computers

MDPI

Article

SOD: A Corpus for Saudi Offensive Language Detection Classification

Afefa Asiri * and Mostafa Saleh *

Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia
* Correspondence: aasiri0410@stu.kau.edu.sa (A.A.); msherbini@kau.edu.sa (M.S.)

Abstract: Social media platforms like X (formerly known as Twitter) are integral to modern communication, enabling the sharing of news, emotions, and ideas. However, they also facilitate the spread of harmful content, and manual moderation of these platforms is impractical. Automated moderation tools, predominantly developed for English, are insufficient for addressing online offensive language in Arabic, a language rich in dialects and informally used on social media. This gap underscores the need for dedicated, dialect-specific resources. This study introduces the Saudi Offensive Dialectal dataset (SOD), consisting of over 24,000 tweets annotated across three levels: offensive or non-offensive, with offensive tweets further categorized as general insults, hate speech, or sarcasm. A deeper analysis of hate speech identifies subtypes related to sports, religion, politics, race, and violence. A comprehensive descriptive analysis of the SOD is also provided to offer deeper insights into its composition. Using machine learning, traditional deep learning, and transformer-based deep learning models, particularly AraBERT, our research achieves a significant F1-Score of 87% in identifying offensive language. This score improves to 91% with data augmentation techniques addressing dataset imbalances. These results, which surpass many existing studies, demonstrate that a specialized dialectal dataset enhances detection efficacy compared to mixed-language datasets.

Keywords: natural language processing (NLP); Saudi dialect; offensive detection; Arabic language processing; machine learning; deep learning; computational linguistics; dialect analysis; hate speech detection; text classification; data annotation; data augmentation

Citation: Asiri, A.; Saleh, M. SOD: A Corpus for Saudi Offensive Language Detection Classification. *Computers* **2024**, *13*, 211. https://doi.org/10.3390/computers13080211

Academic Editor: Ming Liu

Received: 9 July 2024
Revised: 13 August 2024
Accepted: 19 August 2024
Published: 20 August 2024

1. Introduction

Social media platforms, such as Facebook and X (formerly Twitter), have revolutionized communication, enabling people to connect, share views, news, and ideas on an unprecedented scale. However, this open and free-flowing environment has also become a conduit for the spread of offensive language, defined as "hurtful, derogatory, or obscene comments made by one person to another" [1].

Despite the existence of laws and policies aimed at curbing offensive language, the need for automated detection systems has become increasingly apparent. Most social media platforms now require automated methods to identify and mitigate harmful content, driving significant research interest in this area. Initially, many studies relied on machine learning techniques, employing basic textual features like bags of words and n-grams, which proved effective in identifying offensive language [2–7]. More recently, there has been a shift towards deep learning techniques, which have demonstrated superior performance in detecting offensive language [8–10]. However, while systems for detecting offensive language in English are well developed, research focused on the Arabic language remains limited [11].

Arabic, the fastest-growing language on the internet [12], is known for its morphological richness, where a single root word can generate hundreds of variations. Arabic is generally divided into Standard Arabic (SA) and Dialectal Arabic (DA). SA includes both Classical Arabic (CA) and Modern Standard Arabic (MSA), representing the formal

language, while DA refers to informal speech. The main Arabic dialects include Egyptian, Moroccan, Levantine, Iraqi, Gulf, and Yemeni. The Gulf region, including Saudi Arabia, Kuwait, Bahrain, Qatar, and Oman, shares cultural and linguistic similarities, particularly with the eastern region of Saudi Arabia, though variations exist across other regions [13]. Social media content predominantly features these dialects [14].

X (formerly Twitter) has become a key platform for information sharing and gathering opinionated content on topics such as politics, business, and social issues. As shown in Figure 1, Saudi Arabia ranks as the 9th most active country on Twitter globally and the top Arabic-speaking nation, with 15.5 million users [15]. This prominence underscores the need for a Saudi dialect dataset specifically for offensive language detection, as effective natural language processing (NLP) studies depend on access to appropriate corpora.

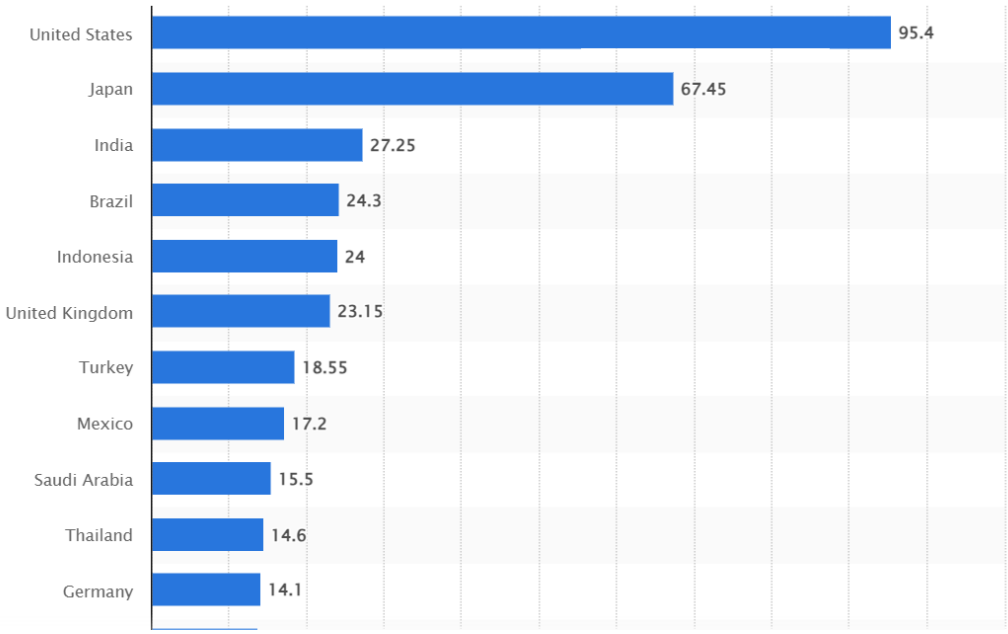

Figure 1. Leading Countries based on number of X, formerly Twitter, users: January 2023, in millions [15].

Most Arabic natural language processing (NLP) tools are tailored for MSA and struggle with DA, as highlighted by Farghaly and Shaalan, who note the impracticality of a singular NLP tool being capable of processing all Arabic variants [16,17]. Consequently, Arabic NLP solutions must designate which variant they are equipped to handle. Notably (see Figure 2), it has been observed that from the year 2017 onwards, there has been a greater inclusion of dialectal Arabic in language corpora compared to MSA.

In this paper, we present the Saudi Offensive Dialect dataset (SOD), consisting of over 24,000 tweets, annotated using a three-tier hierarchical approach. The tweets are first categorized as either offensive or non-offensive. Offensive tweets are then further classified into three categories: general insults, hate speech, and sarcasm. Finally, hate speech tweets are subdivided into three classes: sport, religious–political–racial, and insult–violence. We conducted an in-depth analysis of the SOD dataset, including exploratory data analysis and token analysis, to uncover additional insights. A series of machine learning and deep learning experiments are performed to evaluate the effectiveness of these techniques for offensive language detection. Additionally, we develop and test various data augmentation models to address the challenge of an imbalanced dataset.

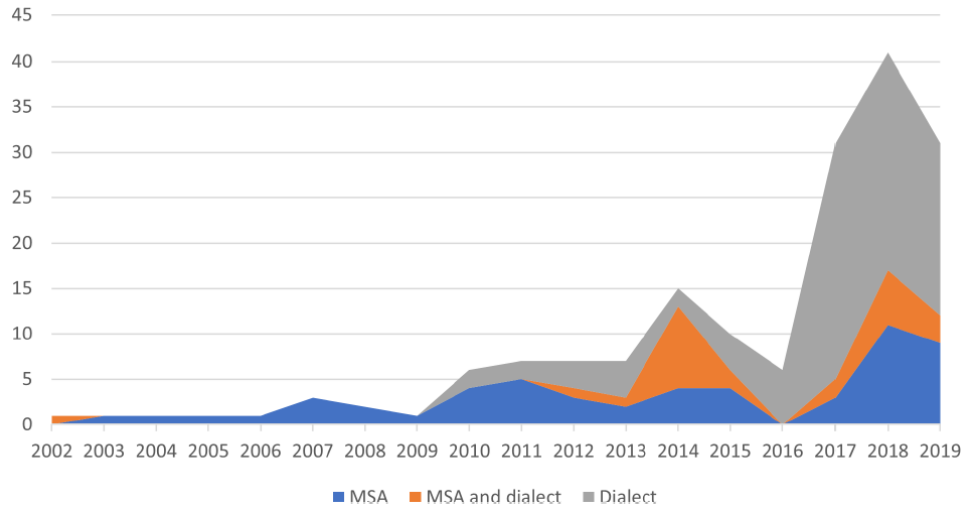

Figure 2. Percentage of Arabic corpora based on the type of corpus, from 2002 to 2019 [18].

The key contributions of this paper are:

- The development of a comprehensive offensive language corpus comprising over 24,000 tweets, representing Saudi dialects from all regions.
- The implementation of a hierarchical annotation system for offensive language detection, enabling both broad classification and deeper, more nuanced categorization.
- An extensive analysis of the linguistic aspects of the Saudi dialect, enhancing understanding of its unique features.
- The evaluation of various NLP tools, including machine learning, traditional deep learning, and transformer-based models, for detecting offensive language.
- The implementation of data augmentation techniques to address the issue of dataset imbalance.

The remainder of the paper is organized as follows: Section 2 reviews related work on Arabic and Saudi data collection, particularly for offensive language detection. Section 3 describes the corpus construction process. Section 4 presents an in-depth analysis of the SOD dataset. Section 5 illustrates the experiments and discusses the results. Section 6 presents the results after implementing data augmentation techniques. Finally, Section 7 concludes the work and discusses future research directions.

2. Literature Review

The proliferation of social media platforms has generated vast amounts of textual data in various languages and dialects, presenting both opportunities and challenges for natural language processing (NLP). In the context of Arabic, known for its rich morphological features and diverse dialects, the need for dialect-specific datasets is crucial for advancing NLP research and applications. This literature review focuses on the efforts to collect and annotate datasets specific to the Saudi dialect, highlighting the contributions of key studies in this area.

Azmi and Alzanin [19] were pioneers in examining the polarity of Saudi public opinion through e-newspaper comments, collecting 815 comments for sentiment classification. Their work underscored the early recognition of the value of analyzing Saudi dialects to gain sentiment insights. Building on this, Al-Harbi and Emam [20] expanded the corpus to 5500 tweets, aiming to refine Arabic sentiment analysis through dialect preprocessing, marking a significant step towards understanding the nuances of Saudi dialects in sentiment analysis.

The efforts by Al-Twairesh et al. [21] to compile over 17,000 tweets for sentiment analysis further exemplified the growing interest in Saudi dialects. Their work not only provided a larger dataset but also highlighted the complexity and richness of the Saudi dialect in expressing sentiments. Similarly, Al-Thubaity et al. [22] contributed by collecting

5400 tweets, enriching resources for sentiment and emotion classification in the Saudi dialect, thereby offering new dimensions for analysis.

Continuing this trend, Alqarafi et al. [23] and Alruily [24] built upon the foundation laid by their predecessors. Alqarafi et al. collected 4000 tweets for sentiment classification, while Alruily amassed 207,452 tweets for linguistic analysis. These contributions emphasized the growing interest and the critical need for comprehensive datasets that capture the linguistic diversity within the Saudi dialect.

Alshalan and Al-Khalifa [8] shifted their focus to hate speech detection, collecting 9316 tweets. This study marked a move towards addressing more specific and socially impactful aspects of language use on social media, reflecting the evolving objectives of dialect-specific dataset collection efforts. Bayazed et al. [25] further advanced this field by classifying 4180 tweets according to dialects and sentiments, demonstrating the importance of dialect-specific approaches in improving the effectiveness of NLP applications.

Finally, Almuqren and Cristea [18] contributed a dataset of 20,000 telecom-related tweets for sentiment classification, highlighting the practical applications of NLP in industry-specific contexts. Their work showcased the versatility and importance of dialect-specific datasets for developing tailored NLP solutions.

As shown in Table 1, the Saudi dialect dataset collection has predominantly focused on sentiment analysis, achieving significant insights into the interaction between language and emotion. While datasets for offensive language detection exist within the broader Arabic context [4,7,10,26–30], there remains a distinct gap for such datasets specifically tailored to the Saudi dialect. Addressing this gap is essential for increasing the inclusivity and safety of digital communication platforms, thereby enriching NLP research and applications across the Arabic linguistic spectrum and ensuring thorough representation of the various Arabic dialects.

Table 1. Summary of Saudi dialect dataset studies.

Ref.	Author/Year	Dataset Size	Main Task/Purpose	Label
[19]	Azmi and Alzanin (2014)	815 comments from two Saudi newspapers	Sentiment classification	Strongly positive, positive, negative, or strongly negative
[20]	Al-Harbi and Emam (2015)	5500 tweets	Sentiment classification	Positive, negative, or neutral
[21]	Al-Twairesh et al. (2017)	17,000+ tweets	Sentiment classification	Positive, Negative, Neutral
[22]	Al-Thubaity et al. (2018)	5400 tweets	Sentiment classification;	Positive, negative, neutral, objective, spam, or not sure;
			Emotion classification	anger, fear, disgust, sadness, happiness, surprise, no emotion, and not sure
[23]	Alqarafi et al. (2018)	4000 tweets	Sentiment classification	Positive or negative
[24]	Alruily (2020)	207,452 tweets	Linguistic analysis	Various linguistic features
[8]	Alshalan and Al-Khalifa (2020)	9316 tweets	Hate speech detection	Normal, abusive, hateful
[25]	Bayazed et al. (2020)	4180 tweets	Dialect classification; Sentiment classification	Hijazi, Najdi, and eastern; positive, negative, or neutral
[18]	Almuqren and Cristea (2021)	20,000 tweets, telecom-related tweets	Sentiment classification	Negative, positive

3. Corpus Generation

This section details our strategy for building the Saudi Offensive Language Dataset (SOD) using the snscrape Python package to collect data from the X platform (formerly Twitter). snscrape is a versatile tool for scraping social networking services, capable of

collecting various data types, including user profiles, hashtags, and searches. Our data-scraping efforts on the X platform considered the following:

- **Regions and Locations:** Data was collected from all 13 regions of Saudi Arabia, with search radii adjusted between 50 km and 300 km around major cities to capture relevant locations, while excluding extraneous areas.
- **Specific Timeframes:** Data was gathered over the past four years (2019–2022) to ensure comprehensive coverage and avoid biases toward specific periods or topics, such as COVID-19 or the World Cup 2022.
- **Query-Driven Approach:** In addition to predetermined geographical boundaries and timeframes, we followed a four-step approach:
  - *General Collection:* The initial phase involved scraping data without specific filters or predefined seeds.
  - *By Emoji*: Data was collected based on specific emojis that could indicate potentially offensive language.
  - *By Keyword:* Tweets containing keywords indicative of the Saudi dialect and potential offensive language were targeted.
  - *By Hashtag:* We selected specific hashtags popular within the Saudi Twitter community, indicative of broader conversations, for data extraction.

Table 2 details the specific query-driven criteria used during data collection. These queries were carefully selected to represent various types of offensive language, each with its own context and implications:

- **Emojis:** These visual symbols are among the most common in offensive Arabic tweets. The list provided in the table is sourced from existing literature [31]. Notably, while these emojis are representative of broader Arabic culture, they may not specifically reflect the nuances of the Saudi dialect. For example, the emoji " 🙏 ", ranked as the most used emoji for offensive language, is generally understood as raindrops in Saudi culture and is often paired with prayers and supplications.
- **Keywords:** These words and phrases encompass a wide range of topics, many of which relate to fine-grained hate classes identified in prior research [32]. Their selection includes general offensive remarks, racial or nationality-based slurs, sports-related ideologies, social class descriptors, and gender-based terms. For instance, terms like "يلعن" (Yilan, "Curse") and "كلب" (Kalb, "Dog") are categorized as general offenses, while "يمني" (Yamani, "Yemeni") and "مصري" (Masri, "Egyptian") relate to race or nationality.
- **Hashtags:** The hashtags in our collection touch upon themes similar to those in the keywords, including general offenses, race or nationality, sports ideologies, social classes, and gender topics.

For a visual representation of the data collection process, refer to the flowchart in Figure 3.

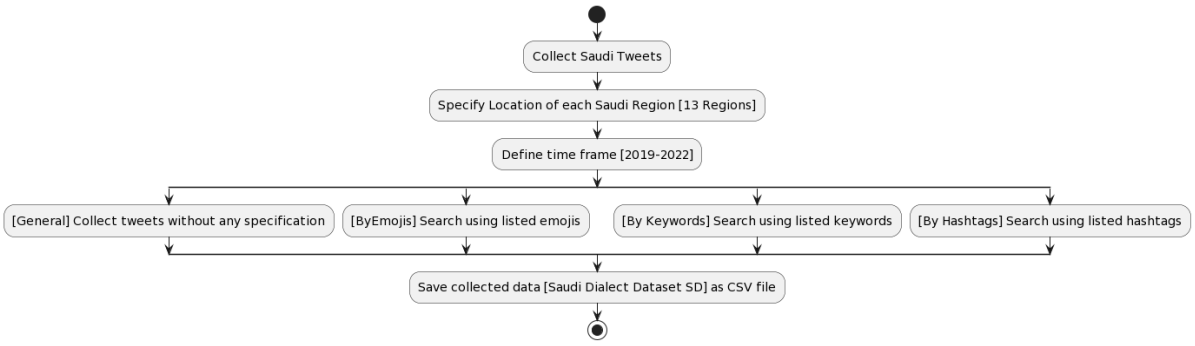

Figure 3. Workflow of the data collection process.

To provide a clear understanding of our collection framework, we present Table 3, which offers examples from each data collection category. It includes two examples for each category—one offensive and one non-offensive.

Table 2. Query-driven criteria: emojis, keywords, and hashtags.

Query-Driven	List	Translation
Emojis	🤧, 🐃, 🐷, 🫣, 🥿, 🐅, 🐶, 🐽, 🐄, 🐯, 🐑, 🐏, 👎, 😒, 🤬, 🤮, 😾, 😡	Curse, Rudeness, Dirty, Rude, Dog, Donkey, Pig, Trash, Spit on you, Despicable, Lowly, You animal, Filthy, You Yemeni, You Egyptian, You Indian, You Bengali, Cap, [Direct names for Football Clubs and Regional Terms]
Keywords	خسيس, تفو عليك, زبالة, خنزير, حمار, كلب, وقح, وسخ, قلة أدب, يلعن, طاقية, يابنقالي, ياهندي, يامصري, يابماني, يايمني, نجس, ياحيوان, نذل, نصر, سحالي, تماسيح, طحالب, نصراوي, اتحاد, هلال, أهلي, دوري يلو, ياحضر, يابدو, ياحضري, يابدوي, شبابي, أهلاوي, هلال, هلالي, اتحادي, نجدي, حجازي, أعرابي, بقايا حجاج, طرش بحر, ياخضيري, ياقروي, ياسني, ياشيعي, ذباب الكتروني, يابيض, زيود, زيدي, حساوي, قصيمي, شيعي, سني, إرهابي, داعشي, ياإرهابي, ياوطنجي, ياأخواني, ياداعشي, رجل, الرجال, البنات, يابنت, ياحريم, حرمة, ليبرالي, أخواني, إيراني, الأولاد, أولاد اليوم, بنات اليوم, الرياجيل	#AlShabab, #AlNassr, #AlHilal, #AlAhli, #AlIttihad, #SaudiLeague, #Crocodiles, #Algae, #Tigers, #Urban, #Bedouin, #Feminism, #Feminist, #Masculinity, #Masculine, #Independent, #Free
Hashtags	#الشباب, #النصر, #الهلال, #الأهلي, #الاتحاد, #السعودي_الدوري, #نسوي, #نسوية, #بدو, #حضر, #النمور, #الطحالب, #التماسيح, #ذكوري, #ذكورية, #مستقلة, #حرة	Curse, Rudeness, Dirty, Rude, Dog, Donkey, Pig, Trash, Spit on you, Despicable, Lowly, You animal, Filthy, You Yemeni, You Egyptian, You Indian, You Bengali, Cap, [Direct names for Football Clubs & Regional Terms]

Table 3. Exemplary data from collection categories.

Tweet	Location	Search by
@User والله من ضعافة النفس وقلة المروءه (Translation: Truly, this is due to a weak spirit and lack of chivalry.)	Dammam	General
@User مغربية خاشعة كالندى تُرطب الروح تبارك الله ✨♡، (Translation: A sunset, humble like the dew, refreshing the soul. God bless ✨♡,)	Jeddah	General
😡اقسم بالله لو يخلو الحمير يسوقو حيسوقو احسن من البهايم الي هنا (Translation: I swear if they let donkeys drive, they would drive better than the idiots here.😡)	Medina	Emojis
🐅😨يارب يخلص هالشهر بسرعه و اروح اشوف لوسي (Translation: Oh God, I hope this month ends quickly so I can go see Lucy 😨🐅.)	Riyadh	Emojis
مندسين مافيه عاشق هلالي الا قال معوضين خير ومبروك للوحده جماهير .هذول .اخي بدر هؤلاء ليسوا هلاليين الزعيم الصادقه حملت الاعبين المسؤليه الامور الاداريه والفني في منتهى الروعه لاتلتفتوا لم يريد الايقاع وشق الصف اتركوهم ولاتصغون لثرثرتهم (Translation: Brother Badr, these are not Hilal supporters. These are infiltrators. No true Hilal fan would say anything but 'we will be compensated for the best' and 'congratulations to Al-Wehda'. The true fans of the leader held the players responsible. Administrative and technical matters are superb. Do not pay attention to those who want to cause division and break ranks. Leave them and do not listen to their chatter.)	Najran	Keywords

Table 3. *Cont.*

Tweet	Location	Search by
@User @User اكرم الحمار منه ع الاقل الحمار يستفاد منه هذا خنزير الله يكرمك احد يعرف رقم حسابه البنكي 😂😂😂 (Translation: It's better to honor the donkey than him, at least the donkey is beneficial. This one's a pig, God bless you. Does anyone know his bank account number? 😂😂😂)	Tabuk	Keywords
@User الشباب#التعاون#الاتحاد#الأهلي#التصر#السعودي_المنتخب# !!الهلال مدللين يلتحقون على كيفهم#لاعبي (Translation: The #Alhilal players are pampered and join as they please!! #Saudi_national_team #Alahli #Alittihad #Altaawon #Alshabab)	Buraidah	Hashtag
الهلال#لا زلت أقول ، الحمد لله على نعمة 🤍 (Translation: I still say, thank God for the blessing of #Alhilal 🤍.)	Albaha	Hashtag

3.1. Data Annotation

From our data collection process, 28,000 tweets were selected for annotation. The annotation task for this dataset was structured hierarchically, spanning three primary levels, as shown in Figure 4:

- **Level 1 [Offensive vs. Non-offensive]:** At this foundational level, each tweet was assessed for its overall tone to determine whether it was offensive or not.
- **Level 2 [Offensive Tweets]:** Focus was placed only on tweets identified as offensive in Level 1, which were then annotated into:
  - *General Insult:* Speech that is simply offensive but poses no risk to others, generally NOT considered a human rights violation [33].
  - *Sarcasm:* The use of remarks that clearly mean the opposite of what they say, made to hurt someone's feelings or to criticize something in a humorous way [4].
  - *Hate Speech:* Becomes a human rights violation if it incites discrimination, hostility, or violence towards a person or a group defined by their race, religion, ethnicity, or other factors [33].
- **Level 3 [Hate Speech Tweets]:** Focus was only on tweets identified as hate speech in Level 2. It was then classified into more specific types of hate speech: racial, gender-based, sports-related, political/religious, vulgar, violence-related, and others.

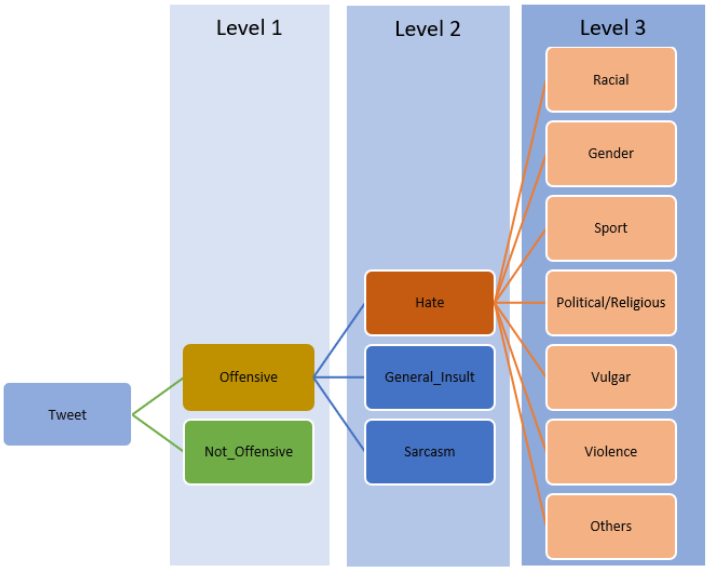

Figure 4. Hierarchical annotation structure.

3.1.1. Annotation Tools

We utilized Dagshub platform (https://dagshub.com), an innovative platform designed for data science tasks. It facilitates collaborative annotation with features such as real-time synchronization and an intuitive interface, streamlining the overall annotation process.

3.1.2. Annotators Guideline

Multiple Saudi annotators participated in this project. To ensure consistency and address any ambiguities, an initial briefing session was conducted. During this session, annotators were introduced to the annotation guidelines as outlined in Table 4, with illustrative examples provided to clarify potential uncertainties. Additionally, a live annotation session was conducted and monitored using Dagshub to ensure the annotators' comprehensive understanding of the process and to address any real-time queries.

Table 4. Annotation guideline table.

Label	Meaning	Description & Examples (If Applicable)
Tweet Type	**Offensive**	Contains any offensive terms, either general or specific to an individual or group. General Example (Arabic): 'حظي مع الأفلام زباالة'. Translation: 'My luck with movies is trash'. Specific Example (Arabic): 'الله ياخذك يا متخلف يا طرش بحر'. Translation: (Derogatory terms directed at a specific person)
	Non-Offensive	Neutral words that aren't meant to offend. Example (Arabic): 'صباح الخير، أحوال الطقس اليوم كأننا في لندن'. Translation: 'Good morning, today's weather feels like we're in London'.
Type of Offense	**Hate Speech**	Any term that shows hatred towards individuals or groups based on their race, gender, orientation, etc.
	Sarcasm	Words that appear complimentary but are meant sarcastically. Example (Arabic): 'سوبهان الله.. خف علينا يا أجمل واحد في العالم'. Translation: 'Ease up on us, the most beautiful person in the world... Praise be to God'. (Meant sarcastically)
	General Insult	The tweet contains abrasive language not directed at anyone specific. Example (Arabic): 'حظي مع الأفلام زباالة'. Translation: 'My luck with movies is trash'.
Type of Hate Speech	**Racial**	Hatred is directed towards a specific race. Example (Arabic): 'ياعبد ياأسود، ياايماني، ياحضري، يابدوي'. Translation: - ياعبد: "O slave" (This is a derogatory term.) - ياأسود: "O black" (This can be seen as a derogatory reference to skin color.) - ياايماني: "O Yemeni" (Referring to someone from Yemen.) - ياحضري: "O urban" (Referring to someone from an urban area.) - يابدوي: "O Bedouin" (Referring to a desert-dwelling Arab.)
	Gender	Insults are based on the opposing gender. Example (Arabic): 'أنتم يا الرجال، الحريم دائما'. Translation: 'You men, women always...
	Sport	Any insult due to sports affiliations, either towards teams or individuals in the sports field.
	Political/Religious	Insults stemming from religious or political differences.
	Vulgar	Any sexually explicit content.
	Insult–Violence	Direct insult towards someone without relating to the above reasons. Example (Arabic): 'ياحمار ما تفهم!'. Translation: 'You donkey, you don't understand!'
	Other	Undefined category; ideally chosen very rarely.
Any Note		In this column, the reasons for deletion are recorded, whether it's due to a lack of understanding requiring further review or if something in the tweet was noticed.

To further ensure the reliability of our annotations, we calculated the Fleiss' Kappa statistic for the annotated dataset. The result was over 70%, which is considered a good

level of agreement among the annotators, demonstrating the consistency and reliability of the annotations across the dataset.

3.2. Post-Annotation Data Refinement

Based on the annotators' notes, tweets that were either incomprehensible or written in a non-Arabic language were removed, reducing the dataset from 28,000 to 24,500 tweets.

In Level 3, the hate speech tweets were initially classified into six categories: racial, gender, sport, political/religious, vulgar, and violent. To optimize the dataset, some categories were merged, resulting in three primary categories: religious/politics/racial, Sports, and violence/insult. Gender-specific data were categorized under man/woman.

Using Python, the dataset was further processed by replacing user mentions with the string "USER", URLs with "URL", and newline indicators with "NL". These steps standardized and anonymized the dataset while preserving the structure of the tweets, ensuring that the dataset was both comprehensive and refined, setting the foundation for the next phases of our study.

3.3. Annotation Category Examples

Upon completing the data collection and annotation processes, it is essential to summarize and present the final results. This section highlights representative examples from each category, as shown in Table 5, offering insights into the nuances and distinctions within our annotated dataset.

Table 5. Saudi Offensive Dataset [SOD] annotation examples.

Label	Category	Example 1	Example 2
Offensive	Non-Offensive	أسأل الله أن يرزقني ويرزقكم فوق مآ تتمنون سعآدة تسع الكون بأكملهNLNL".NLNL سنة سعيدة عليكم متابعينيNLNL#السنه_الجديده_2022 (Translation: I ask God to bless me and you beyond our wishes. Happiness that fills the entire universe.)	تغيب عن عيني ولا تغيب روحك عن روحي (Translation: You disappear from my eyes, but your soul never leaves mine.)
	Offensive	الفار حرمني من الفرحه حسبي الله بس يا عيني يا ماني (Translation: The VAR deprived me of joy, I rely on God, oh my eyes, oh my fate)	صبرنا_نفاذ_قبل_معين_اقيلوا#اذا عفوه من منصبه جابو العنّ منه (Translation: If they pardoned him from his position, they brought disaster from him #Remove_Mu'een_Before_Our_Patience_Runs_Out)
Hate Speech	General Insult	شششنووو القرف الي انا شفته البيوم؟؟؟ (Translation: What the hell did I see today?)	الحين ليش المباره بجدة؟ انا ملعب الجوهره يسبب لي تشاؤم والعياذ بالله (Translation: Why is the match in Jeddah now? The Jewel Stadium brings me pessimism, I seek refuge in God)
	Sarcasm	#الارجنتين_كرواتياNL الاسم: دي بول الوظيفه: الحارس الشخصي لميسي (Translation: #Argentina_Croatia Name: De Paul, Occupation: Messi's personal guard)	سؤال للبنات الشخص اللي قالكNL هيحارب الدنيا عشانك لسى يحارب والا ماتNL (Translation: A question for the girls, the person who told you "He would fight the world for you" , is he still fighting or did he die?)
	Hate Speech	قسم بالله حكم ضايع (Translation: I swear to God, the referee is lost)	USER دور مقزز ومقرفNLتاريخه الفني زفت في زفت (Translation: USER, his artistic history is crap upon crap, a disgusting and repulsive role)

Table 5. *Cont.*

Label	Category	Example 1	Example 2
Hate Speech Types	**Religious/ Politics/ Racial**	جرب تزور الشعب السعودي علشان يقومون ب USER واجبك سوف يستقبلونك👞👞👞👞 افهم وخل الحمير لي وراك تفهم الشعب ا لسعودي محمد بن سلمان وأنتم مو أهل رد معرو ف انتم ناكرين للمعروف منافقين خونه (Translation: USER, try visiting the Saudi people to fulfill your duty, they will receive you with shoes👞👞👞👞. Understand and leave the donkeys behind you. The Saudi people understand Mohammed bin Salman, and you are not people of gratitude, you are ungrateful, hypocrites, traitors)	حتى ف ه من يدعي أنه سني يبوس جزمة ا USER USER USER لخميني ويقدم أخته وأمه متعه لي الفرس خذ حزب الإخوان الخونه خذ حماس هذول يدعون أنهم سنه ولكنهم جنو د لي الفرس يحاربون الإسلام والع رب ومعهم حزب الإصلاح اليمني كلهم يبوسون جزمة الخميني (Translation: USER USER USER, there are even those who claim to be Sunni, they kiss Khomeini's shoes, and offer their sisters and mothers for pleasure to the Persians. Take the traitor Muslim Brotherhood, take Hamas. These claim to be Sunnis, but they are soldiers for the Persians, fighting Islam and the Arabs, and with them is the Yemeni Reform Party, all of them kissing Khomeini's shoe)
	Sports	الفار حرمني من الفرحه حسبي الله بس يا عيني يا ماني❤ (Translation: The VAR deprived me of joy, I rely on God, oh my eyes, oh my fate ❤)	اقذر مهاجم كوستا (Translation: The dirtiest attacker, Costa)
	Violence/Insult	لعنة الله وسخط عليه ابن ال USER🐷 (Translation: May God's curse and wrath be upon him, son of the 🐷)	عندنا إذا واحد يبي يسب واحد ياثورNLآخر لأنه ما يفهم أو ما يستوعب الكلام يقول له🐂 وغاب عنه هذا الذي 🐄🐂أو يابقره يسب أنه هو لم ينتج ولا شيء شاهد وتأملNLما أنتجه الثور أو أنتجته البقره!!👇👇 URL (Translation: In our culture, if someone wants to insult another because they don't understand or grasp something, they call them "bull" or "cow".)
	Gender: Male–Female	وبعدين يجي يقول البنات مايعرفو يسوقو☺ (Translation: And then he comes saying girls don't know how to drive ☺)	اغلب بنات اليوم كذا اجل تيهاUSER تصحى الصباح تسوي لك الفطور والظهر تسوي لك الغدا حلم ابليس في الجنه (Translation: Most of the girls these days are like this, do you expect her to wake up in the morning, prepare breakfast for you and then lunch? It's like Iblis's dream in paradise)

4. Descriptive Analysis

In this section, we examine the Saudi Offensive Dataset (SOD). The analysis begins with an exploratory data analysis (EDA), where we review basic statistics, tweet lengths, and word counts to gain an initial understanding of the data. We then proceed to token analysis, conducting a thorough review of unigrams, bigrams, and trigrams, as well as a focused examination of emojis and their associated sentiments.

4.1. Exploratory Data Analysis (EDA)

This section presents the exploratory data analysis (EDA) of the Saudi Offensive Dataset (SOD) to obtain basic statistics, analyze tweet lengths, and assess word counts, providing insights into the linguistic patterns and characteristics of Saudi offensive language. The results of this analysis are illustrated in Table 6 and Figures 5–7.

Table 6. Exploratory data analysis of SOD.

	All Tweets		Offensive Tweets				Hate Speech Tweets	
Tweet	**Offensive**	**Non-Offensive**	**General Insult**	**Hate Speech**	**Sarcasm**	**Sport**	**Religious–Political–Racial**	**Insult–Violence**
Count	7008	17,509	1588	4707	717	1500	825	2274
				Word				
Avg. length	16.08	17.48	11.77	17.96	13.30	19.33	24.67	14.57
Median	12	13	9	14	10	16	22	11
Minimum	1	1	1	1	1	1	2	1
Max	63	61	61	63	54	60	63	63
				Character				
Avg. length	92.93	108.32	65.56	103.89	81.74	117.40	141.96	80.89
Median	67	82	47	79	62	97	127	57
Minimum	2	2	2	5	10	6	13	5
Max	319	360	290	319	297	319	307	301

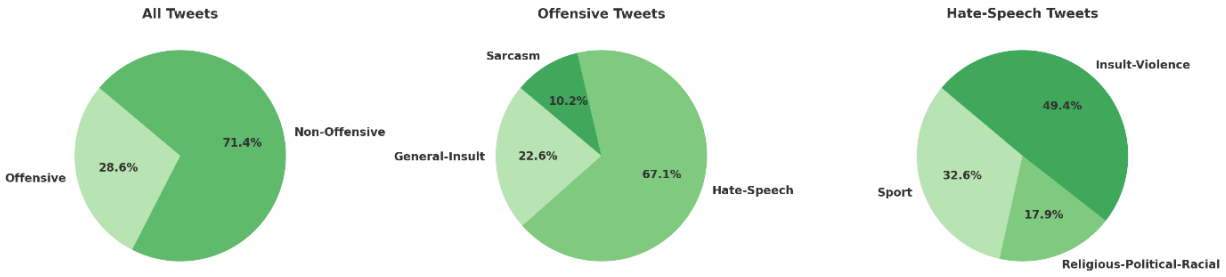

Figure 5. Saudi Offensive Dataset [SOD]—classes percentage.

The EDA reveals that the majority of tweets are categorized as non-offensive, suggesting that everyday linguistic exchanges are more prevalent than those containing hate speech or insults, as shown in the class distributions in Figure 5. Non-offensive tweets exhibit higher average and median word and character lengths compared to offensive tweets, indicating that they may be more detailed or conversational. For hate speech tweets, there is a clear trend: more complex or serious topics (such as religious–political–racial) tend to have longer tweets in terms of both words and characters. Conversely, general insult tweets are shorter, possibly reflecting a more impulsive or less thought-out nature.

4.2. Token Analysis

In this section, we delve into the analysis of linguistic tokens derived from the SOD. Token analysis is a fundamental aspect of computational linguistics and text mining, essential for understanding linguistic patterns and usage in natural language processing (NLP). A token, in the context of NLP, refers to a sequence of characters grouped as a meaningful semantic unit for processing [34]. Typically, tokens represent words, numbers, or punctuation marks.

- **Unigrams:** Unigrams are the simplest form of n-gram analysis, where 'n' denotes the number of contiguous items in a sequence. For unigrams, the sequence consists of individual tokens or words. Analyzing unigrams allows us to assess the frequency and distribution of standalone words within the text corpus [34].
- **Bigrams:** Bigrams build on unigrams by analyzing pairs of contiguous tokens. This approach provides insights into common two-word phrases or collocations within the dataset, offering a deeper understanding of language structure and contextual usage [34].
- **Trigrams:** Trigrams involve the analysis of triples of contiguous tokens, further enriching the context captured by the analysis. Trigrams help identify common phrases or

expressions, offering a more nuanced view of language patterns compared to unigrams and bigrams [34].

- **Emojis:** The analysis also extends to emojis, ideograms, and smileys used in electronic messages and web pages. Emojis have become integral to online communication, often conveying emotions and replacing traditional text. Their analysis provides unique insights into the emotional undertones and sentiments within the dataset [35].

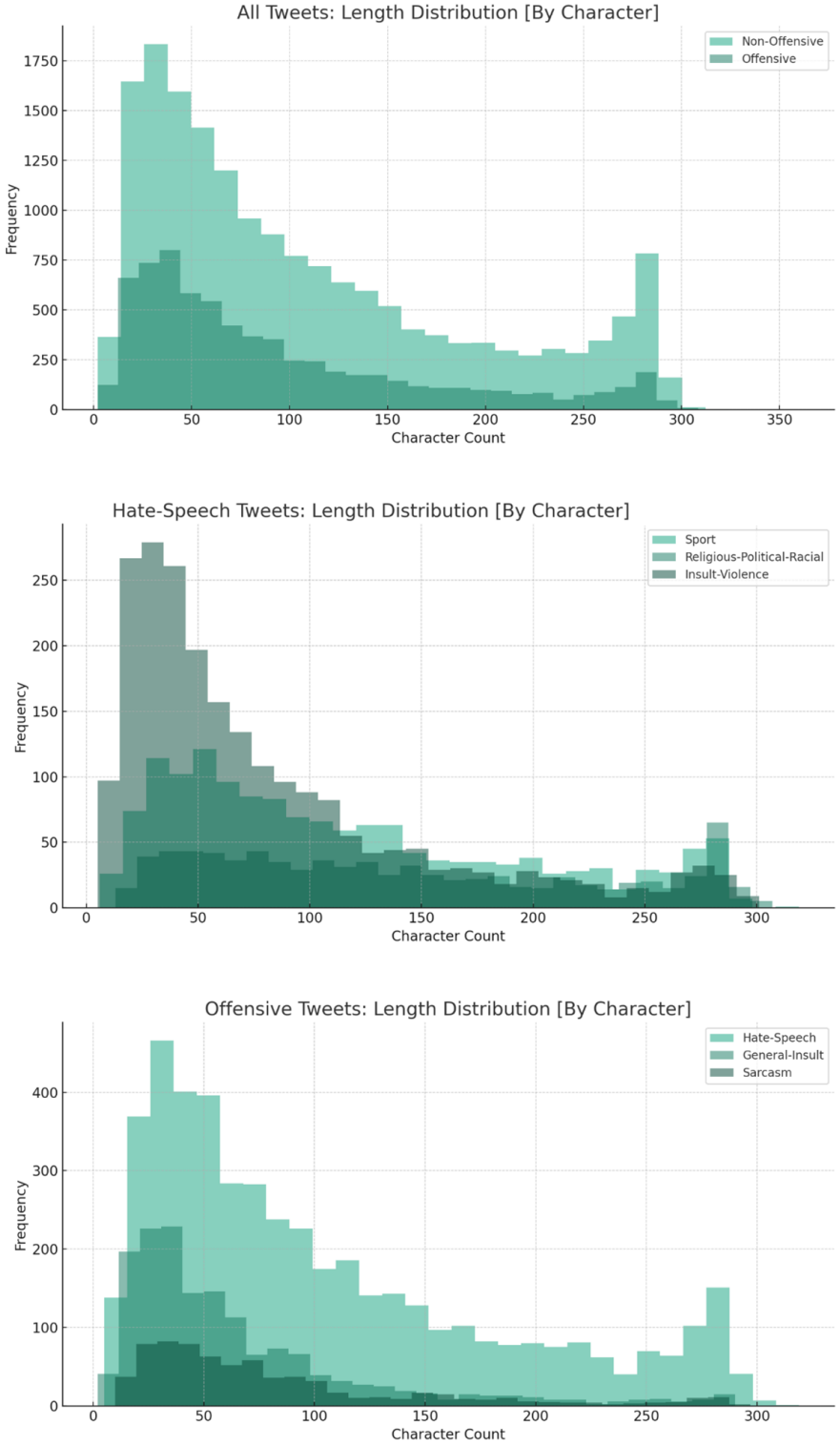

Figure 6. Saudi Offensive Dataset [SOD]—tweet length distribution.

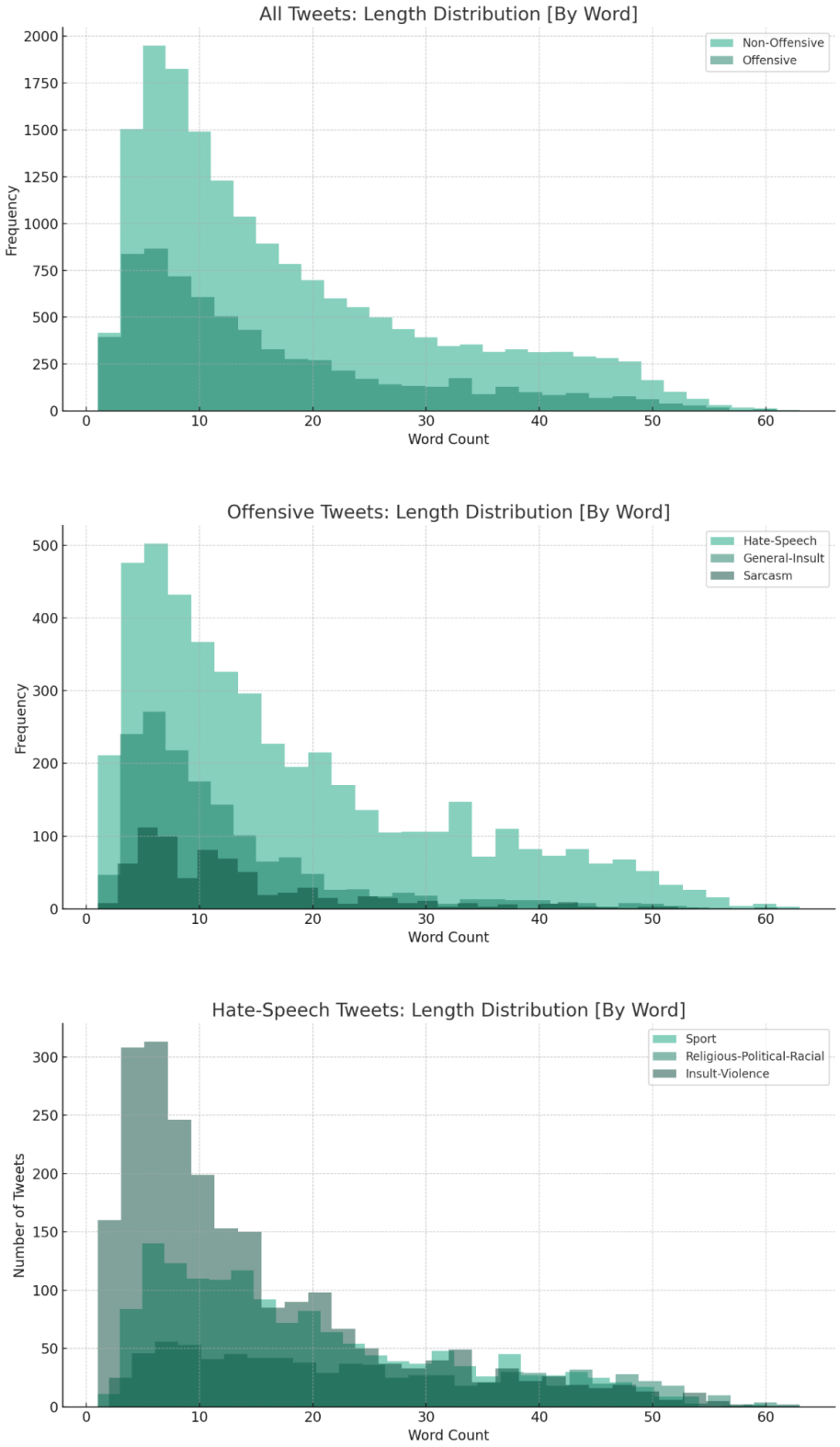

Figure 7. Saudi Offensive Dataset [SOD]—word count distribution.

4.2.1. All Tweets: Offensive or Not Offensive

In the analysis of all tweets, as shown in Table 7 and Figure 8, we observe distinct differences in how offensive and non-offensive language is used, reflecting varying communication styles. The frequent use of "USER" in both offensive and non-offensive contexts

suggests that these tweets originate from a social media platform where direct addressing is common.

Table 7. Top 10 token analysis—offensive.

All Tweets							
Offensive				Non-Offensive			
Unigram	**Bigram**	**Trigram**	**Emoji**	**Unigram**	**Bigram**	**Trigram**	**Emoji**
USER	USER USER	USER USER USER	😠	USER	USER USER	USER USER USER	🤍
من	الله يلعن	الله ونعم الوكيل	👎	من	الله USER	ان شا الله	🖇️
في	USER الله	كذابين كذابين كذابين	😂	في	شا الله	USER USER الله	😠
URL	USER والله	USER USER والله	🐅	URL	في كل	صلى الله عليه	😂
على	قلة ادب	حسبي الله ونعم	😠	الله	كل شي	الله عليه وسلم	👎
الله	USER هذا	USER USER لا	💩	على	الف مبروك	عام وانتم بخير	🤍
و	حسبي الله	USER USER هذا	🍆	و	USER URL	عام ونتم بخير	🔳
يا	USER لا	USER USER الله	😂😂	لا	USER لا	USER USER لا	🔞
ما	تفو عليك	يا USER USER	🖇️	ما	كل عام	ما شا الله	💔
لا	اقسم بالله	حسبي الله USER	😠😠	كل	الشباب#نادي	NLUSER NLUSER NLUSER	🐃

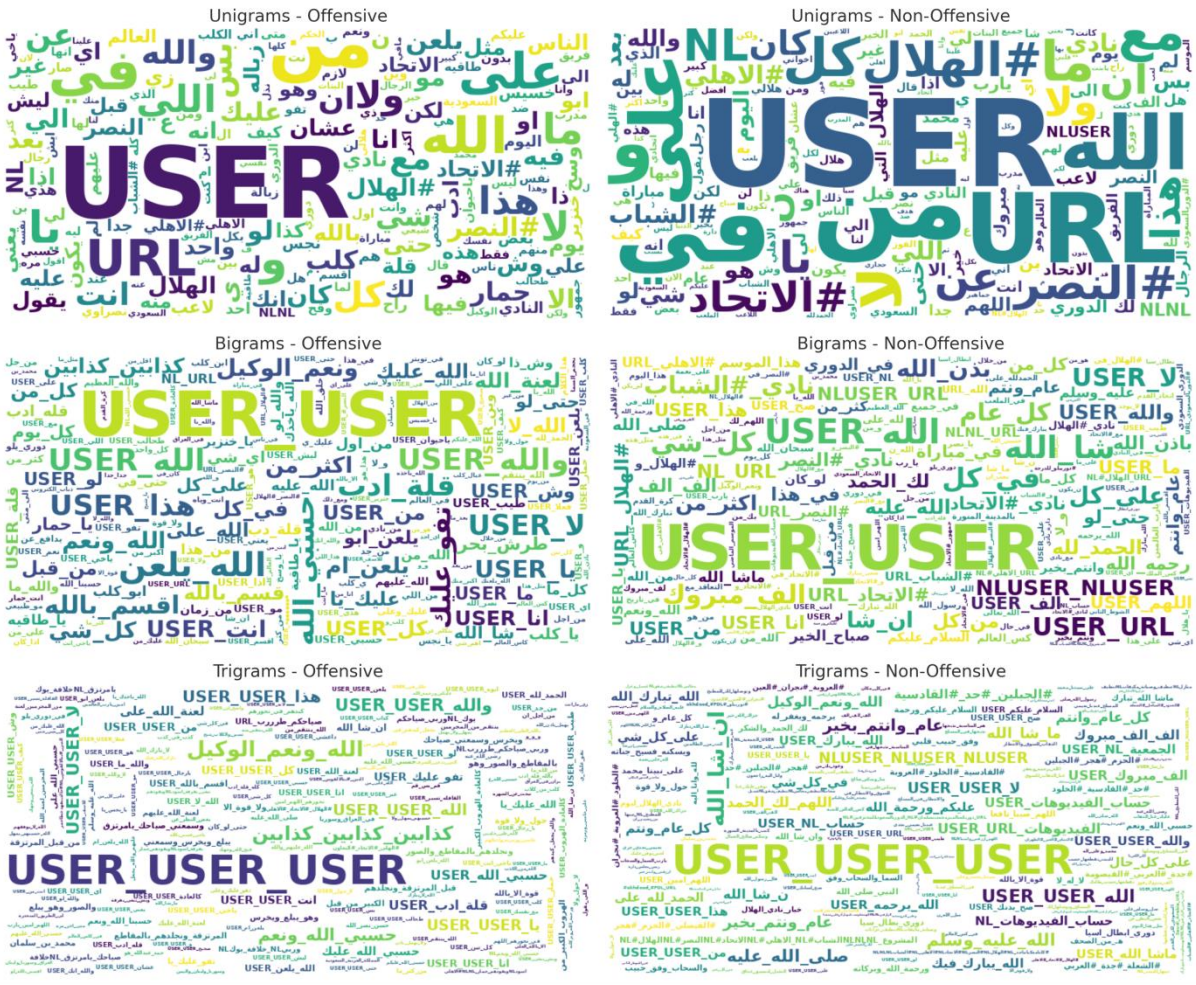

Figure 8. N-gram word cloud—all tweets.

In the offensive category, we note the presence of more aggressive language, including phrases like "الله يلعن" (may God curse) and "قلة ادب" (lack of manners), along with stronger expressions in trigrams. Emojis like 😠 (angry face) and 👎 (thumbs down) further confirm the negative sentiment and confrontational tone in these texts.

On the other hand, the non-offensive category consists of milder expressions, often with religious connotations like "إن شاء الله" (God willing) and "الله ونعم الوكيل" (God is the best guardian), frequently found in bigrams and trigrams. Emojis like 💙 (blue heart) and 💦 (sweat droplets) suggest a softer, potentially more positive or neutral emotional tone.

4.2.2. Offensive Tweets: General Insult, Hate Speech, and Sarcasm

In our detailed examination of offensive tweets, shown in Table 7 and Figure 9, we categorize offensive language into three groups: general insult, hate speech, and sarcasm. The presence of "USER" indicates personal engagement and direct addressing in this largely interactive and potentially confrontational dataset.

Figure 9. N-gram word cloud—offensive tweets.

The word clouds for Unigrams, Bigrams, and Trigrams reveal an increase in the complexity of offensive language. Bigrams and Trigrams expose phrases like "الله يلعن" (may God curse) and "كذابين كذابين كذابين" (liars liars liars), showing a progression from individual cursing to collective insults.

Emojis, as illustrated in Table 8, serve as non-verbal extensions of offensive language. The angry face 😠 and other negative symbols frequently accompany hate speech, emphasizing the aggressive tone and amplifying expressions of disdain and contempt. The Sarcasm category is notable for its higher use of laughing emojis 😂, suggesting ridicule or irony. This aligns with sarcasm's subtler and often humorous form of offense.

Table 8. Top 10 token analysis—offensive tweets.

Offensive Tweets											
General Insult				Hate Speech				Sarcasm			
Unigram	Bigram	Trigram	Emoji	Unigram	Bigram	Trigram	Emoji	Unigram	Bigram	Trigram	Emoji
user	user user	user user user	😠	user	user user	user user user	😠	user	user user	user user user	😂
من	قلة ادب	كذابين كذابين كذابين	😠	من	هذا user	الله ونعم الوكيل	👎	url	دوري يلو	في دوري يلو	😂😂
url	الله يلعن	الله ونعم الوكيل	💩	في	الله user	user user والله	🤦	من	الدوري السعودي	user user طحالب	😂😂😂
في	كذابين كذابين	قلة ادب user	👎	الله	والله user	user user لا	💩	nl	user طحالب	user user يا	😠
الله	يلعن ام	حسبي الله ونعم	🤦	على	الله يلعن	حسبي الله ونعم	😂	في	الهلال nl	nlالهلال nl	😂😂😂😂
على	user الله	قله ادب user	😂	url	حسبي الله	user user هذا	😠	الهلال	والله user	nlالاهلي url	🤦
ما	user قلة	الله user user	😠😠😠	يا	اقسم بالله	user user الله	🫳	النصر	user يا	ترجيع ترجيع ترجيع	🖼
يلعن	تفو عليك	من user user	🫳	ما	user لا	لعنة الله على	🫁	على	طاقية يا	في كرة القدم	😠
لا	حسبي الله	وش user user	😠😠	nl	تفو عليك	محمد بن سلمان	😠😠	يا	طاقيه يا	على أدلة دارون	👎
ادب	قله ادب	حول ولا قوة	😠😠	هذا	النصر nl	user user يا	😂😂	nlnl	nl nl	حافظ مش فاهم	🦃

4.2.3. Hate Speech Tweets: Sport, Religious–Political–Racial, and Insult–Violence

When we examine the hate speech tweets using Table 9 and Figure 10, we observe that the token analysis across three hate speech categories—sport, religious–political–racial, and insult–violence—reveals distinct language patterns that reflect the central themes of each category.

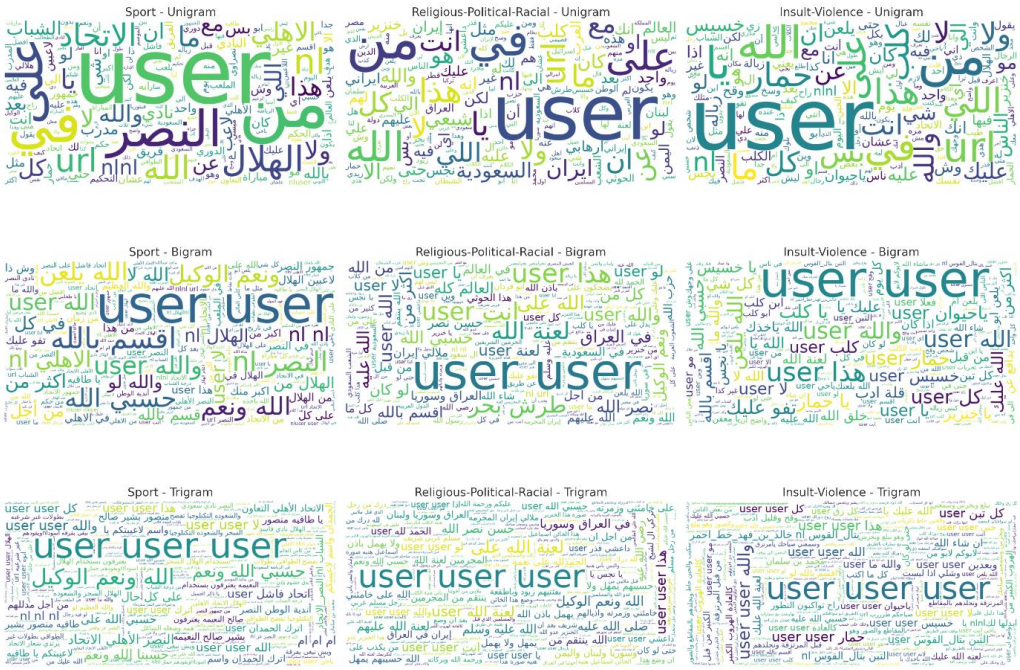

Figure 10. N-gram word cloud—hate speech tweet.

Table 9. Top 10 token analysis—hate speech tweets.

Hate Speech Tweets											
Sport				Religious–Political–Racial				Insult–Violence			
Unigram	Bigram	Trigram	Emoji	Unigram	Bigram	Trigram	Emoji	Unigram	Bigram	Trigram	Emoji
user	user user	user user user	😠	user	user user	user user user	👎	user	user user	user user user	😠
من	nl النصر	الله ونعم الوكيل	👎	من	هذا user	لعنة الله على	🐂	من	الله user	لا user user	🐂
النصر	اقسم بالله	الدوري مع وليد	😂	في	انت user	user لعنة الله	💩	الله	هذا user	الله user user	💩
في	حسبي الله ونعم	حسبي الله ونعم	👎👎	الله	لعنة الله	user user هذا	💩	يا	والله user	user user والله	👎
الهلال	الله يلعن	النصر ضمك الدوري	😠	على	طرش بحر	الله ونعم الوكيل 😠		في	لا user	الله ونعم الوكيل	😠
على	user والله	محمد بن سلمان	😂😂	url	الله على	حسن نصر الله 😂		url	تفو عليك	حسبي الله ونعم	😂
nl	nl الاهلي	user user والله	😠😠 😠	هذا	user والله	SAفي العراق وسوريا		على	الله يلعن	user user يا	💩
الاتحاد	nl الهلال	النصر الاتفاق دوري	😠😠	يا	user الله	الله عليه وسلم	🫀	ما	حسبي الله	user user هذا	😠😠
الله	الله ونعم	حسبي الله عليك	💩	ما	نصر الله	لعنة الله عليهم	▨	اللي	user يا	الله عليك يا	🫀
url	مع وليد	user user لا	🫀	لا	اكثر من	صلى الله عليه	😈	حمار	يا حمار	user user حمار	😈

The "Sport" category frequently mentions "USER" and specific sports clubs like "النصر" and "الهلال". This indicates passionate discussions and potentially heated debates in sports-related conversations. In the "Religious-Political-Racial" category, we encounter a blend of religious references and politically charged language, with phrases like "لعنة الله على" (curse of God on) and references to political figures, indicating discussions deeply rooted in socio-political and religious sentiments.

In contrast, the "Insult-Violence" category exhibits a more aggressive tone, with Bigrams and Trigrams expressing curses and insults, supported by aggressive emojis like 😠 (angry face) and 👎 (thumbs down) in all categories. These emojis underscore the strong emotions and confrontational nature of communication.

5. Data Experiment

To establish a baseline system for the Saudi Offensive Language Dataset (SOD), we conducted a series of experiments. The following subsections detail the experimental setup, present the results, and provide a comparative analysis with other relevant datasets.

5.1. Experimental Setting

We applied three categories of computational models to the classification tasks: machine learning (ML), traditional deep learning (DL), and transformer-based deep learning models. The Saudi Offensive Language Dataset (SOD) was split into an 80/20 ratio for training and testing across all experiments. Evaluation metrics included accuracy, precision, recall, F1-Score, and F1-Macro to comprehensively assess model performance.

- **Machine Learning (ML):** We employed various classifiers, including support vector machine (SVM), Gaussian naive Bayes, multinomial naive Bayes, random forest, logistic regression, and K-nearest neighbors. Text data was vectorized using term frequency-inverse document frequency (TF-IDF) with n-grams ranging from unigrams to trigrams. The scikit-learn library was utilized for model training, with default hyperparameters unless otherwise specified.

- **Traditional Deep Learning (DL):** Three models—feedforward neural network (FFNN), convolutional neural network (CNN), and gated recurrent unit (GRU)—were implemented using TensorFlow and Keras. The text data was tokenized and padded to a maximum sequence length of 128 tokens, with an embedding layer of 100 dimensions applied across all models. The FFNN included a dense layer with 128 units, the CNN used 128 convolutional filters with a kernel size of 5, and the GRU incorporated a single GRU layer with 128 units. All models were trained for 3 epochs using the Adam optimizer and a binary cross-entropy loss function.
- **Transformer-Based DL Model (AraBERT):** The aubmindlab/bert-base-arabertv02-twitter model was fine-tuned on the SOD dataset. Tokenization was conducted using HuggingFace's AutoTokenizer, with a maximum sequence length of 128 tokens. The model was trained using the Adam optimizer with a learning rate of 2×10^{-5} and epsilon set to 1×10^{-8} over 2 epochs. To enhance model robustness, 5-fold cross-validation was employed, and metrics were computed using the HuggingFace Trainer API.

5.2. Experimental Results

Given the class imbalance in the dataset, F1-Score and F1-Macro were prioritized over accuracy as the key evaluation metrics [36]. As shown in Table 10 and Figure 11, the transformer model (AraBERT) consistently outperformed both ML and traditional DL models across all tasks. This performance underscores the model's advanced capability in capturing complex linguistic patterns and nuances specific to the Saudi dialect.

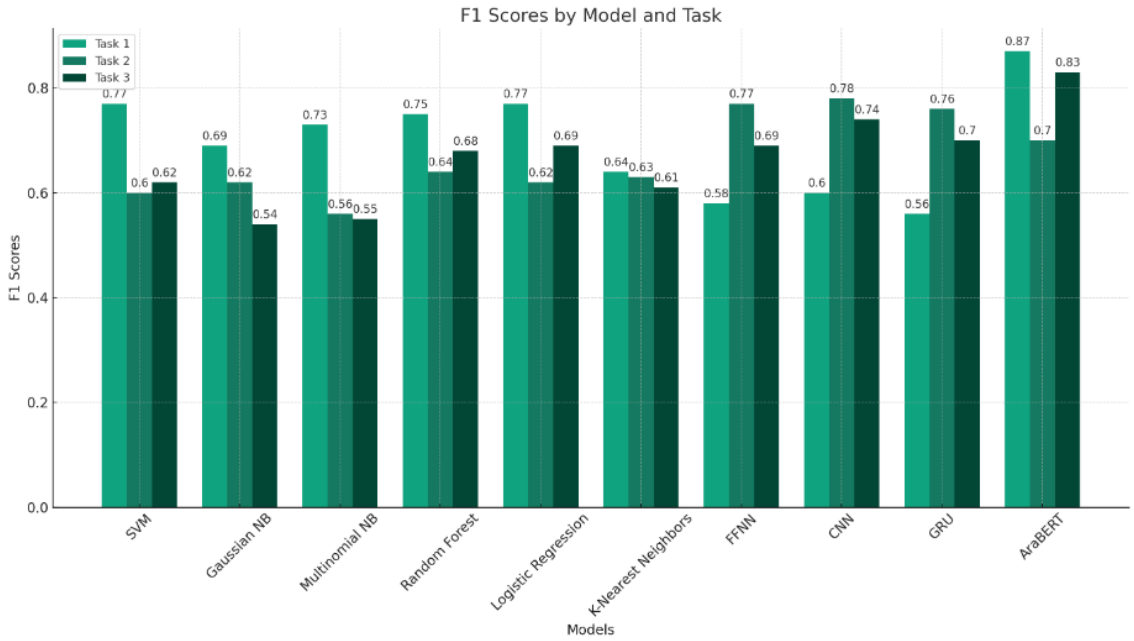

Figure 11. Performance of ML, DL, and transformer model in Saudi dialect tweet classification.

Traditional DL models exhibited mixed results but generally outperformed ML models in multi-class classification tasks. However, ML models demonstrated reasonable performance in the simpler binary classification of offensive and non-offensive tweets, proving to be both sufficient and efficient, particularly when computational resources are limited.

Table 10. Performance of ML, DL, and transformer-based models in Saudi dialect tweet classification.

Classification Task	Model	Classifier	Accuracy	Precision	Recall	F1-Score	F1-Macro
All Tweets: Offensive or Non-Offensive	**ML**	SVM	0.80	0.81	0.80	0.77	0.69
		Gaussian naive Bayes	0.67	0.71	0.67	0.69	0.63
		Multinomial naive Bayes	0.78	0.81	0.78	0.73	0.62
		Random forest	0.78	0.79	0.78	0.75	0.65
		Logistic regression	0.80	0.80	0.80	0.77	0.69
		K-nearest neighbors	0.71	0.65	0.71	0.64	0.49
	DL	FFNN	0.78	0.65	0.53	0.58	0.72
		CNN	0.79	0.65	0.55	0.60	0.73
		GRU	0.77	0.61	0.52	0.56	0.70
	Transformer	AraBERT	**0.87**	**0.87**	**0.87**	**0.87**	**0.84**
Offensive Tweets: General Insult, Hate Speech, or Sarcasm	**ML**	SVM	0.70	0.69	0.70	0.60	0.34
		Gaussian naive Bayes	0.64	0.61	0.64	0.62	0.43
		Multinomial naive Bayes	0.69	0.66	0.69	0.56	0.28
		Random forest	0.70	0.66	0.70	0.64	0.43
		Logistic regression	0.71	0.69	0.71	0.62	0.38
		K-nearest neighbors	0.66	0.63	0.66	0.63	0.45
	DL	FFNN	0.70	**0.79**	0.75	0.77	0.66
		CNN	0.70	**0.79**	**0.77**	**0.78**	**0.67**
		GRU	0.69	0.79	0.74	0.76	0.65
	Transformer	AraBERT	**0.74**	0.75	0.74	0.70	0.48
Hate Speech Tweets: Sport, Religious–Political–Racial, or Insult–Violence	**ML**	SVM	0.68	0.77	0.68	0.62	0.50
		Gaussian naive Bayes	0.54	0.57	0.54	0.54	0.52
		Multinomial naive Bayes	0.62	0.76	0.62	0.55	0.43
		Random rorest	0.71	0.77	0.71	0.68	0.59
		Logistic regression	0.72	0.74	0.72	0.69	0.61
		K-nearest neighbors	0.64	0.63	0.64	0.61	0.54
	DL	FFNN	0.69	0.70	0.70	0.69	0.66
		CNN	0.74	0.75	0.74	0.74	0.73
		GRU	0.69	0.70	0.70	0.70	0.66
	Transformer	AraBERT	**0.83**	**0.83**	**0.83**	**0.83**	**0.81**

The bolded values represent the highest performance achieved by a classifier for each specific dataset and classification task.

5.3. Comparative Analysis

In this section, we compare the results of our study on the Saudi Offensive Language Dataset (SOD) with other significant studies focused on offensive language detection in Arabic tweets. Specifically, we analyze the results from the OSACT5 Shared Task on Arabic Offensive Language and Hate Speech Detection [32] and the study by Mubarak et al. [27].

The datasets used in these studies encompass a broader range of Arabic dialects, whereas our dataset is specifically tailored to the Saudi dialect. Table 11 below summarizes the performance metrics of our study compared to the best results from the OSACT5 Shared Task [32] and the study by Mubarak et al. [27].

The results demonstrate that our model, fine-tuned on the Saudi dialect dataset, performs as well as or better than the top-performing models from the OSACT5 Shared Task [32] and the study by Mubarak et al. [27]. This highlights the importance of dialect specificity, suggesting that focusing on a particular dialect can significantly enhance the effectiveness of offensive language detection models.

Table 11. Comparative performance metrics for Arabic offensive language detection models.

Reference	Model	Preprocessing Methods	Accuracy	Precision	Recall	F1-Score
Our Study (SOD)	AraBERT: Transformer model pre-trained on Arabic, fine-tuned on Saudi dialect dataset.	Replaced user mentions with "USER", URLs with "URL", and newline indicators with "NL".	0.87	0.87	0.87	0.87
Best of OSACT5 Shared Task [37]	Ensemble: Combines MARBERT (without emojis), AraBERT-Large-Twitter, QARiB, and others.	Removed non-Arabic letters, punctuation marks, digits, Arabic diacritics, repeated characters, and replaced URL, @USER, and e-mail.	0.87	0.86	0.85	0.85
Mubarak et al. [38]	AraBERT (TF-IDF + FastText): Combines TF-IDF and FastText embeddings with AraBERT.	Performed text tokenization, removed URLs, numbers, tweet-specific tokens (mentions, retweets, and hashtags); normalized Arabic letters.	0.85	0.83	0.83	0.83

6. Data Augmentation

The SOD dataset exhibits a class imbalance, with offensive language making up about one-third of the total data. This imbalance reflects the distribution of offensive language in real-world social media contexts [37]. To address this issue, data augmentation techniques were employed to enhance the diversity and quality of the data without requiring additional collection [38]. As a result, the SOD dataset expanded to over 35,000 tweets, ensuring equal representation across all classes.

Advanced augmentation approaches using transformations were tested but did not generate effective augmented data, likely due to the dialectical nature of the dataset, which lacks formal construction and writing rules. On the other hand, simple augmentation techniques, such as random deletion, word swapping, and punctuation insertion, significantly improved the performance of the detection system.

Table 12 highlights the effectiveness of these simple techniques compared to the baseline model, which used the unbalanced dataset without augmentation. The random punctuation insertion technique, which maintains word order while slightly altering sentence structure, showed the most significant improvement across all metrics (accuracy, precision, recall, F1-Score, and F1-Macro), achieving a score of 0.91. The random swap technique also demonstrated notable performance gains, outperforming the baseline model.

Table 12. Data augmentation techniques on SOD dataset.

Augmentation Techniques	Accuracy	Precision	Recall	F1-Score	F1-Macro
Baseline (No Augmentation)	0.87	0.87	0.87	0.87	0.84
Random Deletion	0.86	0.86	0.85	0.85	0.85
Random Swap	0.89	0.89	0.89	0.89	0.89
Random Punctuation Insertion	**0.91**	**0.91**	**0.91**	**0.91**	**0.91**

7. Conclusions

This study presents the Saudi Offensive Dialect dataset (SOD), comprising over 24,000 tweets, which marks a significant advancement in Arabic natural language processing (NLP), particularly in detecting offensive language within the Saudi dialect. The hierarchical annotation approach—from general offensive language to specific categories like general insults, hate speech, sarcasm, and further into hate speech subtypes—highlights the dataset's comprehensive nature and its potential for nuanced analysis.

The implementation of machine learning, traditional deep learning, and transformer-based models, particularly the AraBERT model, has achieved notable success. With data augmentation techniques addressing dataset imbalances, our models attained up to 91% accuracy in offensive language detection. This performance surpasses many existing efforts in this domain and underscores the value of dialect-specific datasets in enhancing detection accuracy compared to mixed-language datasets.

This paper's contributions include the development of a robust corpus, the introduction of a hierarchical annotation framework, and insights into the unique linguistic characteristics of the Saudi dialect. We have evaluated various NLP tools for identifying offensive language and employed data augmentation strategies to address dataset imbalances. These efforts aim to provide foundational insights and practical tools for further research in Arabic language processing.

This work encourages the creation of similar dialect-specific datasets within the Arabic linguistic domain, suggesting that such focused studies can lead to more effective NLP applications. It challenges the prevailing notion of expanding datasets to include more dialects, instead promoting the refinement of tools and techniques for individual dialects to maximize performance. The documented success in this study advocates for a more concentrated approach in future NLP research, focusing on enhancing and tailoring solutions to specific dialectal nuances rather than broadening the scope of current models.

Author Contributions: Conceptualization, A.A.; Methodology, A.A.; Software, A.A.; Formal analysis, A.A.; Investigation, A.A.; Data curation, A.A.; Writing—original draft, A.A.; Supervision, M.S.; Project administration, M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: We have made a GitHub (https://github.com/Afefa-Asiri/SOD-A-Corpus-for-Saudi-Offensive-Language-Detection-Classification, accessed on 20 August 2024) repository available for this work, which includes the code and resources used in the study. The dataset used in this research will be available on demand by contacting the authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Nobata, C.; Tetreault, J.; Thomas, A.; Mehdad, Y.; Chang, Y. Abusive Language Detection in Online User Content. In Proceedings of the 25th International Conference on World Wide Web, Montréal, QC, Canada, 11–15 April 2016; ACM Press: New York, NY, USA, 2016; pp. 145–153. [CrossRef]
2. Xiang, G.; Fan, B.; Wang, L.; Hong, J.; Rose, C. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In Proceedings of the 21st ACM International Conference on Information and Knowledge Management–CIKM'12, Maui, Hawaii, USA, 29 October–2 November 2012; ACM Press: New York, NY, USA, 2012; p. 1980. [CrossRef]
3. Abozinadah, E.A.; Mbaziira, A.V.; Jones, J.H.J. Detection of Abusive Accounts with Arabic Tweets. *Int. J. Knowl. Eng.* **2015**, *1*, 113–119. [CrossRef]
4. Mouheb, D.; Ismail, R.; Al Qaraghuli, S.; Al Aghbari, Z.; Kamel, I. Detection of Offensive Messages in Arabic Social Media Communications. In Proceedings of the 2018 International Conference on Innovations in Information Technology (IIT), Al Ain, United Arab Emirates, 18–19 November 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 24–29. [CrossRef]
5. Chowdhury, A.G.; Didolkar, A.; Sawhney, R.; Shah, R.R. ARHNet-Leveraging Community Interaction for Detection of Religious Hate Speech in Arabic. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Florence, Italy, 28 July–2 August 2019; Association for Computational Linguistics: Pennsylvania, PA, USA, 2019; pp. 273–280. [CrossRef]
6. Magdy, W.; Darwish, K.; Weber, I. #FailedRevolutions: Using Twitter to study the antecedents of ISIS support. *First Monday* **2016**, *21*, 1–14. [CrossRef]
7. Haidar, B.; Chamoun, M.; Serhrouchni, A. A Multilingual System for Cyberbullying Detection: Arabic Content Detection using Machine Learning. *Adv. Sci. Technol. Eng. Syst. J.* **2017**, *2*, 275–284. [CrossRef]
8. Alshaalan, R.; Al-Khalifa, H. Hate Speech Detection in Saudi Twittersphere: A Deep Learning Approach. In Proceedings of the Fifth Arabic Natural Language Processing Workshop, Barcelona, Spain, 12 December 2020; Zitouni, I., Abdul-Mageed, M., Bouamor, H., Bougares, F., El-Haj, M., Tomeh, N., Zaghouani, W., Eds.; Association for Computational Linguistics: Pennsylvania, PA, USA, 2020; pp. 12–23. Available online: https://aclanthology.org/2020.wanlp-1.2 (accessed on 19 November 2023).
9. Alshalan, R.; Al-Khalifa, H. A Deep Learning Approach for Automatic Hate Speech Detection in the Saudi Twittersphere. *Appl. Sci.* **2020**, *10*, 8614. [CrossRef]

10. Mohaouchane, H.; Mourhir, A.; Nikolov, N.S. Detecting Offensive Language on Arabic Social Media Using Deep Learning. In Proceedings of the 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), Granada, Spain, 22–25 October 2019; pp. 466–471. [CrossRef]
11. Al-Hassan, A.; Al-Dossari, H. Detection of Hate Speech in Social Networks: A Survey on Multilingual Corpus. In Proceedings of the 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), Granada, Spain, 22–25 October 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 466–471. [CrossRef]
12. Habash, N.Y. Introduction to Arabic Natural Language Processing. *Synth. Lect. Hum. Lang. Technol.* **2010**, *3*, 1–187. [CrossRef]
13. Abozinadah, E.A.; Jones, J.J.H. Improved Micro-Blog Classification for Detecting Abusive Arabic Twitter Accounts. *Int. J. Data Min. Knowl. Manag. Process.* **2016**, *6*, 17–28. [CrossRef]
14. Darwish, K.; Magdy, W. Arabic Information Retrieval. *Found. Trends® Inf. Retr.* **2014**, *7*, 239–342. [CrossRef]
15. Countries with Most X/Twitter Users 2023 | Statista. Available online: https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/ (accessed on 22 January 2024).
16. Habash, N.; Eskander, R.; Hawwari, A. A Morphological Analyzer for Egyptian Arabic. In Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology, Montréal, QC, Canada, 7 June 2012; Cahill, L., Albright, A., Eds.; Association for Computational Linguistics: Pennsylvania, PA, USA, 2012; pp. 1–9. Available online: https://aclanthology.org/W12-2301 (accessed on 14 February 2024).
17. Farghaly, A.; Shaalan, K. Arabic Natural Language Processing. *ACM Trans. Asian Lang. Inf. Process.* **2009**, *8*, 1–22. [CrossRef]
18. Almuqren, L.; Cristea, A. AraCust: A Saudi Telecom Tweets corpus for sentiment analysis. *PeerJ Comput. Sci.* **2021**, *7*, e510. [CrossRef] [PubMed]
19. Azmi, A.M.; Alzanin, S.M. Aara'–A system for mining the polarity of Saudi public opinion through e-newspaper comments. *J. Inf. Sci.* **2014**, *40*, 398–410. [CrossRef]
20. Al-Harbi, W.; Emam, A. Emam Effect of Saudi Dialect Preprocessing on Arabic Sentiment Analysis. *Int. J. Adv. Comput. Technol. (IJACT)* **2015**, *4*, 6.
21. Al-Twairesh, N.; Al-Khalifa, H.; Al-Salman, A.; Al-Ohali, Y. AraSenTi-Tweet: A Corpus for Arabic Sentiment Analysis of Saudi Tweets. *Procedia Comput. Sci.* **2017**, *117*, 63–72. [CrossRef]
22. Al-Thubaity, A.; Alharbi, M.; Alqahtani, S.; Aljandal, A. A Saudi Dialect Twitter Corpus for Sentiment and Emotion Analysis. In Proceedings of the 2018 21st Saudi Computer Society National Computer Conference (NCC), Riyadh, Saudi Arabia, 25–26 April 2018; pp. 1–6. [CrossRef]
23. Alqarafi, A.; Adeel, A.; Hawalah, A.; Swingler, K.; Hussain, A. A Semi-supervised Corpus Annotation for Saudi Sentiment Analysis Using Twitter. In Proceedings of the BICS 2018: 9th International Conference on Brain Inspired Cognitive Systems, Xi'an, China, 7–8 July 2018; Springer International Publishing: Berlin/Heidelberg, Germany, 2018; pp. 589–596. [CrossRef]
24. Alruily, M. Issues of Dialectal Saudi Twitter Corpus. *Int. Arab. J. Inf. Technol.* **2019**, *17*, 367–374. [CrossRef]
25. Bayazed, A.; Torabah, O.; AlSulami, R.; Alahmadi, D.; Babour, A.; Saeedi, K. SDCT: Multi-Dialects Corpus Classification for Saudi Tweets. *Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*, 216–223. [CrossRef]
26. Abozinadah, E.A.; Jones, J.H., Jr. A Statistical Learning Approach to Detect Abusive Twitter Accounts. In Proceedings of the International Conference on Compute and Data Analysis, in ICCD'17, Lakeland, FL, USA, 19–23 May 2017; Association for Computing Machinery: New York, NY, USA, 2017; pp. 6–13. [CrossRef]
27. Mubarak, H.; Darwish, K.; Magdy, W. Abusive Language Detection on Arabic Social Media. In Proceedings of the First Workshop on Abusive Language Online, Vancouver, BC, Canada, 4 August 2017; Association for Computational Linguistics: Pennsylvania, PA, USA, 2017; pp. 52–56. [CrossRef]
28. E Abdelfatah, K.; Terejanu, G.; A Alhelbawy, A. Unsupervised Detection of Violent Content in Arabic Social Media. *Comput. Sci. Inf. Technol. (CS IT)* **2017**, 1–7. [CrossRef]
29. Alakrot, A.; Murray, L.; Nikolov, N.S. Towards Accurate Detection of Offensive Language in Online Communication in Arabic. *Procedia Comput. Sci.* **2018**, *142*, 315–320. [CrossRef]
30. Albadi, N.; Kurdi, M.; Mishra, S. Are they Our Brothers? Analysis and Detection of Religious Hate Speech in the Arabic Twittersphere. In Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Barcelona, Spain, 28–31 August 2018; pp. 69–76. [CrossRef]
31. Mubarak, H.; Hassan, S.; Chowdhury, S.A. Emojis as anchors to detect Arabic offensive language and hate speech. *Nat. Lang. Eng.* **2023**, *29*, 1436–1457. [CrossRef]
32. Mubarak, H.; Al-Khalifa, H.; Al-Thubaity, A. Overview of OSACT5 Shared Task on Arabic Offensive Language and Hate Speech Detection. In Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection, Marseille, France, 25 June 2022; pp. 162–166. Available online: https://aclanthology.org/2022.osact-1.20 (accessed on 6 December 2022).
33. What Is Hate Speech? Rights for Peace. Available online: https://www.rightsforpeace.org/hate-speech (accessed on 26 December 2022).
34. Daniel, J.; Martin, J.H.; Peter, N.; Stuart, R. *Speech and Language Processing*, 3rd ed.; Pearson: London, UK, 2023.
35. Novak, P.K.; Smailović, J.; Sluban, B.; Mozetič, I. Sentiment of Emojis. *PLoS ONE* **2015**, *10*, e0144296. [CrossRef]

36. Ibrahim, M.; Torki, M.; El-Makky, N. Imbalanced Toxic Comments Classification Using Data Augmentation and Deep Learning. In Proceedings of the 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, USA, 17–20 December 2018; pp. 875–878. [CrossRef]
37. Mubarak, H.; Rashed, A.; Darwish, K.; Samih, Y.; Abdelali, A. Arabic Offensive Language on Twitter: Analysis and Experiments. In Proceedings of the Sixth Arabic Natural Language Processing Workshop, Kyiv, Ukraine (Virtual), 19 April 2021; pp. 126–135.
38. Alkadri, A.M.; Elkorany, A.; Ahmed, C. Enhancing Detection of Arabic Social Spam Using Data Augmentation and Machine Learning. *Appl. Sci.* **2022**, *12*, 11388. [CrossRef]