Fine-Grained Arabic Offensive Language Classification with Taxonomy, Sentiment, and Emotions

Natalia Vanetik [1] natalyav@sce.ac.il
Marina Litvak [1] marinal@sce.ac.il
Chaya Liebeskind [2] liebchaya@gmail.com
[1]Shamoon College of Engineering, Beer-Sheva, Israel
[2]Jerusalem College of Technology, Jerusalem, Israel

Abstract

Offensive language detection in Arabic is a challenging task because of the unique linguistic and cultural characteristics of the Arabic language. This study introduces a high-quality annotated dataset for classifying offensive language in Arabic, based on a structured taxonomy, categorizing offensive content across seven levels, capturing both explicit and implicit expressions. Utilizing this taxonomy, we re-annotate the FARAD-500 dataset, creating reFarad-500, which provides fine-grained labels for offensive texts in Arabic. A thorough dataset analysis reveals key patterns in offensive language distribution, emphasizing the importance of target type, offense severity, and linguistic structures. Additionally, we assess text classification techniques to evaluate the dataset's effectiveness, exploring the impact of sentiment analysis and emotion detection on classification performance. Our findings highlight the complexity of Arabic offensive language and underscore the necessity of extensive annotation frameworks for accurate detection. This paper advances Arabic natural language processing (NLP) in resource-constrained settings by enhancing the recognition of hate speech and fostering a deeper understanding of the linguistic and emotional dimensions of offensive language.

1 Introduction

Arabic offensive language detection is a crucial but difficult undertaking that necessitates a thorough comprehension of both linguistic structures and cultural context. Direct insults and hate speech are examples of explicit offensive content. Implicit offensive content necessitates a more thorough contextual study to determine its intent. Even though Arabic language processing has improved due to recent developments in natural language processing (NLP), current categorization frameworks frequently lack the granularity required to effectively capture objectionable statements in various Arabic dialects.

We present a high-quality data set that is annotated in accordance with the ArSOL taxonomy (Liebeskind et al., 2024), but unlike the original ArSOL work, we reapply it with stricter guidelines and expanded multi-label capability, to overcome these issues. It comprises seven hierarchical levels, distinguishing between explicit and implicit offenses, and categorizing offensive content based on target presence, vulgarity, offense severity, and type.

As part of this study, we start from the FARAD-500 dataset, which aggregates Arabic offensive language texts from multiple sources, but our work departs from it by systematically correcting annotation inconsistencies. The original FARAD-500 dataset provides valuable offensive language examples, but several issues limit its utility: inconsistent application of labels, overuse of vague categories, and lack of multi-label annotations for complex instances. Our re-annotation aims to address these issues by applying the ArSOL taxonomy rigorously and by instructing annotators to distinguish between overlapping categories when relevant. This effort resulted in reFarad-500, a dataset that enhances classification precision across different offensive categories. We analyze the dataset using various NLP techniques, including sentiment analysis and emotion detection, to explore their role in improving offensive language classification.

Through an extensive evaluation pipeline, we assess the quality and utility of the dataset by training text classification models on it and evaluating their performance using standard metrics. We analyze results across multiple annotation levels to examine how the structure of the annotation scheme impacts classification performance. Our results demonstrate the advantages of a structured annotation approach and offer important insights into the

Proceedings of the Workshop on Beyond English: NLP for all Languages in an Era of LLM from Texts associated with RANLP 2025, pages 110–119, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-105-9-013

patterns of offensive language in Arabic.

We believe that incorporating sentiment analysis and emotion detection can provide additional information about the speaker's emotional context, even though the majority of previous work has been on explicitly recognizing offensive words. If one is aware of the emotions that accompany offensive remarks, it would be possible to categorize different levels of offensiveness more accurately (Mnassri et al., 2023b).

By concentrating on Arabic, this work helps close the gap in natural language processing for medium- and low-resource languages. Despite being extensively spoken, Arabic, a morphologically rich language, is nevertheless underrepresented in high-quality annotated datasets for offensive material. Our method tackles important issues like the lack of data, consistent annotations, and the intricate relationship between social context and linguistic structure. We offer a fine-grained, reusable resource and experimental approach that is applicable to other languages with comparable limitations by re-annotating an existing dataset and incorporating sentiment and emotion features.

2 Related Work

Various taxonomies classify offensive language at different levels. The term "offensive language" has been defined in diverse ways in prior research; in this paper, we adopt a broad definition stating that offensive language is any form of communication that intentionally or unintentionally conveys hostility, disrespect, or harm toward individuals or groups.

Works of Zampieri et al. (2019a,b) classify content as offensive or not, then as targeted insults or threats, and finally by target type (groups, individuals, etc.). The Nexus Linguarum Working Group (Lewandowska-Tomaszczyk et al., 2021) defined offensive and non-offensive language, targeted and non-targeted insults, and explicit versus implicit language with two primary levels and four sub-levels. Lewandowska-Tomaszczyk et al. (2022) tested a schema for explicit and implicit language and proposed a simplified, unified approach with direct and implied offensiveness in (Lewandowska-Tomaszczyk, 2023). The authors demonstrate that the SOL taxonomy helps identify offensive language in English by showing that its categories align with semantic patterns in word embeddings and yield consistent annotations with high inter-annotator agreement. Liebeskind et al. (2023) have shown that this taxonomy can be successfully applied to Hebrew.

Despite their differences in granularity and structure, these taxonomies all aim to formalize the idea of offensive language. The definition of offensive content is still debatable and complex, though. Related concepts including hate speech, toxicity, abusive language, and incivility have been used in earlier research; meanings range from overtly hostile utterances to more subdued expressions like sarcasm, stereotyping, or exclusionary discourse. In this work, we adopt a more expansive conceptualization that acknowledges sociolinguistic variation in the expression and perception of offense, particularly in morphologically rich and culturally diverse languages like Arabic, and that takes into account both explicit and implicit forms of offense.

To formalize this view, we rely on a structured taxonomy introduced in (Liebeskind et al., 2024) provides a comprehensive framework for categorizing Arabic offensive language. To simplify addressing it in the paper, we denote it by the ArSOL taxonomy. This seven-level taxonomy refines and extends the framework proposed by Lewandowska-Tomaszczyk et al. (2023) and builds on Zampieri et al. (2019a,b) to capture both explicit and implicit offensive language. The taxonomy categorizes offensive language into seven levels: Levels 1 to 6 focus on explicit categories, while Level 7 addresses implicit language. In this study, we focused on Levels 1–6 due to data limitations. Level 7 will be addressed in future extensions. Figure 1 depicts levels 1-6 of ArSOL with English translations.

Multiple datasets for offensive language detection in Arabic have been introduced, reflecting the linguistic and cultural diversity of Arab-speaking regions. Early datasets focused on specific hate speech types: for example, Albadi et al. (2018) contains texts targeting religious prejudice, while Aref et al. (2020) created a dataset focused on religious hate speech concerning the Sunni-Shia divide. Expanding thematic scope, Mulki and Ghanem (2021) developed the Let-Mi dataset, which provides versatile examples of misogynistic behavior.

Other datasets use multidimensional annotation frameworks to capture complex phenomena. Ousidhoum et al. (2019) presented a multilingual dataset annotated for hostility, directness, and target attributes such as religion or sexual orientation. Similarly, Ahmad et al. (2024) released a multi-class

dataset of tweets categorized into four sentiment-based hate speech classes.

Researchers have also explored platform-specific data sources, including YouTube comments (Alakrot et al., 2018) and news articles (Chowdhury et al., 2020; Mubarak et al., 2017). Furthermore, several studies address offensive language in different Arabic dialects (Mulki et al., 2019; Haddad et al., 2019; Mubarak et al., 2020; Litvak et al., 2021; Essefar et al., 2023; Alhazmi, 2023). The FARAD-500 dataset, proposed by Liebeskind et al. (2024), focuses on Modern Standard Arabic (MSA) and Levantine dialects and contains 500 offensive texts annotated according to the ArSOL taxonomy.

Our work complements and extends prior efforts by re-annotating the FARAD-500 dataset to improve annotation accuracy and balance across offense categories. The refined dataset, reFarad-500, ensures a more balanced representation of offensive language types, facilitating improved model training and evaluation. We also evaluate the effectiveness of the ArSOL taxonomy by training text classification models on the refined dataset. In addition, we investigate how sentiment analysis and emotion detection can assist offensive language detection. Lastly, the re-annotated dataset is made publicly available to support further advances in Arabic NLP and offensive language identification. Although extensive research exists on Arabic offensive language detection, few studies explore integrating sentiment and emotion analysis to enhance classification. We investigated their role in enhancing offensive language classification since previous work (Plaza-del Arco et al., 2021; Mnassri et al., 2023a; Samghabadi et al., 2020; Elmadany et al., 2020; Althobaiti, 2022) shows the advantages of combining these approaches.

3 The reFarad-500 Dataset

3.1 Data Preprocessing

We used the FARAD-500 dataset of (Liebeskind et al., 2024) as a starting point, but our work substantially revises and extends it. FARAD-500 was generated from 16 existing Arabic offensive language datasets, ensuring alignment with ArSOL taxonomy. Table 1 lists the datasets and specifies taxonomy levels for which the data is originally annotated (cases, where not all options of a taxonomy level are used, are marked with an asterisk). Most datasets are annotated at level 1 (offensive or not) and partially at level 5 (offense strength), primarily focusing on hate speech. Other taxonomy levels, such as target presence and offense aspects, lack annotation. Furthermore, because of style variations, the dataset only contains texts from Facebook and Twitter, leaving out sources like YouTube and news articles. However, FARAD-500's partial and sometimes inaccurate taxonomy coverage limits its suitability for fine-grained ArSOL-based classification.

Paper	Source	Levels
(Albadi et al., 2018)	Twitter	5*
(Ousidhoum et al., 2019)	Twitter	1, 2, 3, 5*
(Mulki et al., 2019)	Twitter	1, 5*
(Zampieri et al., 2020)	Twitter	1, 4, 5*
(Mubarak et al., 2017)	Twitter	1, 5*
(Aref et al., 2020)	Twitter	5*
(Ahmad et al., 2024)	Twitter	1, 5*
(Mulki and Ghanem, 2021)	Twitter	1, 5*, 6*
(Litvak et al., 2021)	Twitter	1
(Alhazmi, 2023)	Twitter	1

Table 1: Data sources of FARAD-500 (* indicates partial annotation).

3.2 Re-annotated dataset reFarad-500

The original FARAD-500 annotations contain partial category coverage and misclassifications, reducing consistency with the ArSOL taxonomy (Liebeskind et al., 2024). We therefore re-annotated the dataset using explicit criteria: correcting label misapplications, ensuring full category coverage, and allowing multi-label assignment when multiple aspects occur in a text. In multiple cases, texts that clearly met the criteria for some of the labels were either misclassified or left unlabeled. To enhance classification accuracy and consistency, this dataset underwent a meticulous re-annotation process to address annotation errors and ambiguities with the help of native Arabic speakers fluent in Modern Standard Arabic (MSA) and Levantine Arabic. The annotators were provided with comprehensive instructions and examples. The main errors we strive to fix were a lack of attention to several aspects expressed in one text, and the use of Other aspect in an erroneous way when other aspects are present in the text. We denote the resulting dataset by reFarad-500. We guided the annotators to mark the aspect as Other only if no other aspect is applicable. Additionally, at level 5, we instructed the annotators to mark each text as either Hate speech or Insult, and then annotate it separately as Threat and as Discredit if necessary to allow for multi-label annotation. We did it to capture the fact that a single offensive text can simultaneously serve multiple

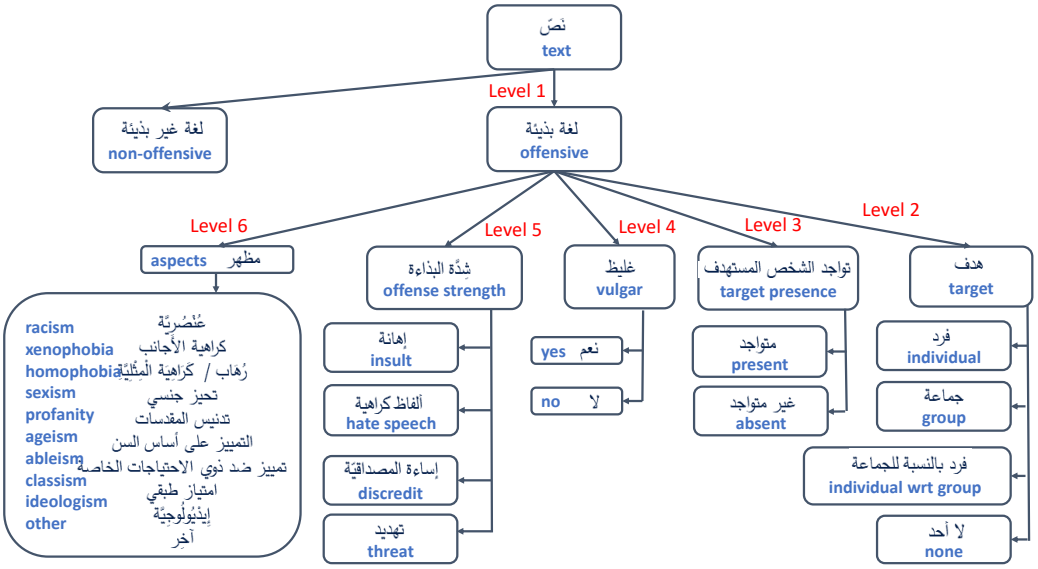

Figure 1: ArSOL taxonomy. Levels 1 through 6 include classifications such as non-offensive vs. offensive (Level 1), the target of the offense (Level 2), the presence or absence of the target (Level 3), vulgar vs. non-vulgar language (Level 4), the severity of the offense (Level 5), and specific types of explicit offenses (Level 6).

functions, such as insulting an individual while also discrediting a group or issuing a threat.

The re-annotation process involved three native Arabic speakers with academic backgrounds in linguistics and NLP. We measured annotator agreement throughout the process using Cohen's Kappa on the full dataset (Table 2). We report Cohen's Kappa coefficient (Cohen, 1960) on the entire dataset of 500 texts. Cohen's Kappa is a statistical measure used to assess the degree of agreement between two raters or annotators when dealing with categorical or ordinal data. The highest level of agreement was found in the Ideologism aspect of level 6, indicating consistency in identifying religiously offensive content. Good to excellent agreement was also noted for Sexism, Xenophobia, and Religion. However, agreement levels were lower for some taxonomy aspects, such as Homophobia and Other, implying difficulties in discovering the intent behind the offense in these specific cases. Still, these agreement values are superior to those reported in (Liebeskind et al., 2024), thus justifying the need for re-annotation. In Table 3 we show the agreement between original and new annotations for levels present in both datasets. We see that the aspects most affected by re-annotation are Homophobia", Ableism, and Other, while no changes were made for target presence and vulgarity levels of the taxonomy. Two examples of texts where the annotation was modified appear in Table 4.

level	name	kappa
2	target	0.931
3	target presence	0.775
4	vulgar	0.881
5	offense strength (avg)	0.926
6	racism	0.725
6	xenophobia	0.714
6	homophobia	0.203
6	sexism	0.798
6	other	0.251
6	ageism	1.000
6	ableism	0.127
6	classism	0.322
6	ideologism	0.824
6	religion	0.755

Table 2: Annotation agreement for the reFarad-500 dataset.

level name	kappa
target	0.996
target presence	1.000
vulgar	1.000
racism	0.895
xenophobia	0.846
homophobia	0.203
sexism	0.817
other	0.259
ageism	1.000
ableism	0.160
classism	0.406
ideologism	0.863
religion	0.788

Table 3: Kappa agreement values for FARAD-500 and reFarad-500 datasets.

text	translation	aspect
نتن ريزنخ اذهو يفو بلكلا سب كناسل حص	Well said, but the dog is loyal, and this one is a filthy pig.	ableism
..ثويد هجوز زيط قشيد لاخ	An uncle who f**s his wife's a**—he's a cuckold.	homophobia

Table 4: Examples of re-annotated texts.

4 Dataset Analysis

4.1 Sentiment Analysis

Sentiment analysis (SA), which helps find sentiment patterns that might be associated with offensive expressions, is an essential tool for text analysis, especially when it comes to identifying offensive language. In offensive language detection, SA can provide additional context by distinguishing between neutral, aggressive, and harmful content, offering a better understanding of intent. For this purpose, we considered two state-of-the-art Arabic pre-trained transformer models: camelBERT (Inoue et al., 2021) and araBERT (Antoun et al., 2020). camelBERT is a state-of-the-art transformer-based model for Arabic that was pretrained on diverse Arabic corpora across dialects and Modern Standard Arabic (MSA), and fine-tuned for various downstream tasks, including sentiment analysis, achieving high accuracy on benchmark datasets ASTD (Nabil et al., 2015) and LABR (Aly and Atiya, 2013). araBERT, similarly, was pretrained on over 200 million Arabic sentences and fine-tuned on multiple sentiment analysis benchmarks, consistently outperforming earlier models and establishing itself as a strong baseline for Arabic NLP tasks.

We applied both camelBERT and araBERT models to the reFarad-500 dataset for sentiment classification. The results, presented in Table 5, show that araBERT produced better quality predictions with more negative labels, while camelBERT assigned a neutral sentiment label to the vast majority of texts, which is surprising. Therefore, we selected araBERT for further analysis. The sentiment distribution over categories of level 6 of the ArSOL taxonomy is depicted in Figure 2.

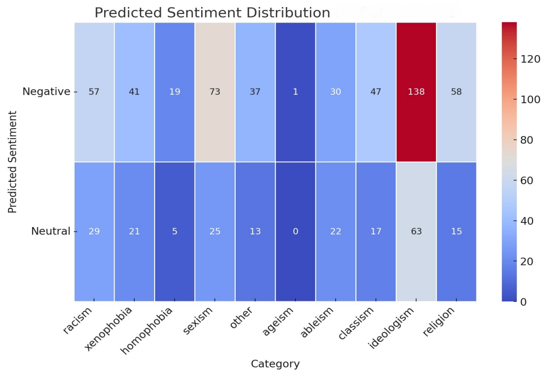

Figure 2: Sentiments seen at level 6 of ArSOL taxonomy.

4.2 Emotion Analysis

We also want to investigate whether incorporating emotion detection into offensive language classification will lead to more accurate classification results. Research has indicated that the use of emotional characteristics enhances the identification of hate speech (Plaza-del Arco et al., 2021).

We tested three emotion detection models for Arabic: (1) `hatemnoaman/bert-base-arabic-finetuned-emotion` model (Noaman, 2023) for emotion detection in Arabic which was fine-tuned from `asafaya/bert-base-arabic` on the Emotone dataset (Al-Khatib and El-Beltagy, 2018) and uses eight emotion labels (anger, disgust, fear, joy, neutral, sadness, surprise, trust); (2) `kiroloskhela/Sentiment-Bert` model (Khela, 2023) trained on Emotone (Al-Khatib and El-Beltagy, 2018) and the SetFit/Emotion (Tunstall et al., 2022) datasets with five emotions (disgust, joy, anger, fear, sadness; and (3) araBERT fine-tuned the Arabic Emotion Dataset (Almahdawi and Teahan, 2019) with five emotions (amused, confident, disgust, empathetic, fear). Distribution of detected emotions is shown in Table 6. Because the first two models assign almost all texts the Disgust label, we have elected to proceed with the third model (fine-tuned araBERT). By including emotion recognition in fine-grained offensive language classification, we hope for a better understanding of the intent and severity of offensive text.

4.3 Offensive Language Classification

To evaluate the effectiveness of the simplified taxonomy, we performed the classification of texts in the reFarad-500 dataset for every taxonomy level. We also study the effect of Sentiment Analysis (SA) and Emotion Detection (ED) on classification accuracy by using their output for classification. We first generate text representations for the original texts in Arabic, then optionally enhance them with

Model	Neutral	Positive	Negative
camelBERT	490	5	5
araBERT	269	0	231

Table 5: Sentiment classification results on the reFarad-500 dataset using two Arabic BERT models.

model	num of texts per emotion
asafaya/bert-base-arabic	disgust(448), joy(19), surprise(14),, neutral(7), sadness(5),fear(4), trust(3)
kiroloskhela/Sentiment-Bert	disgust(480), joy(10), anger(6), fear(2), sadness(2)
fine-tuned araBERT	amused(9), confident(378), disgust (82), empathetic(12), fear(19)

Table 6: Distribution of detected emotions.

SA or ED output, and split them into training and test sets with the 80%/20% split ratio. Then we apply classification models and report average results (precision, recall, F1 measure, and accuracy). The pipeline of our approach is shown in Figure 3. We only considered categories with at least 10 samples in a minority class, which excluded aspects such as Homophobia and Ageism.

4.3.1 Text Representations and Models

For the offensive language classification, we used three text representations: word n-grams of sizes 1 to 3 (denoted by n-grams in Table 7), tf-idf vectors, and BERT sentence embeddings (denoted by SE). We used two SE models – the Arabic bert-base-arabertv02 model denoted by the araBERT SE model (Antoun et al., 2020) and the multilingual bert-base-multilingual-cased model denoted by mlBERT SE (Devlin et al., 2019). We also investigated the enhancement of these representations with SA and ED labels.

We have applied eXtreme Gradient Boost (XGB) (Chen, 2015), Random Forest (RF) (Pal, 2005), and Logistic Regression (LR) (Kleinbaum et al., 2002) classifiers to these text representations. As baselines, we also applied mlBERT and arBERT and fine-tuned them on the training part of the data.

4.3.2 Results

Table 7 contains the results of classification for levels 2-6 of the taxonomy, showing the difference in performance on syntactic text representations versus semantic representations. At level 6, we classified each offensive aspect separately. In every case, we report the results of the representation-classifier combo that achieved the best accuracy for categories with 10 or more texts in a minority class. Semantic representations (arBERT SE, mlBERT SE) generally outperform syntactic representations (n-grams, tf-idf) in both accuracy and F1-score, highlighting the advantage of contextual embeddings. We can observe that all traditional models consistently outperform both BERT baselines, as can be seen in Table 8 (note some classes were omitted because they had less than 10 texts in the minority class as required by BERT models). We also observed that mlBERT performed better than arBERT in most cases.

Table 10 contains the results of the evaluation of reFarad-500 data enhanced with sentiments predicted by the model described in Section 4.1; the prediction was performed with a train/test split of 80%/20%. The accuracy values in this table indicate that adding SA had minimal impact on most categories, with some slight improvements (Vulgar and Ideologism) but also some decreases (Religion). Representation-wise, n-grams performed consistently well across multiple categories, often achieving competitive or higher accuracy compared to semantic representations like arBERT and mlBERT sentence embeddings. Classifier-wise, XGBoost (XGB) remained the best-performing model.

Table 9 contains the results of the evaluation of reFarad-500 data enhanced with emotions predicted by the model described in Section 4.2; the prediction was performed with train/test split of 80%/20%. The results indicate that incorporating emotion detection (ED) slightly improved accuracy in most categories, particularly in Vulgar and Offense strength, suggesting that emotions contribute to better classification of offensive content intensity. However, for some categories like Target presence and Ideologism, the accuracy changes were minimal, implying that ED might not significantly affect these aspects of offensive language. Overall, the results indicate the potential of emotion-aware models in enhancing fine-grained offensive language classification. Table 11 shows classification results for the reFarad-500 data enhanced by both sentiment analysis (SA) and emotion detection (ED) data. In no case were the results better than the results for the data enhanced only by SA or only by ED.

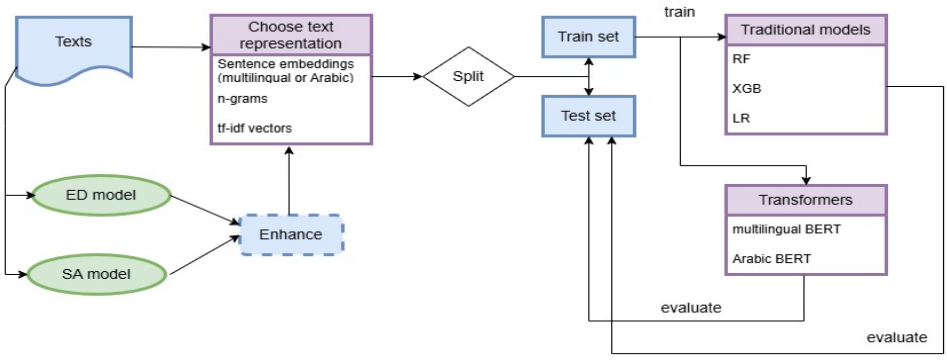

Figure 3: Offensive language classification pipeline.

	semantic text representations				syntactic text representations			
level	repr.	model	F1	acc	repr.	model	F1	acc
(2) target	arBERT SE	LR	0.5802	0.7300	n-grams	XGB	0.5019	0.6800
(3) target presence	mlBERT SE	XGB	0.5921	0.7200	n-grams	XGB	0.6685	0.7800
(4) vulgar	arBERT SE	RF	0.5739	0.8200	n-grams	LR	0.6755	0.8300
(5) hate speech/insult	tf-idf	LR	0.5128	0.6300	arBERT SE	RF	0.7460	0.7500
(5) discredit	mlBERT SE	LR	0.7288	0.7300	n-grams	XGB	0.7527	0.7600
(6) racism	arBERT SE	XGB	0.5765	0.7900	n-grams	RF	0.4350	0.7700
(6) xenophobia	mlBERT SE	LR	0.5821	0.8700	n-grams	RF	0.4624	0.8600
(6) homophobia	arBERT SE	XGB	0.4872	0.9500	n-grams	RF	0.4872	0.9500
(6) sexism	arBERT SE	XGB	0.6678	0.8400	n-grams	XGB	0.6114	0.8300
(6) other	arBERT SE	XGB	0.6094	0.9400	n-grams	XGB	0.6633	0.9300
(6) ageism	arBERT SE	XGB	1.0000	1.0000	n-grams	XGB	1.0000	1.0000
(6) ableism	arBERT SE	RF	0.4681	0.8800	tf-idf	XGB	0.5392	0.8800
(6) classism	arBERT SE	RF	0.4475	0.8100	n-grams	RF	0.4475	0.8100
(6) ideologism	mlBERT SE	LR	0.6484	0.6800	tf-idf	XGB	0.6239	0.6700
(6) religion	arBERT SE	XGB	0.5924	0.8800	tf-idf	XGB	0.6801	0.8900

Table 7: Comparison of classification results with semantic and syntactic text representations across taxonomy levels (best scores marked in gray).

level	model	F1	acc	model	F1	acc
(2) target	arBERT	0.3133	0.3780	mlBERT	0.2510	0.4180
(3) target presence	arBERT	0.5258	0.5480	mlBERT	0.5738	0.6060
(4) vulgar	arBERT	0.6576	0.7180	mlBERT	0.4576	0.6480
(5) hate speech/insult	arBERT	0.6094	0.6400	mlBERT	0.4781	0.4800
(5) discredit	arBERT	0.5478	0.5500	mlBERT	0.4916	0.5100
(6) racism	arBERT	0.5530	0.6580	mlBERT	0.3764	0.5220
(6) xenophobia	arBERT	0.4191	0.6420	mlBERT	0.5003	0.8480
(6) homophobia	arBERT	0.5589	0.9140	mlBERT	0.5146	0.9460
(6) sexism	arBERT	0.6695	0.7740	mlBERT	0.4820	0.6660
(6) religion	arBERT	0.5973	0.7360	mlBERT	0.4607	0.6240
(6) ableism	arBERT	0.5524	0.7260	mlBERT	0.4480	0.7300
(6) classism	arBERT	0.3826	0.5180	mlBERT	0.4950	0.7920
(6) ideologism	arBERT	0.7023	0.7140	mlBERT	0.5221	0.5340
(6) other	arBERT	0.5926	0.7660	mlBERT	0.5505	0.7820

Table 8: Fine-tuning of BERT models (best scores are marked in gray).

	no ED				with ED			
level	repr.	model	F1	acc	repr.	model	F1	acc
(2) target	arBERT SE	XGB	0.4887	0.7300	arBERT SE	XGB	0.4887	0.7300
(3) target presence	n-grams	XGB	0.6685	0.7800	n-grams	XGB	0.6593	0.7700
(4) vulgar	n-grams	LR	0.6755	0.8300	n-grams	LR	0.7137	0.8500
(5) hate speech/insult	arBERT SE	XGB	0.7052	0.7300	arBERT SE	RF	0.6614	0.7300
(5) discredit	ngrams	XGB	0.7029	0.7200	ngrams	XGB	0.7029	0.7200
(5) threat	arBERT SE	RF	0.4975	0.9900	arBERT SE	RF	0.4975	0.9900
(6) racism	arBERT SE	XGB	0.5765	0.7900	mlBERT SE	XGB	0.5847	0.8000
(6) xenophobia	mlBERT SE	LR	0.5821	0.8700	mlBERT SE	LR	0.5821	0.8700
(6) sexism	arBERT SE	XGB	0.6678	0.8400	arBERT SE	XGB	0.6678	0.8400
(6) religion	arBERT SE	XGB	0.6094	0.9400	n-grams	XGB	0.6842	0.9400
(6) ableism	arBERT SE	RF	0.4681	0.8800	arBERT SE	RF	0.4681	0.8800
(6) classism	arBERT SE	RF	0.4475	0.8100	arBERT SE	RF	0.4475	0.8100
(6) ideologism	mlBERT SE	LR	0.6484	0.6800	mlBERT SE	XGB	0.6494	0.7000
(6) other	n-grams	XGB	0.6464	0.8900	n-grams	LR	0.6464	0.8900

Table 9: Performance comparison of offensive language classification with and without emotion detection (ED) on the reFarad-500 dataset (best scores are marked in gray).

	no SA				with SA			
level	repr.	model	F1	acc	repr.	model	F1	acc
(2) target	arBERT SE	XGB	0.4887	0.7300	arBERT SE	LR	0.5802	0.7300
(3) target presence	n-grams	XGB	0.6685	0.7800	n-grams	XGB	0.6685	0.7800
(4) vulgar	n-grams	LR	0.6755	0.8300	n-grams	LR	0.6863	0.8400
(5) hate speech/insult	arBERT SE	XGB	0.7052	0.7300	arBERT SE	XGB	0.6881	0.7200
(5) discredit	ngrams	XGB	0.7029	0.7200	ngrams	XGB	0.7220	0.7300
(5) threat	arBERT SE	RF	0.4975	0.9900	arBERT SE	RF	0.4975	0.9900
(6) racism	arBERT SE	XGB	0.5765	0.7900	arBERT SE	XGB	0.5765	0.7900
(6) xenophobia	mlBERT SE	LR	0.5821	0.8700	mlBERT SE	LR	0.5821	0.8700
(6) sexism	arBERT SE	XGB	0.6678	0.8400	arBERT SE	XGB	0.6678	0.8400
(6) religion	arBERT SE	XGB	0.6094	0.9400	n-grams	LR	0.6464	0.8900
(6) ableism	arBERT SE	RF	0.4681	0.8800	arBERT SE	RF	0.4681	0.8800
(6) classism	arBERT SE	RF	0.4475	0.8100	arBERT SE	RF	0.4475	0.8100
(6) ideologism	mlBERT SE	LR	0.6484	0.6800	mlBERT SE	LR	0.6658	0.7000
(6) other	n-grams	XGB	0.6464	0.8900	arBERT SE	XGB	0.6094	0.9400

Table 10: Performance comparison of offensive language classification with and without sentiment analysis (SA) on the reFarad-500 dataset (best scores are marked in gray).

level	repr.	model	F1	acc	level	repr.	model	F1	acc
(2) target	arBERT SE	XGB	0.4783	0.7300	(6) homophobia	arBERT SE	XGB	0.6299	0.9500
(3) target presence	n-grams	XGB	0.6250	0.7600	(6) sexism	arBERT SE	XGB	0.6678	0.8400
(4) vulgar	n-grams	LR	0.6863	0.8400	(6) other	arBERT SE	XGB	0.6094	0.9400
(5) hate speech/insult	arBERT SE	XGB	0.7013	0.7300	(6) ageism	arBERT SE	XGB	1.0000	1.0000
(5) discredit	ngrams	XGB	0.7220	0.7300	(6) ableism	arBERT SE	RF	0.4681	0.8800
(5) threat	arBERT SE	RF	0.4975	0.9900	(6) classism	arBERT SE	RF	0.4475	0.8100
(6) racism	arBERT SE	XGB	0.5504	0.7900	(6) ideologism	mlBERT SE	LR	0.6349	0.6700
(6) xenophobia	mlBERT SE	LR	0.5821	0.8700	(6) religion	n-grams	LR	0.6464	0.8900

Table 11: Performance of offensive language classification with combined SA and ED enhancement on the reFarad-500 dataset.

5 Conclusions

This paper studies various levels of offensive language in Arabic following the ArSOL taxonomy of explicit offensive language. For this purpose, we re-annotate the existing dataset of (Liebeskind et al., 2024) to produce a quality dataset reFarad-500 covering multiple categories of Arabic offensive language. By applying various deep learning models, we assessed their effectiveness in detecting offensive content, and our experiments demonstrated that transformer models outperform traditional classifiers, highlighting their potential for this task. We also explored emotion detection and sentiment analysis to capture the emotional tone and subjective sentiment of offensive texts, showing that these methods are not merely auxiliary but integral to a comprehensive offensive language detection framework. The reFarad-500 dataset, together with full annotation guidelines, is freely available for research purposes at https://github.com/NataliaVanetik/OffensiveLanguageDatasetInArabicFinegrainAnnotation.

Future research should focus on expanding the dataset, integrating additional language resources, and enhancing classification models. The proposed taxonomy and annotation framework are designed to be adaptable, making them applicable not only to other Arabic dialects but also to languages with similar challenges in computational resources, data availability, and linguistic tooling. By combining taxonomy-driven annotation with semantic signals such as sentiment and emotion, this work offers a transferable foundation for offensive language detection across diverse linguistic contexts.

Ethics Statement and Limitations

This study re-annotates publicly available Arabic offensive language data from online platforms such as Twitter and Facebook, using only anonymized texts without identifiable information. Native Arabic speakers, trained with comprehensive guidelines, conducted the re-annotation to ensure consistency, minimize bias, and maintain cultural sensitivity. We acknowledge the subjective nature of offense and encourage ethical consideration when using models trained on this dataset.

The reFarad-500 dataset has several limitations: its size (500 texts) restricts large-scale model training and category diversity; the focus on explicit offense excludes implicit cases for future work; category distributions remain imbalanced to preserve real-world patterns; and coverage is limited to Modern Standard Arabic and Levantine dialects, reducing generalizability to underrepresented varieties such as Maghrebi or Gulf Arabic.

References

Ashraf Ahmad, Mohammad Azzeh, Eman Elnagi, Qasem Abu Al-Haija, Dana Halabi, Abdullah Aref, and Yousef AbuHour. 2024. Hate Speech Detection in the Arabic Language: Corpus Design, Construction and Evaluation. *Frontiers in Artificial Intelligence*, 7:1345445.

Azalden Alakrot, Liam Murray, and Nikola S Nikolov. 2018. Dataset construction for the detection of anti-social behaviour in online communication in Arabic. *Procedia Computer Science*, 142:174–181.

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.

Ali Alhazmi. 2023. Hate Speech Dataset for the Saudi Dialect. Mendeley Data.

Amer J Almahdawi and William J Teahan. 2019. A new arabic dataset for emotion recognition. In *Intelligent Computing-Proceedings of the Computing Conference*, pages 200–216. Springer.

Maha Jarallah Althobaiti. 2022. BERT-based approach to Arabic hate speech and offensive language detection in Twitter: exploiting emojis and sentiment analysis. *International Journal of Advanced Computer Science and Applications*, 13(5).

Mohamed Aly and Amir Atiya. 2013. LABR: A large scale arabic book reviews dataset. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 494–498, Sofia, Bulgaria. Association for Computational Linguistics.

Amr Al-Khatib and Samhaa R. El-Beltagy. 2018. *Emotional Tone Detection in Arabic Tweets*, volume 10762 of *Lecture Notes in Computer Science*, page 105–114. Springer, Cham.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Flor Miriam Plaza-del Arco, Sercan Halat, Sebastian Padó, and Roman Klinger. 2021. Multi-task learning with sentiment, emotion, and target detection to recognize hate speech and offensive language. *arXiv preprint arXiv:2109.10255*.

Abdullah Aref, Rana Husni Al Mahmoud, Khaled Taha, Mahmoud Al-Sharif, et al. 2020. Hate speech detection of arabic shorttext. In *9th International Conference on Information Technology Convergence and Services (ITCSE 2020)*, pages 81–94. Computer Science & Information Technology.

Tianqi Chen. 2015. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4.

Shammur Absar Chowdhury, Hamdy Mubarak, Ahmed Abdelali, Soon-gyo Jung, Bernard J Jansen, and Joni Salminen. 2020. A multi-platform Arabic news comment dataset for offensive language detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6203–6212.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Abdelrahim Elmadany, Chiyu Zhang, Muhammad Abdul-Mageed, and Azadeh Hashemi. 2020. Leveraging affective bidirectional transformers for offensive language detection. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 102–108.

Kabil Essefar, Hassan Ait Baha, Abdelkader El Mahdaouy, Abdellah El Mekki, and Ismail Berrada. 2023. OMCD: Offensive Moroccan Comments Dataset. *Language Resources and Evaluation*, pages 1–21.

Hatem Haddad, Hala Mulki, and Asma Oueslati. 2019. T-HSAB: A Tunisian hate speech and abusive dataset. In *International Conference on Arabic Language Processing*, pages 251–263. Springer.

Go Inoue, Hassan Sajjad, Fath Elrahman Saleh, Abdelrahim Elmadany, and Preslav Nakov. 2021. Camelbert: Open large-scale language models for arabic. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4706–4721. Association for Computational Linguistics.

Kirolos Khela. 2023. Sentiment-bert: Arabic sentiment analysis model. `https://huggingface.co/kiroloskhela/Sentiment-Bert`. Accessed: 2025-07-21.

David G Kleinbaum, K Dietz, M Gail, Mitchel Klein, and Mitchell Klein. 2002. *Logistic regression*. Springer.

Barbara Lewandowska-Tomaszczyk. 2023. A simplified taxonomy of offensive language (SOL) for computational applications. *Konin Language Studies*, 10(3):213–227.

Barbara Lewandowska-Tomaszczyk, Anna Bączkowska, Chaya Liebeskind, Giedre Valunaite Oleskeviciene, and Slavko Žitnik. 2023. An integrated explicit and

implicit offensive language taxonomy. *Lodz Papers in Pragmatics*, 19(1):7–48.

Barbara Lewandowska-Tomaszczyk, Slavko Žitnik, Anna Bączkowska, Chaya Liebeskind, Jelena Mitrović, and Giedrė Valūnaitė Oleškevičienė. 2021. LOD-connected offensive language ontology and tagset enrichment. In *R. Carvalho & R. Rocha Souza, R. (Eds.), Proceedings of the Workshop and Tutorials held at LDK 2021 co-located with the 3rd Language, Data and Knowledge Conference*, volume 3064, pages 135–150. CEUR Workshop Proceedings.

Barbara Lewandowska-Tomaszczyk, Slavko Žitnik, Chaya Liebeskind, Giedrė Valūnaitė-Oleškevičienė, Anna Bączkowska, Paul A Wilson, Marcin Trojszczak, Ivana Brač, Lobel Filipić, and Ana Ostroški Anić. 2022. Annotation scheme and evaluation: The case of offensive language. *Rasprave*.

Chaya Liebeskind, Ali Afawi, Marina Litvak, and Natalia Vanetik. 2024. Classifying offensive language in arabic: a novel taxonomy and dataset. *Lodz Papers in Pragmatics*, 20(2):433–462.

Chaya Liebeskind, Natalia Vanetik, and Marina Litvak. 2023. Hebrew offensive language taxonomy and dataset. *Lodz Papers in Pragmatics*, 19(2):325–351.

Marina Litvak, Natalia Vanetik, Yaser Nimer, Abdulrhman Skout, and Israel Beer-Sheba. 2021. Offensive language detection in semitic languages. In *Multimodal Hate Speech Workshop*, volume 2021, pages 7–12.

Khouloud Mnassri, Praboda Rajapaksha, Reza Farahbakhsh, and Noel Crespi. 2023a. Hate speech and offensive language detection using an emotion-aware shared encoder. In *ICC 2023-IEEE International Conference on Communications*, pages 2852–2857. IEEE.

Khouloud Mnassri, Praboda Rajapaksha, Reza Farahbakhsh, and Noël Crespi. 2023b. Hate speech and offensive language detection using an emotion-aware shared encoder. *CoRR*, abs/2302.08777.

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, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. 2020. Arabic offensive language on Twitter: Analysis and experiments. *arXiv preprint arXiv:2004.02192*.

Hala Mulki and Bilal Ghanem. 2021. Let-Mi: An Arabic Levantine Twitter Dataset for Misogynistic Language. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 154–163.

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.

Mahmoud Nabil, Mohamed Aly, and Amir Atiya. 2015. ASTD: Arabic sentiment tweets dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2515–2519, Lisbon, Portugal. Association for Computational Linguistics.

Hatem Noaman. 2023. Improved emotion detection framework for arabic text using transformer models. *Advanced Engineering Technology and Application*, 12(2):1–11.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684.

Mahesh Pal. 2005. Random forest classifier for remote sensing classification. *Journal of Remote Sensing*, 26(1):217–222.

Niloofar Safi Samghabadi, Afsheen Hatami, Mahsa Shafaei, Sudipta Kar, and Thamar Solorio. 2020. Attending the emotions to detect online abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 79–88.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts. *arXiv*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Cagri Coltekin. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447. Association for Computational Linguistics.