So Hateful! Building a Multi-Label Hate Speech Annotated Arabic Dataset

Wajdi Zaghouani[1], **Hamdy Mubarak**[2], **Md. Rafiul Biswas**[1]
[1]Hamad Bin Khalifa University, Qatar
[2]Qatar Computing Research Center, HBKU, Qatar

wzaghouani@hbku.edu.qa , hmubarak@hbku.edu.qa, mdbi30331@hbku.edu.qa

Abstract

Social media enables widespread propagation of hate speech targeting groups based on ethnicity, religion, or other characteristics. With manual content moderation being infeasible given the volume, automatic hate speech detection is essential. This paper analyzes 70,000 Arabic tweets, from which 15,965 tweets were selected and annotated, to identify hate speech patterns and train classification models. Annotators labeled the Arabic tweets for offensive content, hate speech, emotion intensity and type, effect on readers, humor, factuality, and spam. Key findings reveal 15% of tweets contain offensive language while 6% have hate speech, mostly targeted towards groups with common ideological or political affiliations. Annotations capture diverse emotions, and sarcasm is more prevalent than humor. Additionally, 10% of tweets provide verifiable factual claims, and 7% are deemed important. For hate speech detection, deep learning models like AraBERT outperform classical machine learning approaches. By providing insights into hate speech characteristics, this work enables improved content moderation and reduced exposure to online hate. The annotated dataset advances Arabic natural language processing research and resources.

Keywords: Hate Speech Detection, Offensive language, Dialectal Arabic, Multi-label Annotation

1. Introduction

Disclaimer: Due to the nature of this work, some examples may contain offensive, hateful, violent, and profane content. It is imperative to understand that such content does not reflect the authors' viewpoints in any way. Instead, we hope this work can help in detecting and preventing the spread of such harmful content.

Social media has revolutionized how people connect and share their thoughts, opinions, and experiences with just a few clicks. This has made information easily accessible, which is undoubtedly beneficial. However, the downside is that negative information and opinions are also easily spread. Hate speech is one negative form of expression that is prevalent on social media platforms, including Twitter, widely used and known for its limited-character posts. Hate speech is defined as statements that attack and belittle a person based on their group affiliation, including ethnicity, race, gender identity, political affiliation, spirituality, and religion (Pereira-Kohatsu et al., 2019).

Over the past few years, automatic hate speech detection models have been developed and described in the literature (e.g., Watanabe et al. 2018; Waseem and Hovy 2016). The creation of these models has become essential as the abundance of posts constantly created online by users makes manual flagging and deletion unfeasible. Hate speech often arises with the emergence of events around the world, and social media users come together through hashtags to either support a group or voice hate speech, often through echo chambers (Bagavathi et al., 2019). In 2015, the German government secured an agreement from social media platforms, including Twitter, to delete all hate speech targeting refugees within 24 hours of its occurrence on the platform (Lomas, 2017), recognizing the potential of hate speech to turn social media platforms into unsafe environments for targeted individuals and groups. Hateful content is known to spread faster and further than other content on social media (Mathew et al., 2019), with hateful users being densely connected and engaging highly with each other. Identifying and removing hate speech from social media is challenging for moderators (Masud et al., 2022), highlighting the urgency for automated management of hate speech.

Detecting hate speech in Arabic social media posts is a challenging task due to the diverse nature of the Arabic language. The language has various dialects that differ from each other and from Modern Standard Arabic (MSA). Moreover, social media users often post their thoughts, opinions, and information in their own dialect, which vary in terms of typing conventions. Since Arabic dialects were predominantly spoken and not written until the recent technological revolution, dialect typing can be inconsistent. Considering that Arabic is spoken in 25 countries worldwide, it is crucial to annotate and analyze Arabic tweets for hate speech.

Arab users tend to use emojis differently in

LREC-COLING 2024, pages 15044–15055
20-25 May, 2024.

their tweets compared to English-speaking cultures (Mubarak et al., 2023). Moreover, Hakami et al. demonstrated this cultural influence on the perception and use of emojis online by recruiting Arab annotators to identify sentiment in emojis without any textual context.

Similar to recent research (e.g., Waseem et al. 2017), the current project differentiates between hate speech and generally offensive language. This approach aims to reconcile research that focuses on only one aspect or the other to reach a consensus. We also look at hate speech directed at an individual, organization, or entity, as well as a particular outgroup (based on gender, race, religion, etc.). In addition, our annotations take sarcasm and irony into account, as hate speech is sometimes implicit and cannot be detected through lists of words or n-grams (Waseem et al., 2017). The main contributions of this paper are as follows:

1. We have created the largest multi-label, fine-grained Arabic hate speech dataset to date.

2. Our dataset is unique and versatile, with each tweet annotated with nine labels, such as sentiments, emotions, and valence, etc. This multi-purpose labeling makes it conducive for various research studies.

3. We have thoroughly documented the dataset's collection, creation, annotation methodology, and guidelines. This comprehensive documentation simplifies the process of reproducing these steps for future projects.

4. We conducted a comprehensive corpus analysis of the dataset, shedding light on the distinct features of Arabic hate speech discourse.

5. We have carried out experiments with several classification techniques, providing valuable insights that can contribute to advancements in this domain.

2. Related Works

Numerous research efforts have been dedicated to the development and examination of Arabic corpora for a range of NLP applications. Studies by Ahmed et al. (2022) and Zaghouani (2014) have detailed surveys on accessible Arabic corpora. Rosso et al. (2018) reviewed Arabic author profiling and irony detection. Charfi et al. (2019) created a detailed multi-dialectal Arabic corpus, and Rangel et al. (2020) performed an in-depth study on the variation of language and demographic aspects within Arabic. Furthermore, Abbes et al. (2020) unveiled a corpus focusing on dialectal Arabic irony, sourced from Twitter.

For hate speech, several studies have investigated the issue using different annotation approaches. For instance, Waseem and Hovy (2016) specified 16 target groups, while others used general labels for hate and profanity (Waseem et al., 2017). Qian et al. improved on basic hate speech annotation by classifying it into 40 hate ideologies that fall into 13 different categories. Similarly, Chatzakou et al. aimed to detect bullying and aggression on Twitter by labeling tweets as bullying, aggressive, spam, or normal.

Some researchers explored hate speech detection in Arabic social media using various computational approaches. For example, Kaddoura et al. examined deep learning methods like CNNs and LSTMs for classifying the sentiment of Arabic tweets. Their experiments found that LSTMs outperformed CNNs, and both surpassed traditional machine learning classifiers. Mohaouchane et al. compared classifiers like CNN-LSTM, CNN, and Bi-LSTM for detecting offensive language in Arabic social media posts. Their analysis determined that CNN-LSTM achieved the highest recall.

Ousidhoum et al. selected a list of emotions, including shock, sadness, disgust, anger, fear, confusion, and indifference, and asked annotators to choose based on how the tweet made them feel. The present project also includes a range of emotions, but the focus is on identifying the emotion in the tweet, rather than how it made the annotator feel. Moreover, Mubarak et al. tried to detect hate speech towards various groups, including gender, race, ideology, social class, religion, and disability, using emojis as anchors to build a dataset of 13K tweets, with 35% of the tweets labeled as offensive and 11% as hate speech.

Shapiro et al. detected offensive tweets, judged whether they contained hate speech, and identified the type from six classes. They tested various models and found that MarBERTv2 outperformed other BERT-based models trained on Arabic. Similarly, Bennessir et al. found that MARBERT with Quasi-recurrent neural networks (QRNN) performed best. Albadi et al. explored religious hate speech in Arabic Twitter. They looked at religious hate speech issues, gathered 6,000 tweets discussing religions, classified them with crowdsourcing, studied the labeled data, and identified primary hate targets. Their method used feature selection for religion-related phrases and scores indicating how well they identified hate or not. Their analysis found the best performance using pre-trained word embeddings with simple Recurrent Neural Networks (RNN) with Gated Recurrent Units (GRU).

Several studies have developed Arabic hate speech datasets and corpora. Abdelali et al. presented the Farasa segmentation tool for preprocessing Arabic text to improve comprehension. Al-

Hassan and Al-Dossari surveyed existing multilingual hate speech datasets, highlighting the lack of large-scale Modern Standard Arabic resources. Mubarak et al. released an annotated corpus of 17K Arabic tweets labeled for abusive language. Mulki et al. compiled the L-HSAB dataset containing Levantine Arabic tweets labeled as hateful, abusive, or normal.

A study in Jordan examined tweets on racism, journalism, sports, terrorism, and Islam to develop a hate speech tool (Aljarah et al., 2021). They cleaned the data by removing non-Arabic characters, numbers, symbols, punctuation, hashtags, and more. Their analysis found that RF classifiers performed best at detecting hate speech.

Targeted hate speech has also been investigated. Albadi et al. focused on detecting religious hate speech in Arabic Twitter. They identified primary hate targets and found deep learning models performed best. Almaliki et al. experimented with transformer architectures for Arabic hate speech detection, achieving strong results with Arabic BERT-Mini.

In summary, various studies have explored different approaches to annotating and detecting hate speech on Arabic Twitter. They gathered or analyzed tweets, detected or classified hate speech and offensive content, identified types of hate speech, explored models and features to improve detection, and evaluated different methods. By measuring performance, they aimed to develop effective ways of recognizing hate speech in Arabic on social media. These studies collectively underscore the importance of developing annotated Arabic corpora and applying deep learning for reliable hate speech detection within the challenges of informal dialectal social media posts. Our work builds on prior literature through a large-scale annotated dataset and experiments with transformers like AraBERT.

3. Data Collection and Annotation

3.1. Data Collection and Corpus Description

We collected 60M Arabic tweets between August 12, 2020, and October 4, 2020, and took a random sample of 600K tweets (1% of the original tweets). The tweets were pre-processed to remove duplicates and near-duplicates, as well as short tweets with less than 5 Arabic words and long tweets with more than 80 Arabic words. The tweets cover a variety of topics and themes, reflecting the diverse interests and perspectives of Arabic Twitter users during the specified time frame.

As observed in Mubarak et al. (2017), only 1–2% of Arabic tweets are abusive. The percentage of hate speech is even smaller. To make the annotation more efficient, we applied the following steps to increase the ratio of harmful content (offensive language, hate speech, spam, etc.) using publicly available ASAD tools (Hassan et al., 2021):

- We took random samples from the tweets labeled by ASAD based on the confidence scores for offensiveness. The distribution is as follows: 4,000 tweets from the high-confident offensive tweets (confidence range is 80%–100%), 4,000 tweets from the average confidence scores (confidence range is 60%–79%), and 2,000 tweets from the low-confident scores (confidence range is 39%–1%).
- We took random samples of 4,000 positive tweets, 4,000 negative tweets, and 2,000 neutral tweets.
- We included all the tweets that have hate speech (11,000), all the tweets that have adult content (2,600), and all the spam tweets (210).

It is important to note that there is an overlap between the extracted tweets from the above selection steps, for example, a tweet can be offensive, and have a negative sentiment at the same time.

After merging all the selected tweets, we ended up with 15,965 tweets for our annotation. We opted to create a multi-label dataset and not only focus on offensive discourse and hate speech. We asked the annotators to label the tweets for the 13 categories (or questions) described in **Table 1**.

We developed a clear data collection and annotation pipeline. We started by writing the guidelines to specify the process of annotating the labels. The project was led by an experienced annotation manager.

Annotators and Training: The annotators involved in this study originated from various Arabic-speaking regions across the Middle East and North Africa and possessed a strong Arabic language background. Sixteen annotators contributed to the initial round of evaluating the tweets, which included both text and emojis. Each tweet was independently judged by multiple annotators, ranging from one to five annotators per tweet, based on their ability to understand the dialect in the written tweets.

Training: Selected annotators underwent intensive training over two to three weeks, developing a nuanced and shared understanding of the guidelines, procedures, complex concepts, and Arabic tweets. They completed designed practice questions for each category, studied guidelines and examples, and met regularly to discuss difficulties, strengthen perspectives, and revise as needed. The manager evaluated comprehension and provided retraining when necessary.

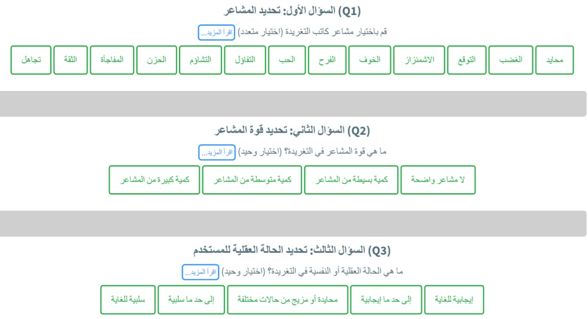
Categories	Description/Values
Q1. Emotions	Choosing from 12 options like anger, anticipation, sadness, etc., or neutral
Q2. Emotion Intensity	No, small, moderate or large amount
Q3. Sentiment	Very positive to very negative or neutral/mixed
Q4. Offensive Language	Whether the tweet contains offensive language and if directed to a target
Q5. Hate Speech Target	Individual, group, or other entity
Q6. Hate Speech Type	Race, religion, ideology, gender, or social class
Q7. Vulgarity	Whether the offensive tweet contains profanity
Q8. Violence	If the offensive tweet promotes violence
Q9. Effect	Whether the tweet is positive (inspiring), negative (depressed), or none
Q10. Sarcasm or Irony	Whether the tweet uses words, often in a humorous way, to mock someone or something
Q11. Humor	Not funny, somewhat funny or very funny
Q12. Factuality	Whether the tweet contains verifiable information and is important
Q13. Spam	Annoying advertising or requests

Table 1: Data annotation categories

Revision: During the revision phase, the manager analyzed errors, unresolved cases, feedback, and updated guidelines to maximize quality, consistency, and consensus in annotation decisions. The manager thoroughly analyzed errors, unresolved cases, feedback, and issues to continuously improve the guidelines, examples, training, and process documentation.

Production: In the production phase, the manager assigned annotators to appropriate tasks and datasets, monitoring quality through flags, reviews, and regular meetings. Annotators typically worked independently but met regularly. The manager controlled quality, handling flags and messages, preprocessing files, and assigning final tasks. We fostered open communication through a message board for questions, comments, issues, and feedback to improve the guidelines over time. Revision also persisted during the production stage.

The annotators performed their annotation via MicroMappers, an online annotation management platform (Figure 1).



السؤال الأول: تحديد المشاعر (Q1)

قم باختيار مشاعر كاتب التغريدة (اختيار متعدد)

السؤال الثاني: تحديد قوة المشاعر (Q2)

ما هي قوة المشاعر في التغريدة؟ (اختيار وحيد)

السؤال الثالث: تحديد الحالة العقلية للمستخدم (Q3)

ما هي الحالة العقلية أو النفسية في التغريدة؟ (اختيار وحيد)

Figure 1: A screenshot from the Arabic version of the annotation interface (Showing the first three questions).

The annotation platform allowed immediate access to annotation guidelines, examples, and procedures. Although working independently, annotators frequently met to discuss difficult cases, resolve disagreements, and strengthen shared perspectives. The annotation manager reviewed and oversaw all work to enforce consistency while enabling independent judgment. Afterward, the annotation manager assigned the tasks to the available annotators as a first human pass.

3.2. Annotation Guidelines

Creating clear guidelines is crucial for annotation projects to ensure data consistency, accuracy, and reliability, as highlighted by Zaghouani et al.. This approach, focusing on Arabic diacritized texts (Zaghouani et al., 2016a) and translation editing (Zaghouani et al., 2016b), emphasizes the need for precise guidelines to produce quality datasets. Such methods not only streamline annotation but also boost the resulting corpora's utility and trustworthiness, contributing significantly to Arabic language resources (Bouamor et al., 2018; Habash et al., 2018; Zaghouani and Charfi, 2018).

We adopted an iterative approach to developing the guidelines, revising them multiple times to ensure consistency and address difficult cases. Some key issues we considered include:

- Defining the categories of emotion detection, emotion intensity, hate speech detection, the effect of the tweet, sarcasm, humor, fact-checking, and spam for Arabic tweets. The definitions were detailed and paired with examples for every choice in the category, including the basic emotions like joy, anger, fear, disgust, surprise, and sadness.

- Addressing issues specific to tweets, such as hashtags, mentions, links, emojis, etc. How do they impact the annotations? The guidelines indicated when and how to consider their meaning.

- Providing examples and non-examples for each annotation category to illustrate the guidelines. Both correct and incorrect annotations of example tweets were included.

- Conducting trial annotations of sample tweets from each category (sentiment, emotion, etc.) to revise the guidelines as needed. Ensuring consistent annotations across samples before finalizing the guidelines.

- Annotators had guidelines that contained the definition and an example for each choice they had to make. For example, for the emotion of disgust, annotators had the following definition and example:
  يشمل ذلك أيضا عدم الاهتمام والكراهية والبغض. مثال:
  انا مريض نفسي وفيه ناس يستفزوني: الاشمئزاز
  Disgust: This also includes indifference, hatred, and loathing. Example: I am mentally ill and there are people who provoke me

Table 2 shows the 13 labels (questions and sub-questions) in our annotation guidelines with the possible options to be selected by the annotators.

Category	**Guidelines**	**Example**
Emotions	The annotators select emotions expressed from 12 options: neutral, anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, trust.	أشعر بالإحباط وخيبة الأمل من هذا الوضع "I'm frustrated and disappointed with this situation." Labels: Anger, Pessimism
Emotion Intensity	The annotators judge the intensity of emotions in the tweet as none, small, moderate, or large.	أنا قلقة بعض الشيء بشأن امتحاني غدا "I'm a little worried about my exam tomorrow." Label: Small amount
Sentiment	The annotators label the overall sentiment of the tweet as very positive, somewhat positive, neutral/mixed, somewhat negative, or very negative.	أنتم أسوأ قناة في الأخبار "You are the worst news channel." Label: Very negative
Offensive Language	The annotators determine if the tweet contains offensive language and if it's directed to a target.	أنت حقير ومثير للاشمئزاز. "You are despicable and loathsome." Labels: Yes, offensive; Yes, directed
Hate Speech Target	The annotators identify if the hate speech target is an individual, a group, or other entities.	هؤلاء الرياضيون قمامة لا قيمة لهم "Those athletes are worthless trash." Label: Group
Hate Speech Type	The annotators specify the group targeted by hate speech based on common characteristics.	كل الرجال حثالة ويجب أن يلقوا في السجن "All men are scum and should be thrown in jail." Label: Gender
Vulgarity	The annotators mark if the offensive tweet contains profanity.	وأنت مال أمك يا إبن الوسخة "It's none of your business, you son of a bitch." Label: Yes, profanity
Violence	The annotators label if the offensive tweet promotes violence.	أيها الرئيس، اقتل كل المعارضين "Boss, kill all the opponents." Label: Yes, violence

Effect	The annotators judge if the tweet has a positive/inspiring or negative effect.	ابقي رأسك مرفوعا. المستقبل مشرق "Keep your head up. The future is bright." Label: Positive effect
Sarcasm or Irony	The annotators identify if the tweet contains sarcasm or irony directed to a target.	اتفضل قول يا أذكى إخواتك! "You're so incredibly smart and talented!" Label: Yes sarcasm
Humor	The annotators label if the tweet is not funny, somewhat funny, or very funny.	اوعى تتخلى عن حلمك خليك نايم لحد ما تكمله "Don't give up on your dream, stay asleep until you complete it" Label: Very funny
Factuality	The annotators label if the tweet contains verifiable information and if it's important.	سيتم إعلان نتائج الانتخابات اليوم "The election results will be announced today." Labels: Verifiable information, Important
Spam	The annotators label if the tweet contains spam (advertising or requests).	بحاجة الى المال على وجه السرعة. الرجاء المساعدة من خلال التبرع على هذا الموقع "Need money urgently. Please help by donating on this site." Label: Spam

Table 2: Annotation Guidelines

Item	**N**	**%**
Offensive language?		
Yes	2793	17.5%
No	13171	82.5%
Directed?		
Yes, directed	2348	84.7%
No, not directed	445	15.3%
Target		
Individual	963	41.0%
People with common features	1090	46.4%
Organization, company, situation, or topic	299	13.9%
What do they have in common?		
Ideology, politics, sports	747	68.5%
Class, social status, or profession	83	7.6%
Religion or sect	144	13.3%
Gender	43	3.9%
Origin, race, or nationality	73	3.5%
Obscene language?		
Yes	874	31.3%
No	1919	69.7%
Advocates for violence?		
Yes	201	7.1%
No	2592	92.9%

Table 3: Offensive Language Results from Manual Annotation

3.3. Annotation Analysis

Table 3 below lists the frequency distribution of the offensive annotation with the various labels annotated. The results of the manual annotation of offensive language in the dataset reveal some interesting insights. Offensive language was present in a sizable minority (17.5%) of the samples, indicating that offensive content is a relevant issue to study in this dataset. Of the offensive examples, the vast majority (84.7%) were directed at a target rather than being general offensive expressions.

This suggests that when offensive language appears in this dataset, it is generally used to target specific individuals or groups rather than simply for shock value.

The most common targets of offensive speech were people with common features (46.4%) and individuals (41.0%). This indicates that offensive expressions in this dataset tend to focus on identity-based attributes like ideology, gender, race, etc. Only 13.9% targeted organizations or abstract concepts. Understanding these patterns of targeting can help guide strategies for mitigating harmful speech. Among targeted offensive expressions, the most common attributes attacked were ideology and politics (68.5%), followed distantly by origin/race/nationality (13.3%). This aligns with broader findings that political and racial discussions tend to attract higher levels of offensiveness online. Gender-based targeting was comparatively rare (3.9%) in this dataset.

Additionally, while a moderate portion (31.3%) of offensive examples contained obscene language, only 7.1% advocated violence. This suggests that overt threats are less common than cursing or slurs when expressing offense in this dataset.

The results in Table 4 provide insight into the perceived impact of tweets in the hate speech dataset on the well-being and sentiments of readers/annotators. Out of the total tweets annotated, 26.54% were rated as frustrating by annotators. This indicates that over a quarter of the tweets had some level of negative effect on the annotators' state of mind or emotions. In contrast, 15.51% of tweets were deemed motivating, suggesting a more positive emotional response from annotators. However, the majority of tweets (56.93%) were rated as neither frustrating nor motivating by annotators. This implies that more than half of the content did not evoke a strong sentimental reaction either way. Annotators likely considered most tweets neutral in tone and impact on their personal well-being. Still, the sizable frustrating subgroup points to the presence of harmful language that could detrimentally affect readers' mental health if exposed over time. In summary, while a majority of tweets were effectively neutral, a significant portion contained potentially frustrating content according to annotator judgments. This highlights the need for hate speech detection systems that can identify not just overtly aggressive language, but also subtle linguistic patterns that may harmfully impact users' psychological well-being. More analysis of the specific frustrating tweets could reveal distinguishing attributes to improve classification performance.

Effect of the Tweet on the Reader's Wellbeing/Annotator's Sentiment	N	%
Frustrating	4237	26.54
Motivating	2637	15.51
Neither frustrating nor motivating	9089	56.93

Table 4: Sentiment Results from Manual Annotation

The fact-checking annotations in Table 5 reveal that most tweets (63.83%) contained no verifiable factual information. Only 17.45% were found to have checkable information, though a majority of those were still deemed unimportant to the public. This suggests hate speech tweets rarely make the truth claims, but when they do, the claims are often subjective or lack significance.

Fact-Checking	N	%
No information	10190	63.83%
Contains information, but not verifiable	2987	18.71%
Contains information that is verifiable	2787	17.45%
Important to the public?		
Yes, important	1594	57.19%
No, not important	981	35.2%

Table 5: Fact-Checking Information

The categorized annotations in Table 7 provide additional context. A majority of tweets contained some amount of discernible emotion intensity, indicating hate speech frequently carries strong sentiment. Additionally, over 90% were labeled for sarcasm/irony and humor, implying attempts at subtle rhetorical attacks rather than overt hostility. Specific product/service spam was rarely detected, and valence was distributed across positive, negative and neutral.

Data Annotation Categories	N
Emotion (Pessimism, sadness, confidence, joy and others)	12301
Emotion Intensity (Small, large, average amount of feelings and others)	9075
Sarcasm, irony, and ridicule (yes, no)	9036
Humor, joking (Yes, but not funny and others)	9257
Spam Detection (Yes, specific product or service, and others)	9004
Valence (positive, negative. Neutral, and others)	9059

Table 6: Data Annotation Categories

These annotation results in both tables demonstrate the complex nature of hate speech across subjective claims, varied emotions, and figurative language use. Most tweets did not express verifiable facts or commercial spam. The prevalence of sarcasm and humorous intent highlights the need for nuanced linguistic analysis. Even tweets with positive valence may carry implicitly harmful messages. Effective detection systems must leverage these annotation types to identify harmful speech in its many manifestations.

3.4. Annotation Evaluation

To measure the quality of the annotation, we calculated the Inter-Annotator Agreement (IAA) using Cohen's Kappa. The results of the Kappa calculations for each label are presented in Table 7. The IAA was calculated on 500 tweets by performing pairwise comparisons between all combinations of annotators and averaging the results.

The inter-annotator agreement scores provide an important quality indicator for the multi-label tweet annotation process. As shown in Table 7, the overall average IAA based on Cohen's Kappa is a substantial 0.75, suggesting annotators largely agreed in their labeling decisions. Several categories like offensive content, hate speech type, hate target, vulgarity, factuality, and spam achieved strong agreement above 0.85 Kappa. This indicates clear, consistent guidelines and shared understanding for identifying offensive tweets, specifying

Label	Cohen's Kappa
Emotions	0.4396
Emotion intensity	0.5632
Sentiment	0.9289
Offensive content	0.8863
Hate speech type	0.7664
Hate speech target	0.8972
Vulgarity	0.9024
Violence	0.7304
Effect	0.4896
Sarcasm/irony	0.6377
Humor	0.7010
Factuality	0.8545
Spam	0.9499
Overall	0.7497

Table 7: Inter-Annotator Agreement for Data Annotation Categories

hate speech details, and labeling spam.

Moderately strong agreement around 0.60-0.75 was reached for emotion intensity, sarcasm/irony, and humor detection. The subjective nature of judging emotions and rhetorical devices like humor likely contributed to more variability. Nonetheless, the scores signify satisfactory convergence.

Categories like emotions, sentiment, and effect saw lower agreement in the 0.4-0.5 Kappa range. Discrepancies likely arise from the inherent difficulty in disambiguating subtly different emotions based on limited tweet context, as well as identifying mixed sentiment. The lower effect score suggests possible interpreter variation in judging a tweet's impact as positive or negative.

While some subjective categories naturally displayed greater variability between annotators, the IAA results overall confirm effective training, guidelines, and quality control to ensure consistent dataset annotations. Higher agreement for objective attributes like spam indicates clearer instruction. Lower agreement prompts refinement of murkier categories. However, averaging 0.75 agreement despite the linguistic complexity highlights the rigor of the annotation methodology and training process.

In Table 8 , we provide some cases of annotation disagreement.

3.5. Data Annotation through Machine and Deep Learning

In this section, we analyzed the applicability of our annotated dataset for detecting hate speech and offensive language through an automated approach. We examined two approaches: traditional machine learning and deep learning. Our aim was to develop and evaluate a model that can accurately identify hate speech in Arabic tweets. By focusing on a monolingual model, we aimed to tailor the model to handle the specific challenges of Arabic text processing, such as dialectal variation, code-switching, and the use of non-standard orthography. While multilingual models leverage cross-lingual similarities, they can sometimes be less effective for languages like Arabic, where understanding context, idiomatic expressions, and cultural references is crucial for hate speech detection. The process involved several steps as follows.

Example tweet	Annotator Disagreement
الله يبارك فيك حبيبي ي احمد عقبال عندك يارب God bless you [in response to "congratulations"], my dear Ahmed. I wish you the same	One annotator opted for a "neutral" emotional classification, and the two others selected "joy optimism confidence" and "optimism confidence".
اذا طلعنا من حياتكم طلعونا من سوالفكم If we get out of your life, get us out of your conversations/problems	For detecting the sentiment, the three annotators chose "neutral or combination", "very negativ", and "somewhat negative".
العدوان يحاول بكل ما لديه من قوه اعاده القاعده وداعش الى المناطق التي تم طردهم منها The aggression is trying with all its might to return Al-Qaeda and ISIS to the areas from which they were expelled.	The annotators also made three different judgments in this case: "information-verifiable", "no information", and "information-but-not-verifiable".
العميل من يسرق بلده ويسلم مقدراته لبلد آخر لايفعلها الا ذيل او مرتزق او خائن The client is the one who steals from his country and hands over his reins to another country. This can only be done by a tail, a mercenary, or a traitor.	In the section where annotators were asked if the tweet contained offensive or vulgar language, two of them selected "no" while one selected "yes".

Table 8: Sample Cases of Discrepancy Among Annotators

Data Preprocessing: We preprocessed the text by removing unwanted characters, English words, and punctuation. Then, we vectorized the text to a matrix of features using CountVectorizer. We chose two columns: ’Offensive’ and ’HateSpeech’ for the experiment.

Dataset Split: The dataset is unbalanced. There is a significantly larger number of ’No’ labels than ’Yes’ labels for both offensive and hate speech labels. For offensive labels, ’No’ accounts for 82.5%, and for hate speech labels, ’No’ accounts for 93.1%. The dataset was split between 60% for training and 40% for testing.

Machine Learning Approach: We applied Logistic Regression, Support Vector Machine (SVM), Random Forest, Decision Tree, and Gradient Boosting to predict the label between yes and no. Then, we measured the accuracy, precision, recall, and F1-score for each test.

Deep Learning Approach: We applied AraBERT, a Transformer-based model from Antoun et al., for the deep learning approach to predict offensive and hate speech labels. AraBERT is a transformer-based model tailored for Arabic language understanding.

For offensive language detection, Logistic Regression excelled in non-offensive precision (87%) but had a 34% recall for offensive content. Random Forest and Gradient Boost had precision values of 70% and 81% for offensive remarks but struggled with recall. SVM showed moderate precision with 40% recall for offensive comments. Impressively, AraBERT had 100% precision for non-offensive and 49% for offensive content. Table 9 below details these findings.

Model	Label	Precision	Recall	F1-Score	Accuracy
LR	Yes No	0.58 0.87	0.34 0.95	0.43 0.91	0.84
RF	Yes No	0.70 0.86	0.25 0.98	0.37 0.91	0.85
GB	Yes No	0.81 0.84	0.14 0.99	0.23 0.91	0.84
SVM	Yes No	0.49 0.87	0.40 0.91	0.44 0.89	0.82
DT	Yes No	0.42 0.87	0.38 0.88	0.40 0.88	0.79
Arabert	All	0.49	0.99	0.65	0.82

Table 9: Offensive Language Detection

For hate speech detection, models varied in efficacy. Logistic Regression had a 38% precision for hate speech but 94% for non-hate content, with a 14recall rate. Random Forest and Gradient Boost had mid-40s precision with low recall. SVM and Decision Tree had precisions of 27% and 22% respectively. Notably, Arabert showed 57% precision and 66recall for hate speech, outperforming others. Table 10 presents these findings.

Model	Label	Precision	Recall	F1-Score	Accuracy
LR	Yes No	0.38 0.94	0.14 0.98	0.20 0.96	0.93
RF	Yes No	0.47 0.93	0.05 0.99	0.09 0.96	0.93
GB	Yes No	0.41 0.93	0.04 0.99	0.07 0.96	0.93
SVM	Yes No	0.27 0.94	0.24 0.95	0.25 0.95	0.9
DT	Yes No	0.22 0.94	0.19 0.95	0.21 0.95	0.9
Arabert	All	0.57	0.47	0.66	0.83

Table 10: HateSpeech Language Detection

4. Limitation

This study on Arabic hate speech on Twitter provides valuable insights but has limitations: (1) Annotator bias from specific Arabic regions may affect labeling accuracy. (2) Subjectivity in identifying hate speech can impact annotation reliability. (3) Findings might not apply to underrepresented Arabic dialects. (4) Using only Twitter data limits the study’s applicability to other platforms. (5) Detecting implicit hate speech, such as sarcasm, is challenging. (6) Sample selection for model testing could lead to biased performance evaluations. (7) Class imbalance may skew model metrics. (8) Accessing Twitter data now incurs a cost for academic researchers.

5. Conclusions

In this research, we introduce a multi-label, fine-grained Arabic hate speech dataset, the most comprehensive of its kind, annotated across nine dimensions such as sentiment and emotion, enhancing its utility for various studies. We’ve meticulously documented the dataset’s development process, from collection to annotation, providing a detailed roadmap for future replication and research. Our in-depth analysis of the dataset offers novel insights into Arabic hate speech, and our experiments with multiple classification techniques contribute valuable perspectives to the field. This dataset not only advances our understanding of Arabic linguistic patterns but also serves as a valuable asset for researchers and practitioners in Arabic language processing.

6. Data Availability Statement

The guidelines and the SoHateful 1.0 annotated dataset can be obtained by contacting the authors to facilitate future research and reproducibility. The users of the dataset must adhere to the terms and conditions outlined in the repository. To request the dataset for research purposes, please fill the following form: https://forms.gle/S9fZtYjAyLAqFsH19

The dataset is released under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, allowing for its free use, distribution, and adaptation, provided the original work is properly credited.

7. Code Availability Statement

The code for preprocessing the Arabic tweets, feature extraction, and training the machine learning models are available at https://github.com/rafiulbiswas/hatespeech-detection.

Usage of this code must abide by the licensing terms documented in the repository. The repository contains Python notebooks detailing the step-by-step implementation of the natural language processing pipeline and experiments presented in this study. Researchers can use these resources to replicate the approach on new datasets.

8. Acknowledgments

This publication was made possible through the financial support of NPRP grants NPRP13S-0206-200281 from the Qatar National Research Fund, now known as the Qatar Research Development and Innovation Council (QRDI). The findings and opinions expressed herein are solely the responsibility of the authors.

We express our gratitude to the voluntary efforts of the annotators who contributed to this annotation task. A total of 16 annotators from various Arabic-speaking countries participated, comprising both male and female individuals. Their diverse backgrounds and perspectives enriched the annotation process and ensured a comprehensive understanding of the nuances in the Arabic language and cultural context.

9. References

Ines Abbes, Wajdi Zaghouani, Omaima El-Hardlo, and Faten Ashour. 2020. Daict: A dialectal arabic irony corpus extracted from twitter. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6265–6271.

Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. 2016. Farasa: A fast and furious segmenter for arabic. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Demonstrations*, pages 11–16.

Arfan Ahmed, Nashva Ali, Mahmood Alzubaidi, Wajdi Zaghouani, Alaa A Abd-alrazaq, and Mowafa Househ. 2022. Freely available arabic corpora: A scoping review. *Computer Methods and Programs in Biomedicine Update*, 2:100049.

Areej Al-Hassan and Hmood Al-Dossari. 2019. Detection of hate speech in social networks: a survey on multilingual corpus. In *6th international conference on computer science and information technology*, volume 10, pages 10–5121.

Nuha Albadi, Maram Kurdi, and Shivakant Mishra. 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)*, pages 69–76. IEEE.

Ibrahim Aljarah, Maria Habib, Neveen Hijazi, Hossam Faris, Raneem Qaddoura, Bassam Hammo, Mohammad Abushariah, and Mohammad Alfawareh. 2021. Intelligent detection of hate speech in arabic social network: A machine learning approach. *Journal of Information Science*, 47(4):483–501.

Malik Almaliki, Abdulqader M Almars, Ibrahim Gad, and El-Sayed Atlam. 2023. Abmm: Arabic bert-mini model for hate-speech detection on social media. *Electronics*, 12(4):1048.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Arunkumar Bagavathi, Pedram Bashiri, Shannon Reid, Matthew Phillips, and Siddharth Krishnan. 2019. Examining untempered social media: analyzing cascades of polarized conversations. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 625–632.

Mohamed Aziz Bennessir, Malek Rhouma, Hatem Haddad, and Chayma Fourati. 2022. icompass at arabic hate speech 2022: Detect hate speech using qrnn and transformers. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 176–180.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, et al. 2018. The madar arabic dialect corpus and lexicon. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Anis Charfi, Wajdi Zaghouani, Syed Hassan Mehdi, and Esraa Mohamed. 2019. A fine-grained annotated multi-dialectal arabic corpus. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 198–204.

Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Mean birds: Detecting aggression and bullying on twitter. In *Proceedings of the 2017 ACM on web science conference*, pages 13–22.

Nizar Habash, Fadhl Eryani, Salam Khalifa, Owen Rambow, Dana Abdulrahim, Alexander Erdmann, Reem Faraj, Wajdi Zaghouani, Houda Bouamor, Nasser Zalmout, et al. 2018. Unified guidelines and resources for arabic dialect orthography. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Shatha Ali A Hakami, Robert J Hendley, and Phillip Smith. 2022. A context-free arabic emoji sentiment lexicon (cf-arab-esl). In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 51–59.

Sabit Hassan, Hamdy Mubarak, Ahmed Abdelali, and Kareem Darwish. 2021. Asad: Arabic social media analytics and understanding. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 113–118.

Sanaa Kaddoura, Suja A Alex, Maher Itani, Safaa Henno, Asma AlNashash, and D Jude Hemanth. 2023. Arabic spam tweets classification using deep learning. *Neural Computing and Applications*, pages 1–14.

Natasha Lomas. 2017. Facebook, google, twitter commit to hate speech action in germany. *Last accessed: July*.

Sarah Masud, Pinkesh Pinkesh, Amitava Das, Manish Gupta, Preslav Nakov, and Tanmoy Chakraborty. 2022. Half-day tutorial on combating online hate speech: The role of content, networks, psychology, user behavior, etc. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 1629–1631.

Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019. Spread of hate speech in online social media. In *Proceedings of the 10th ACM conference on web science*, pages 173–182.

Hanane Mohaouchane, Asmaa Mourhir, and Nikola S Nikolov. 2019. Detecting offensive language on arabic social media using deep learning. In *2019 sixth international conference on social networks analysis, management and security (SNAMS)*, pages 466–471. IEEE.

Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive language detection on arabic social media. In *Proceedings of the first workshop on abusive language online*, pages 52–56.

Hamdy Mubarak, Sabit Hassan, and Shammur Absar Chowdhury. 2023. Emojis as anchors to detect arabic offensive language and hate speech. *Natural Language Engineering*, 29(6):1436–1457.

Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. 2019. L-hsab: A levantine twitter dataset for hate speech and abusive language. In *Proceedings of the third workshop on abusive language online*, pages 111–118.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. *arXiv preprint arXiv:1908.11049*.

Juan Carlos Pereira-Kohatsu, Lara Quijano-Sánchez, Federico Liberatore, and Miguel Camacho-Collados. 2019. Detecting and monitoring hate speech in twitter. *Sensors*, 19(21):4654.

Jing Qian, Mai ElSherief, Elizabeth Belding, and William Yang Wang. 2018. Hierarchical cvae for fine-grained hate speech classification. *arXiv preprint arXiv:1809.00088*.

Francisco Rangel, Paolo Rosso, Wajdi Zaghouani, and Anis Charfi. 2020. Fine-grained analysis of language varieties and demographics. *Natural Language Engineering*, 26(6):641–661.

Paolo Rosso, Francisco Rangel, Irazu Hernández Farías, Leticia Cagnina, Wajdi Zaghouani, and Anis Charfi. 2018. A survey on author profiling, deception, and irony detection for the arabic language. *Language and Linguistics Compass*, 12(4):e12275.

Ahmad Shapiro, Ayman Khalafallah, and Marwan Torki. 2022. Alexu-aic at arabic hate speech 2022: Contrast to classify. *arXiv preprint arXiv:2207.08557*.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*.

Zeerak Waseem and Dirk Hovy. 2016. Understanding abuse: A typology of abusive language detection subtasks.? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Hajime Watanabe, Mondher Bouazizi, and Tomoaki Ohtsuki. 2018. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE access*, 6:13825–13835.

Wajdi Zaghouani. 2014. Critical survey of the freely available arabic corpora. In *International Conference on Language Resources and Evaluation (LREC'2014), OSACT Workshop. Reykjavik, Iceland, 26-31 May 2014*.

Wajdi Zaghouani, Houda Bouamor, Abdelati Hawwari, Mona Diab, Ossama Obeid, Mahmoud Ghoneim, Sawsan Alqahtani, and Kemal Oflazer. 2016a. Guidelines and framework for a large scale arabic diacritized corpus. In *The Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3637–3643. European Language Resources Association (ELRA).

Wajdi Zaghouani and Anis Charfi. 2018. Guidelines and annotation framework for arabic author profiling. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Wajdi Zaghouani, Nizar Habash, Ossama Obeid, Behrang Mohit, Houda Bouamor, and Kemal Oflazer. 2016b. Building an arabic machine translation post-edited corpus: Guidelines and annotation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1869–1876.

Wajdi Zaghouani, Behrang Mohit, Nizar Habash, Ossama Obeid, Nadi Tomeh, Alla Rozovskaya, Noura Farra, Sarah Alkuhlani, and Kemal Oflazer. 2014. Large scale arabic error annotation: Guidelines and framework.