality traits for hate speech detection in Arabic social media. Int J Electr Comput Eng 13:1979. https://doi.org/10.11591/ijece.v13i2.pp1979-1988

Elzayady H, Mohamed MS, Badran K, Salama G (2022) Improving Arabic hate speech identification using online machine learning and deep learning models. In: Proceedings of seventh international congress on information and communication technology, p 533–541. https://doi.org/10.1007/978-981-19-1610-6_46

Farghaly A, Shaalan K (2009) Arabic natural language processing. ACM Trans Asian Lang Inf Process 8:1–22. https://doi.org/10.1145/1644879.1644881

Faris H, Aljarah I, Habib M, Castillo P (2020) Hate speech detection using word embedding and deep learning in the Arabic language context. In: Proceedings of the 9th international conference on pattern recognition applications and methods. https://doi.org/10.5220/0008954004530460

Guellil I, Adeel A, Azouaou F et al (2020) Detecting hate speech against politicians in Arabic community on social media. Int J Web Inf Syst 16:295–313. https://doi.org/10.1108/ijwis-08-2019-0036

Haddad H, Mulki H, Oueslati A (2019) T-HSAB: a Tunisian hate speech and abusive dataset. Commun Comput Inf Sci. https://doi.org/10.1007/978-3-030-32959-4_18

Haddad B (2020) Arabic offensive language detection with attention-based deep neural networks. In: ACL Anthology. https://aclanthology.org/2020.osact-1.12/

Husain F, Uzuner O (2022) Transfer learning across Arabic dialects for offensive language detection. In: 2022 international conference on Asian language processing (IALP). https://doi.org/10.1109/ialp57159.2022.9961263

Husain F, Uzuner O (2021) A survey of offensive language detection for the Arabic language. ACM Trans Asian Low-Resour Lang Inf Process 20:1–44. https://doi.org/10.1145/3421504

Husain F, Uzuner O (2022a) Investigating the Effect of preprocessing Arabic text on offensive language and hate speech detection. ACM Trans Asian Low-Resour Lang Inf Process 21:1–20. https://doi.org/10.1145/3501398

Husain F (2020) SalamNET at SemEval-2020 Task12: deep learning approach for Arabic offensive language detection. In: arXiv.org. https://doi.org/10.48550/arXiv.2007.13974

Khairy M, Mahmoud TM, Abd-El-Hafeez T (2021) Automatic detection of cyberbullying and abusive language in Arabic content on social networks: a survey. Procedia Comput Sci 189:156–166. https://doi.org/10.1016/j.procs.2021.05.080

Khairy M, Mahmoud TM, Omar A, Abd El-Hafeez T (2023) Comparative performance of ensemble machine learning for Arabic cyberbullying and offensive language detection. Lang Resour Eval. https://doi.org/10.1007/s10579-023-09683-y

Khezzar R, Moursi A, Al Aghbari Z (2023) arHateDetector: detection of hate speech from standard and dialectal Arabic tweets. Discov Int Things. https://doi.org/10.1007/s43926-023-00030-9

Makram K (2022) CHILLAX—at Arabic hate speech 2022: a hybrid machine learning and transformers based model to detect Arabic offensive and hate speech. In: ACL Anthology. https://aclanthology.org/2022.osact-1.25/

Mansur Z, Omar N, Tiun S (2023) Twitter Hate Speech Detection: A Systematic Review of Methods, Taxonomy Analysis, Challenges, and Opportunities. IEEE Access 11:16226–16249. https://doi.org/10.1109/access.2023.3239375

Mohamed MS et al (2023) An efficient approach for data-imbalanced hate speech detection in Arabic social media. Int J Electr Comput Eng 13:6381–6390. https://doi.org/10.3233/JIFS-231151

Mohaouchane H, Mourhir A, Nikolov NS (2019) Detecting offensive language on Arabic social media using deep learning. In: 2019 sixth international conference on social networks analysis, management and security (SNAMS). https://doi.org/10.1109/snams.2019.8931839

Mostafa A (2022) GOF at Arabic hate speech 2022: breaking the loss function convention for data-imbalanced Arabic offensive text detection. In: ACL Anthology. https://aclanthology.org/2022.osact-1.21/

Muaad AY, Hanumanthappa J, Prakash SPS et al (2023) Arabic hate speech detection using different machine learning approach. Adv Intell Comput Data Sci. https://doi.org/10.1007/978-3-031-36258-3_38

Mubarak H, Darwish K (2019) Arabic offensive language classification on Twitter. Lect Notes Comput Sci. https://doi.org/10.1007/978-3-030-34971-4_18

Mubarak H, Darwish K, Magdy W (2017) Abusive language detection on Arabic social media. In: Proceedings of the first workshop on abusive language online. https://doi.org/10.18653/v1/w17-3008

Mubarak H (2020) Overview of OSACT4 Arabic offensive language detection shared task. In: ACL Anthology. https://aclanthology.org/2020.osact-1.7/

Mubarak H (2021) Arabic offensive language on Twitter: analysis and experiments. In: ACL Anthology. https://aclanthology.org/2021.wanlp-1.13/

Mulki H, Haddad H, Bechikh Ali C, Alshabani H (2019) L-HSAB: a Levantine Twitter Dataset for hate speech and abusive language. In: Proceedings of the third workshop on abusive language online. https://doi.org/10.18653/v1/w19-3512

Omar A, Mahmoud TM, Abd-El-Hafeez T (2020) Comparative performance of machine learning and deep learning algorithms for Arabic hate speech detection in OSNs. Adv Intell Syst Comput. https://doi.org/10.1007/978-3-030-44289-7_24

De Paula AFM (2022) UPV at the Arabic hate speech 2022 shared task: offensive language and hate speech detection using transformers and ensemble models. In: ACL Anthology. https://aclanthology.org/2022.osact-1.23/

Rahma A, Azab SS, Mohammed A (2023) A comprehensive survey on Arabic Sarcasm detection: approaches, challenges and future trends. IEEE Access 11:18261–18280. https://doi.org/10.1109/access.2023.3247427

Ruwandika NDT, Weerasinghe AR (2018) Identification of hate speech in social media. In: 2018 18th international conference on advances in ICT for emerging regions (ICTer). https://doi.org/10.1109/icter.2018.8615517

Shannag F, Hammo BH, Faris H (2022) The design, construction and evaluation of annotated Arabic cyberbullying corpus. Educ Inf Technol 27:10977–11023. https://doi.org/10.1007/s10639-022-11056-x

Shannaq F, Hammo B, Faris H, Castillo-Valdivieso PA (2022) Offensive language detection in Arabic social networks using evolutionary-based classifiers learned from fine-tuned embeddings. IEEE Access 10:75018–75039. https://doi.org/10.1109/access.2022.3190960

Waseem Z, Chung WHK (2017) Proceedings of the first workshop on abusive language online. Association for Computational Linguistics, Vancouver, BC