AraTar: A Corpus to Support the Fine-grained Detection of Hate Speech Targets in the Arabic Language

Seham Alghamdi[†,‡], **Youcef Benkhedda**[†], **Basma Alharbi**[◇] **and Riza Batista-Navarro**[†]
[†]Department of Computer Science, The University of Manchester, UK
[‡]Department of Information Systems, University of Jeddah, Saudi Arabia
[◇]Department of Computer Science and Artificial Intelligence, University of Jeddah, Saudi Arabia
{seham.alghamdi, youcef.benkhedda, riza.batista}@manchester.ac.uk, bmalharbi@uj.edu.sa

Abstract

We are currently witnessing a concerning surge in the spread of hate speech across various social media platforms, targeting individuals or groups based on their protected characteristics such as race, religion, nationality and gender. This paper focusses on the detection of hate type (Task 1) and hate target (Task 2) in the Arabic language. To comprehensively address this problem, we have combined and re-annotated hate speech tweets from existing publicly available corpora, resulting in the creation of AraTar, the first and largest Arabic corpus annotated with support for multi-label classification for both hate speech types and target detection with a high inter-annotator agreement. Additionally, we sought to determine the most effective machine learning-based approach for addressing this issue. To achieve this, we compare and evaluate different approaches, including: (1) traditional machine learning-based models, (2) deep learning-based models fed with contextual embeddings, and (3) fine-tuning language models (LMs). Our results demonstrate that fine-tuning LMs, specifically using AraBERTv0.2-twitter (base), achieved the highest performance, with a micro-averaged F1-score of 84.5% and 85.03%, and a macro-averaged F1-score of 77.46% and 73.15%, for Tasks 1 and 2, respectively.

Keywords: Hate speech detection, Arabic language models, Text classification, Annotated corpus

1. Introduction

The widespread propagation of hate speech messages on social media and the anonymity enjoyed by online users who post such messages have had an overwhelming negative impact on those targeted by hate speech (Alsafari et al., 2020a; Aluru et al., 2020). Moreover, hate speech can provoke dangerous reactions and online aggression amongst online users, which, in some cases, can spill over into physical harm to people (Aluru et al., 2020; Abu Farha and Magdy, 2020). Hate speech is defined as discriminating against, or insulting an individual or a group of people based on characteristics such as race, sexual orientation, ethnicity, religion, gender or nationality (ElSherief et al., 2018; Blaya, 2019). In addition to studying and detecting hate speech in general, it is imperative to identify the specific targets of hate speech, e.g., individuals or groups experiencing religious intolerance, racism and misogyny. Natural language processing (NLP) plays a critical role in detecting such content (Waseem and Hovy, 2016).

In this work, we cast Arabic Hate Speech and Target Detection (AHTD) as a text classification problem with two tasks. The first task (Task 1) is detecting hate speech within a message, classifying it according to pre-defined categories which are based on protected characteristics covered by the definition of hate speech: religion-hate (RH), ethnicity-hate (EH), nationality-hate (NH), gender-hate (GH), undefined-hate (UDH)[1] or clean (CL), with the last category pertaining to messages that do not contain hate according to the definition above. This task is considered to be a multi-label classification problem where any number of labels (i.e., the hate categories) can be assigned to a given message. The second task (Task 2) involves identifying the specific target of hate speech according to finer-grained categories under the above-mentioned hate categories. For example, targets for the religion-hate category could be Islam, Christianity or Judaism. This task is considered as a multi-label classification problem, as we cannot assume that every message is directed only towards one target; there are cases when there are multiple targets, hence approaches that assign only one label at a time are insufficient.

Targets are different in each hate category and are defined in this research as the individual or group of people possessing certain protected characteristics who are the subject of hate. The novelty of our work lies in addressing the second task, which thus far has been under-explored with respect to hate speech detection in Arabic. The main contributions[2] of this paper are:

- A new corpus, AraTar, with annotated hate

[1]Pertains to hate types different from RH, EH, NH and GH

[2]Our annotation guidelines, annotations and code are publicly available at https://github.com/SehamAlghamdi/AraTar.

OSACT 2024 Workshop, pages 1–12

types and hate targets, which supports the development of multi-label classification methods for automatically detecting types and targets of hate speech.

- A comparative study conducted to investigate different machine learning-based approaches, including: (1) traditional machine learning-based models, (2) deep learning-based models, and (3) fine-tuning language models (LMs).
- Comparative evaluation of the best performing model on our corpus and on other relevant corpora.

2. Related Work

Despite the abundance of Arabic corpora and approaches proposed for automatic hate speech detection, it is important to note that the number of such resources falls short in comparison to those available in English. While several efforts have been made to develop corpora and detection methods for Arabic hate speech, they primarily focus on distinguishing between hate and non-hate categories, or differentiating hate speech from offensive and abusive language. The development of resources specifically focussing on fine-grained hate speech detection and hate target identification remains limited.

Hate Type Detection (Task 1). Upon conducting a careful literature search, we noted that the majority of the corpora reported in the literature concentrated on detecting hate speech types and formalising the problem as either a multi-class classification problem whereby one out of multiple possible hate types is identified (Mubarak et al., 2023; Duwairi et al., 2021; Alsafari et al., 2020b; Al-Hassan and Al-Dossari, 2022; Anezi, 2022; Yadav et al., 2023), or a binary classification problem focussing on detecting whether a given input text contains a specific type of hate speech or not, e.g., religious hate (Albadi et al., 2018) and ethnicity hate (Alotaibi and Abul Hasanat, 2020). Only one study (Azzi and Zribi, 2022) developed a corpus and approaches compatible with multi-label classification, achieving a 79% micro-averaged F1-score. Seven classes were defined in their corpus to detect racism, sexism, religious hatred, xenophobia, violence, hate, pornography and LGBTQ hate (Azzi and Zribi, 2022).

Hate Target Identification (Task 2). A few studies have investigated the detection of specific targets of hate speech. Aref et al. (2020) and Alraddadi and Ghembaza (2021), for instance, focussed on anti-Islam or Islamophobic speech. They achieved varying levels of performance: F1-scores of 52% and 97% on their SSIT corpus and anti-Islamic corpus, respectively. In another work, the detection of anti-immigrant speech was explored by Mohdeb et al. (2022), obtaining an F1-score of 57% based on their own RED corpus. Speech containing sentiment against women (i.e., misogyny) was investigated in the Arabic Misogyny Identification (ArMI) shared task (Mulki and Ghanem, 2021). Six participating teams used the ArMI corpus, with the highest ranked team achieving a 91% macro-averaged F1-score (Mahdaouy et al., 2022). In a similar vein, the study by Guellil et al. (2022) focussed on women as hate targets, making use of their own Arabic_fr_en corpus. They obtained a macro-averaged F1-score of 86%. It is worth noting that all these studies formalised the detection of hate target as a binary classification problem.

We also noted common limitations among the existing corpora mentioned above. Firstly, the majority of them do not support multi-label classification, dealing with mutually exclusive classes only, thus ignoring the possibility that messages could pertain to multiple hate types or targets. Secondly, there is no standard labelling scheme for the types or targets of hate; each dataset follows a different set of hate types and targets. Furthermore, these existing corpora focussed on either only one type or one target of hate speech; therefore, there is no benchmark corpus for the task of fine-grained hate speech detection that covers multiple existing types and targets of hate speech in Arabic.

3. Data Collection and Annotation

We collected hate tweets from various available corpora and re-annotated them to facilitate a multi-label setting and to identify hate targets.

3.1. Data Collection

Five available corpora were used in collecting hate tweets, described as follows.

Arabic-Twitter corpus (Alsafari et al., 2020b). This is the first corpus that was constructed while considering the task of detecting different hate types. Specifically, four different hate types were explored: religion, ethnicity, gender and nationality hate, as well as offensive speech. It contains 5,340 tweets collected from Twitter where 1,423 tweets belong to the defined hate types. The tweets were obtained through robust search techniques using keywords, hashtags, user profiles, and phrases that defend groups with protected characteristics (as they are typically posted in response to hate-containing tweets which were retrieved to become part of the corpus). The researchers specifically included tweets written in the Gulf Arabic dialect and Modern Standard Arabic. The corpus was manually

annotated by native Arabic speakers, employing a three-level hierarchical annotation scheme for the binary classification of offensive and hate speech, ternary classification of offensive, hate speech and non-hate speech, and multi-class classification of different types of hate and offensive speech.

OSACT5 shared task corpus (Mubarak et al., 2023). The OSACT5 corpus was developed for the fine-grained hate speech detection shared task, consisting of 12,698 tweets where 1,339 tweets were labelled as containing hate.The tweets were collected from Twitter using an emoji-based method, where emojis that are known to often appear in offensive content were used. The annotation process incorporated a hierarchical annotation scheme to address three distinct sub-tasks: (1) offensiveness detection, treated as a binary classification task (offensive or non-offensive); (2) hate speech detection, also approached as a binary classification task (hate or non-hate); and (3) fine-grained hate detection, treated as a multi-class classification task with seven classes: hate based on nationality, race, and ethnicity, hate based on religion and belief, ideological hate, hate based on disability, hate based on social class, hate based on gender and non-hate speech.The tweets were written in both Modern Standard Arabic and various Arabic dialects and were annotated through crowd-sourcing.

Arabic hate-speech corpus (Al-Hassan and Al-Dossari, 2022). This corpus consists of 11K tweets, with 2,605 tweets labelled as containing hate. It was compiled by curating a list of hashtags associated with topics that are known to trigger hateful content. The annotation scheme employs multi-class classification, assigning one of five distinct classes to each tweet, namely religious hate, racial hate, sexism, general hate and no hate. The initial annotation was conducted by a volunteer, followed by a rigorous review process involving two additional volunteers to ensure the accuracy and consistency of the annotations.

Levantine Hate Speech and Abusive Language Dataset (L-HSAB) (Mulki et al., 2019). This corpus contains 5,846 tweets obtained through the Tweepy API and were written in the Lebanese and Syrian dialects. A lexicon-based approach was used to collect tweets from verified or popular political and social public figures' timelines, focussing on entities associated with hate, such as refugees. The annotated tweets in L-HSAB support multi-class classification, categorised into three classes: normal, hate, and abusive. The annotation was carried out by three annotators, who are Levantine native speakers.

3.2. Data Annotation Task

Our annotation task was carried out by three volunteer annotators, all of whom are native Arabic speakers and pursuing a higher education degree at that time. To ensure annotation quality, several meetings took place with the lead annotator (the first author of this paper who is also a native Arabic speaker) and volunteer annotators during the annotation stage. Firstly, a workshop was held with the annotators to describe the task and the process, including a training exercise with a number of example tweets from each category of both tasks. In the workshop, discussions about confusing and ambiguous cases were held. Then, a pilot study was conducted with the annotation team, who were given annotation guidelines and a set of 300 hate tweets that were previously annotated by the lead annotator. The annotators were asked to independently label the tweets using a hierarchical annotation scheme (described below). Their annotations were compared to those of the lead annotator in order to identify the most consistent annotator and to identify cases of disagreement. These cases were then discussed and clarified, and the annotation guidelines were revised accordingly.

Annotation Process. The process of annotation took six months and involved two stages. In Stage 1, 30% of the hate tweets in the corpus (1541 tweets) were annotated by all three annotators. Then, we evaluated the inter-annotator agreement (IAA) or reliability among the annotators. In Stage 2, the remaining 70% (3594 tweets) was divided among annotators for single annotation, each annotating 1198 tweets independently.

The annotation was performed using spreadsheets designed with drop-down lists that allow for multiple selections to support the annotators in annotating the tweets with one or more types or targets of hate.

Annotation Scheme. A hierarchical scheme formed the basis of the annotation of the corpus, shown in Figure 1. This scheme was designed for Task 1 based on the annotation scheme used in the Arabic Twitter corpus (Alsafari et al., 2020b), but refined to consider annotating one or more types of hate and non-predefined hate types, and extended to annotate targets of hate (Task 2).

In the proposed scheme, targets of hate were defined based on recently published work that highlighted the common targets of Arabic hate speech according to religion, nationality, and gender (Mubarak et al., 2023). For ethnicity hate, a pilot study was conducted on 30 tweets from the combined corpus to identify the most common ethnicity targets. Additionally, in the proposed scheme, the issue of annotating hate tweets that do not belong

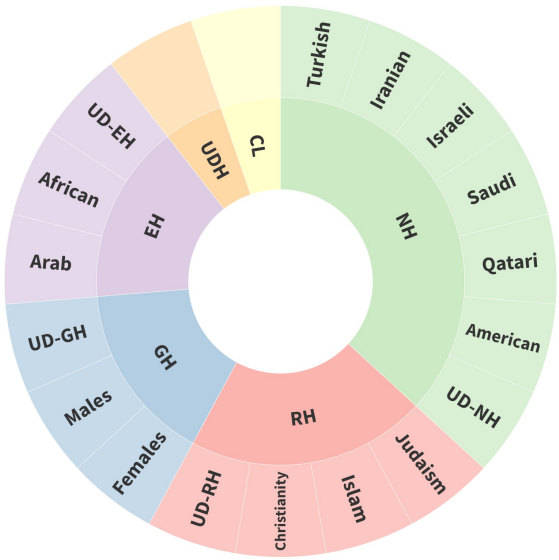

Figure 1: The AraTar Annotation Scheme. Key: RH = religion-hate, EH = ethnicity-hate, NH = nationality-hate, GH = gender hate, UDH = undefined-hate, CL = clean, UD = Undefined.

to the defined types and targets was addressed by defining an undefined-hate (UDH) category and undefined target categories, including undefined-RH (UD-RH), undefined-NH (UD-NH), undefined-EH (UD-EH), and undefined-GH (UD-GH). Figure 1 illustrates our taxonomy, i.e., the hate speech types and target categories in a hierarchical/sunburst form.

Annotation Guidelines. We have developed and validated annotation guidelines to provide our annotators with clear instructions for the tasks. Our annotation guidelines for Task 1 were inspired by the guidelines proposed by Alsafari et al. (2020b). However, we have extended these guidelines to include the annotation of hate types that are not covered in their annotation scheme, as well as the identification of hate targets. Furthermore, our guidelines take into account the annotation of implicit hate: when the type or target of hate is mentioned implicitly, either by using epithets or indirect references to the type or target of hate.

3.3. Annotation Results

As mentioned above, a common set consisting of 30% of the hate tweets in our corpus was independently annotated by the three annotators, thus allowing us to measure inter-annotator agreement (IAA). IAA was calculated using metrics that are suitable for multi-label scenarios such as F1-score (Hripcsak and Rothschild, 2005) and Krippendorff's α (Krippendorff, 1970, 2004), as they consider the distance/difference in annotations across all potential annotation units, regardless of the number of labels or annotators and the nature of annotation (including numeric, categorical and ordinal labels). The results, presented in Table1, show high agreement among the annotators in both Tasks 1 and 2. The average macro-averaged F1-scores are 97.21% and 97.18%, respectively, and the average micro-averaged F1-scores are 98.92% and 98.67% respectively. Similarly, Krippendorff's α is high, i.e., 98.76% in both tasks.

Metrics	Task1	Task2
Avg of Pairwise Macro-F1	97.21	97.18
Avg of Pairwise Micro-F1	98.92	98.67
Krippendorff's α	98.76	98.76

Table 1: IAA for Hate Type Detection (Task 1) and Hate Target Identification (Task 2).

Conflicting cases between the annotators were resolved by the lead annotator. At the end of the annotation process, 6124 tweets were added to the corpus for the clean (CL) category, drawn from the offensive and clean categories of the OSACT5 corpus. Additionally, 81 tweets from different datasets were manually labelled as CL, as closer inspection showed that they did not contain hate speech. Furthermore, 40 tweets were deleted due to duplication. The total number of tweets in AraTar is 11,219, spanning Modern Standard Arabic and a number of dialects including Gulf and Levantine. Figure 2 shows the label distribution according to hate type and hate target. Notably, in AraTar, 7% and 10% of hate tweets were annotated with more than one type of hate and more than one hate target, respectively.

4. Methodology

Upon completion of the annotation of the AraTar corpus, we set out to determine the performance of various classification models on Tasks 1 and 2. In this section, we describe the steps that we took towards this goal, including pre-processing of the tweets in the corpus, selection and design of three different types of classification models, and experimentation with the said models.

4.1. Data pre-processing

To prepare the corpus for analysis, we applied the following text pre-processing steps: removing diacritics, punctuation, repeated characters, symbols, special characters, URLs, English tokens and emojis to reduce noise, and performing letter normalisation by converting the forms of three letters into one form: Alif (أ ، آ ، إ to ا), Hamza (ؤ ، ئ to و ، ى) and Ta Marbouta (ة to ه). Next, the AraTar dataset was

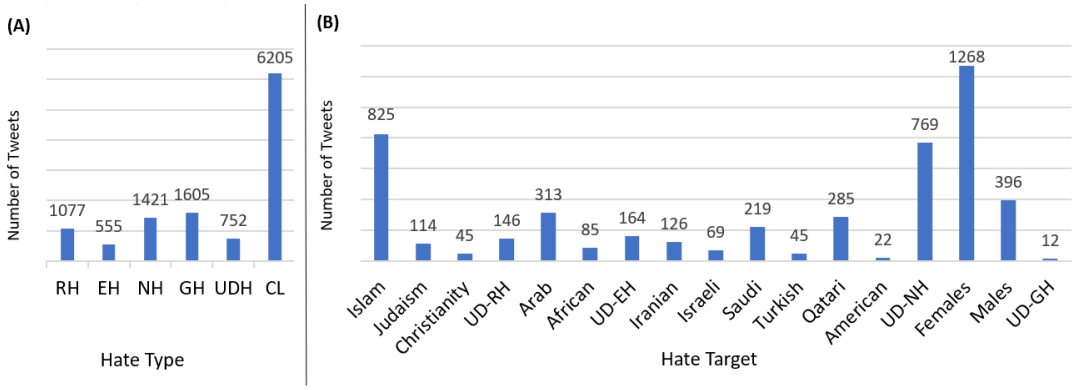

Figure 2: Label distribution in AraTar for the Hate Type Detection (A) and Hate Target Identification (B) tasks.

split using stratified sampling into three subsets: training, validation and test sets with proportions of 70%, 15% and 15% respectively.

4.2. Approaches and Models

We investigated the following machine learning-based approaches on our two tasks.

(1) Traditional Machine Learning-based Approach. We investigate support vector machine (SVM) models (Cortes and Vapnik, 1995) trained on features based on term frequency-inverse document frequency (TF-IDF) (Sparck Jones, 1972). In previous work, satisfactory results were achieved by employing SVM models fed with TF-IDF features, both in detecting hate type (Al-Hassan and Al-Dossari, 2022; Azzi and Zribi, 2022) and detecting hate target as a binary classification problem (Alraddadi and Ghembaza, 2021; Aref et al., 2020).

(2) Deep learning-based Model fed with Contextual Embeddings. We used a Long Short Term Memory (LSTM) model initialised with AraBERTv02-twitter embeddings. Apart from the ability of LSTM models to learn long-term dependencies between words, it has also proven its robustness in capturing and identifying multiple types of hate categories in the work of Al-Hassan and Al-Dossari (2022). LSTMs models have also shown good performance in the experiments conducted by Al-Hassan and Al-Dossari (2022) and Alsafari et al. (2020b) compared to other deep learning algorithms. Furthermore, the work of Alsafari et al. (2020b) demonstrated that the use of contextual word embeddings in LSTMs yields superior results compared to LSTM models with static word embeddings such as fastText and AraVec.

In our own work, we used the contextual word embeddings from AraBERTv02-twitter (base), a model variant of AraBERT (Antoun et al., 2020) that supports dialectal Arabic and is trained on Arabic tweets, as our data is from the Twitter platform. These embeddings were then fed as features to an LSTM model that was built upon the architecture proposed by Alsafari et al. (2020b).

(3) Fine-tuned Language Models. We used state-of-art transformer-based Arabic language models (LMs), namely, MARBERTV2 (Abdul-Mageed et al., 2021) and AraBERTv02-twitter (base and large) which are variants of AraBERTv2 (Antoun et al., 2020) since our data is from Twitter. These language models were used as they were pre-trained on dialectal Arabic tweets and are thus best-suited LMs for the downstream task of Arabic hate speech detection. In addition to that, models based on fine-tuning MARABERTv2 and AraBERTv2 achieved state-of-art results on hate speech detection cast as multi-class classification (AlKhamissi and Diab, 2022; Althobaiti, 2022; Bennessir et al., 2022; Shapiro et al., 2022) and binary classification (Abbes et al., 2021; Mahdaouy et al., 2022; Messaoudi et al., 2021; Mohdeb et al., 2022; Nwesri et al., 2021). On top of each pre-trained LM, we added a linear layer which computes a probability distribution based on the possible classes in the task at hand, i.e., either of Task 1 and Task 2.

4.3. Experimental Setup

Experiments on AraTar. We used identical training, validation and test sets across all five models: SVM, LSTM, MARBERT, AraBERT-base (AraBERT-b) and AraBERT-large (AraBERT-l). For the last four models, we employed the following hyperparameter settings: a maximum sequence length of 90, which considers the maximum sequence length in the corpus; the Adam optimiser; a learning rate of 5×10^{-5}; training for 50 epochs with early stopping based on validation loss; and for the fine-tuning of language models we used 16 as the batch size and binary cross-entropy as the loss function.

As mentioned above, our LSTM model was adopted from the architecture and implementations used in the study by Alsafari et al. (2020b). However, it was instead fed with contextual word embeddings, specifically AraBERTv02-twitter (both base and large variants) and evaluated on our corpus. The obtained results were disappointing, with low micro-averaged F1-scores of 25% and 2% for Task 1, and 43% and 11% for Task 2, using the

base and large models respectively. The macro-averaged F1-scores were also low, at 22% and 1% for Task 1, and 11% and 10% for Task 2 with the base and large models respectively. We thus optimised the hyperparameters used in training the LSTM model. Specifically, we set dropout and recurrent dropout to 0.2, and set batch size to 32.

Experiments on Other Corpora. To assess the performance of our top-performing model on other datasets, we conducted further fine-tuning using the OSACT5 (Mubarak et al., 2023) and Arabic Twitter datasets (Alsafari et al., 2020b), which were described in Section 3.1. The rationale behind choosing OSCAT5 and the Arabic Twitter dataset for comparison lies in their unique attributes. OSCAT5 stands out as the current benchmark corpus in the field, while the developed models using the Arabic Twitter dataset have demonstrated superior performance in previous literature. Furthermore, both datasets offer readily available training and test sets, ensuring the comparability of our experiments. It is worth noting that there are currently no other available corpora specifically focussed on the types of hate speech. Our experiments were conducted using the complete datasets and the original training and test sets provided by the authors. We however excluded Disability hate from OSACT5 due to its limited representation, with only two tweets in the entire corpus. Table 2 summarises the class frequencies in both datasets.

Arabic Twitter		OSACT5	
Classes	**Count**	**Classes**	**Count**
RH	321	Race-HS1	366
EH	382	Religion-HS2	38
NH	368	Ideology-HS3	190
GH	352	Social Class-HS5	101
OFF	437	Gender-HS6	641
Clean	3480	NOT_HS	11359
Total	**5340**	**Total**	**12695**

Table 2: Frequencies of hate types in the Arabic Twitter and OSACT5 corpora.

Additionally, since these corpora have a maximum sequence length close to that in AraTar, we kept the same hyperparameter value for model training to maintain consistency. We also used the same values as before, for the rest of the hyperparameters.[3]

Evaluation Metrics. Following standard practices, we calculated the precision, recall and F1-score to evaluate the performance of the classification models. Additionally, we report the exact match ratio metric (EMR) in our experiments on AraTar, which is commonly used for multi-label scenarios to measure the proportion of predicted outputs that exactly match the ground truth.

5. Evaluation Results

Tables 3 and 4 present the evaluation results in terms of F1-score (F1) for each label and model for Tasks 1 and 2, respectively.[4] Additionally, we provide the combined performance of each model in terms of micro-averaged F1-score (micro-F1), macro-averaged F1-score (macro-F1) and EMR. From the obtained results, it is noticeable that overall, the fine-tuned LMs, particularly the models that use AraBERTv2-twitter (i.e., AraBERT-b and AraBERT-l) obtained superior performance over the SVM and LSTM models in both tasks.

Classes	**SVM**	**LSTM**	**MARBERT**	**AraBERT-b**	**AraBERT-l**
	F1	F1	F1	F1	F1
RH	76.55	82.72	85.29	**86.15**	82.43
EH	63.16	68.70	**79.76**	79.53	78.69
NH	65.16	67.18	75.26	79.07	**80.00**
GH	72.98	77.10	**79.09**	76.28	78.75
UDH	27.27	36.16	48.09	**51.98**	39.53
CL	86.08	87.44	90.10	**91.76**	90.04
Micro-F1	77.78	79.37	83.55	**84.50**	83.56
Macro-F1	65.20	69.88	76.26	**77.46**	74.91
EMR	70.19	72.62	74.26	**81.83**	55.29

Table 3: Evaluation Results for Hate Type Detection (Task 1).

Classes	**SVM**	**LSTM**	**MARBERT**	**AraBERT-b**	**AraBERT-l**
	F1	F1	F1	F1	F1
Islam	82.16	90.91	90.27	90.20	**91.25**
Judaism	48.00	72.73	72.73	**75.86**	74.07
Christianity	22.22	**66.67**	00.00	**66.67**	54.55
UD-RH	23.53	74.07	69.23	**75.00**	71.43
Arab	65.75	79.52	**83.15**	78.65	79.55
African	64.00	69.23	78.57	**90.32**	86.67
UD-EH	48.48	80.00	84.44	82.93	**88.37**
Iranian	34.48	64.71	76.60	**80.95**	76.92
Israeli	00.00	50.00	**61.54**	55.56	50.00
Saudi	26.09	68.66	70.27	**72.46**	72.22
Turkish	00.00	71.43	61.54	**88.89**	87.50
Qatari	69.33	79.52	82.76	**91.67**	89.13
American	00.00	00.00	00.00	**40.00**	00.00
UD-NH	63.58	**85.45**	83.25	84.40	83.65
Females	84.42	**93.44**	91.78	90.34	92.91
Males	62.22	78.10	77.69	79.67	**85.71**
UD-GH	00.00	00.00	00.00	00.00	00.00
Micro-F1	68.87	84.20	83.66	85.03	**86.05**
Macro-F1	40.84	66.14	63.75	**73.15**	69.64
EMR	53.49	77.36	74.26	**77.52**	72.56

Table 4: Evaluation Results for Hate Target Identification (Task 2).

Hate Type Detection (Task 1). The results in Table 3 show that AraBERT-b obtained the highest micro-averaged F1-score of 84.50%, followed by AraBERT-l which obtained 83.56%. Notably, AraBERT-b consistently outperformed other models in Task 1, according to the three metrics for combined performance (micro-F1, macro-F1 and EMR) with a significant margin in terms of EMR.

[3] Implementation details including the hardware and software frameworks that were used in our experiments are provided in Appendix B.

[4] Precision and recall values are reported in Appendix C.

The highest score for EMR in Task 1 is 81.83% (AraBERT-b) followed by 74.26% (MARBERT), resulting in an improvement of 7.5 percentage points. This indicates that AraBERT-b is the most dependable model for accurately identifying various forms of hate in Arabic tweets.

AraBERT-b displayed superior performance in accurately classifying religious hate, undefined hate and clean (CL) categories compared to the other models. However, ethnicity and gender hate were better identified by MARBERT by a margin of 0.23 and 2.81 percentage points, respectively, compared to AraBERT-b.

It is also worth noting that, although not directly comparable, our best model (AraBERT-b) outperformed the model proposed by Azzi and Zribi (2022). Even though a direct comparison might not be entirely apt, their model, often regarded as the state-of-the-art in the literature, obtained a micro-averaged F1-score of 79%. In contrast, our model achieved 84.50% for the same metric. Moreover, to best of our knowledge, in terms of macro-averaged F1-score, AraBERT-b achieved a higher score compared to the majority of the existing models reported in the literature (Alsafari et al., 2020a,b; Al-Hassan and Al-Dossari, 2022; Duwairi et al., 2021; Althobaiti, 2022; Bennessir et al., 2022; Magnossão de Paula et al., 2022; Shapiro et al., 2022; AlKhamissi and Diab, 2022; Albadi et al., 2018, 2019).

Hate Target Identification (Task 2). The best performance was obtained by AraBERT-l, with a micro-averaged F1-score of 86.05%, gaining a 1 percentage point improvement over AraBERT-b which obtained 85.03% on the same metric. For the remaining two overall metrics, AraBERT-b achieved the highest scores of 73.15% and 77.52% in terms of macro-averaged F1-score and EMR. For these two metrics, the next best scores were 69.64% (AraBERT-l) and 77.36% (LSTM). This indicates that AraBERT-b obtained a 4 and 0.16 percentage point improvement on macro-averaged F1-score and EMR, respectively. AraBERT-b demonstrated superior performance in learning nine hate targets: Judaism, Christianity, undefined religious targets, African, Iranian, Saudi, Turkish, Qatari and American. In contrast, for categories like Islam, undefined ethnicity and males, AraBERT-l outperformed AraBERT-b, with improvements of 1.25, 5.44 and 6 percentage points, respectively. For other categories such as Arab, Israeli, undefined nationality and females, either MARBERT or LSTM proved to be superior, showing gains of 4.5, 5.98, 1.05 and 3.1 percentage points over AraBERT-b, respectively.

A notable limitation of the classification models is their difficulty in accurately identifying undefined gender hate targets. AraBERT-b, along with all other models, did not effectively learn to identify this target. This could be attributed to the low number of samples in the training set.

Furthermore, when comparing our best model (AraBERT-b) with those reported in the literature, there is an absence of reporting the micro-averaged F1-score of the published models and a lack of studies that have developed generalised detection models that consider different targets in a multi-label classification task. However, when looking at the macro-averaged F1-score, AraBERT-b performed lower than the majority of published models. This may be attributed to the imbalanced distribution in our dataset and the more complex nature of the multi-label classification task compared to binary classification.

6. Discussion

Error Analysis. We conducted error analysis by inspecting some of the misclassified cases produced by the best model in each task. A total number of 306 samples and 145 samples were misclassified in Tasks 1 and 2, respectively, with 82 overlapping samples. We have four main observations, outlined below.

Disclaimer: Due to the nature of this work, our examples contain hate speech which some readers might find offensive. These do not in any way reflect the researchers' own views or opinions.

(1) Mention of hate targets in a neutral context might mislead the trained classifier: We identified instances where mentions of potential hate targets were used in a neutral context, thus misleading the classifier. For instance, "Houthis" in the tweet *"By God, show us at the borders with the Houthis, O Mas'ood. Two states are with you. Seriously, they are besieged and you couldn't handle them, O'Utaibi, O effeminate"*

واله ورينا في الحدود مع الحوثي يا مسعود ٢ دوله معاكم
علي جماعه محاصره وما قدرتوا عليهم يا عتيبي يا خنيث

(2) Implicit hate: We recognised that in some cases the classification model fails to detect implicit hate, as in the following post with implicit gender hate towards women. *"O [vomiting emoji], Do not believe themselves, butterflies and dancers"*

يع لايصدقون نفسهم قسم الفراشات والغوازي

In the context of this tweet, "butterflies" and "dancers" are allegorically used to refer to women. Such coded language presents challenges for our classifier due to its inherent subtlety. This inability to predict such coded language can be addressed by employing a dataset that captures many examples of such cases.

At times, epithets are mentioned in the content that refers to a hate target. An example is the post: *"I'm tired from cursing and insulting the Be*uins.*

AraTar		Arabic Twitter		OSACT5	
Classes	**F1**	**Classes**	**F1**	**Classes**	**F1**
RH	86.15	RH	81.32	Race-HS1	43.24
EH	79.53	EH	82.03	Religion-HS2	0.00
NH	79.07	NH	83.70	Ideology-HS3	21.62
GH	76.28	GH	76.62	Social Class-HS5	0.00
UDH	51.98	OFF	80.32	Gender-HS6	64.20
CL	91.76	Clean	94.83	NOT_HS	96.16
Micro-F1	84.50	Micro-F1	91.14	Micro-F1	92.44
Macro-F1	77.46	Macro-F1	83.14	Macro-F1	37.54

Table 5: Results of AraBERT-b on the AraTar, Arabic Twitter and OSCAT5 corpora.

They don't know that this sound could explode a child's ear and make them deaf due to this ignorance. Please, Mohammed, find a solution to this drifting"

تعبت وانا العن و اسب الب* و مايدرون ان هالصوت ممكن يفجر اذن الطفل و يصير اصم بسبب هالتخلف تكفى يامحمد شف حل للطعوس

The tweet combines a personal feeling of exhaustion with a negative generalisation about the Bedouins, suggesting ignorance. However, it does not mention any explicit derogatory terms. It mentions "drifting", an epithet used for Bedouins. It is worth noting that the use of an asterisk (*) to mask some characters in the word *"Bedouins"* was likely a means for avoiding detection by Twitter's automatic moderation tools.

(3) Correlation between less presented subtargets and contents might lead to misclassification: For instance, the Islam hate target, Houthi (an extremist Islamic group), was misclassified as nationality hate and as an undefined nationality hate target in Tasks 1 and 2, respectively. We can interpret the reason for this behaviour as the correlation between Houthi and Yaman in the content.

(4) Offensive tweets that were predicted as having a hate type: For example, the following offensive tweet was classified as gender hate: *"We are on time, has many frivolous people, they are disgusting [face with medical mask emoji]."*

نحنُ في زمن كَثُر فيه الخفيفون و الخفيفات، مثيرين للاشمئزاز

Comparison with Other Corpora. Table 5 presents the results of applying AraBERT-b on the other corpora with multi-class classification settings. The motivation for conducting this comparison is two-fold:

(1) Highlight the extent to which AraTar can enable a model to learn the hate type classification task. Upon closer examination of the F1-score for each label, it becomes apparent that performance on AraTar is better than on OSACT5. Moreover, it is noticeable that performance on the social hate and religion hate classes is 0. This can be attributed to their under-representation in OSACT5, making it challenging for the model to learn and generalise effectively to these specific classes. Furthermore, even though the Arabic Twitter corpus has a more balanced distribution, AraTar was able to provide comparable results.

The reason for the reduction in the overall performance on AraTar is the score for the UDH class which is one of the minority classes and has a diversity of tweets that convey different hate types that do not belong to the other categories. These empirical findings lead to the conclusion that classification based on AraTar yields satisfactory results although UDH is difficult to detect. Unlike the Arabic Twitter dataset that supports the detection of only one hate type at a time, AraTar supports the detection of messages containing general hate as well as any number of defined hate types where they exist.

(2) Assess whether our best performing model (AraBERT-b) obtains competitive performance on multi-class classification of hate type, when compared with the state-of-the-art models previously reported for the other corpora. The obtained results demonstrate a significant improvement in the detection performance on the Arabic Twitter dataset achieved by AraBERT-b in terms of the reported macro-averaged F1-score of state-of-art models. Alsafari et al. (2020a) used the Arabic Twitter dataset and achieved the highest macro-averaged F1-score at 80.23% using an ensemble model that employed the BiLSTM architecture with AraBERTv1 embeddings and applying the average value method for aggregation. Our model outperformed this performance by 2.91 percentage points. However, AraBERT-b exhibits a 15.3 percentage point decrease when compared to the state-of-the-art model on OSACT5, which achieved a macro-averaged F1 score of 52.8. This model, which ranked first in the OSACT5 competition (Mubarak et al., 2022), was designed using multi-task learning techniques. Specifically, its architecture consists of a hard parameter-sharing layer composed of AraBERTv2 contextualised text representation models and subtask-specific layers. These subtask-specific layers were fine-tuned using quasi-recurrent neural networks (QRNNs) for each subtask. The model was trained on two tasks: the detection of offensive speech and general hate speech (Magnossão de Paula et al., 2022).

7. Conclusion

We present AraTar, a corpus to support the fine-grained detection of Arabic hate speech targets. It addresses the previously limited scale of Arabic hate speech detection and the lack of unified annotation in previous datasets. Our experiments show that fine-tuning language models, especially AraBERTv2-twitter, yields favourable results for both the Hate Type Detection and Hate Target

Identification tasks. An AraBERT model trained on AraTar also fares well in comparison with the same model architecture trained on other corpora.

Limitations

The main limitations of AraTar lie in the fact that not all Arabic dialects are covered, and that the corpus is confined to tweets. Furthermore, specific targets are under-represented, thus affecting classification performance for these targets. Future work will focus on broadening the scope of the corpus to include diverse dialects and platforms, and on employing data augmentation methods to generate synthetic data to improve the representation of minority hate targets. Another future direction is the enhancement of the capability of models in detecting hate targets by developing a stronger model using techniques such as parameter-efficient tuning (Yang et al., 2022) or ensemble methods as described in the study by Alsafari and Sadaoui (2021).

Bibliographical References

Istabrak Abbes, Eya Nakache, and Moez BenHajHmida. 2021. Context-aware language modeling for arabic misogyny identification. In *Working Notes of the 2021 Forum for Information Retrieval Evaluation (FIRE 2021)*, pages 847–851.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105. Association for Computational Linguistics.

Ibrahim Abu Farha and Walid Magdy. 2020. Multitask Learning for Arabic Offensive Language and Hate-Speech Detection. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 86–90, Marseille, France. European Language Resource Association.

Areej Al-Hassan and Hmood Al-Dossari. 2022. Detection of hate speech in Arabic tweets using deep learning. *Multimedia systems*, 28(6):1963–1974.

Nuha Albadi, Maram Kurdi, and Shivakant Mishra. 2018. 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)*, pages 69–76.

Nuha Albadi, Maram Kurdi, and Shivakant Mishra. 2019. Investigating the effect of combining GRU neural networks with handcrafted features for religious hatred detection on Arabic Twitter space. *Social Network Analysis and Mining*, 9(1):41.

Badr AlKhamissi and Mona Diab. 2022. Meta AI at Arabic Hate Speech 2022: MultiTask Learning with Self-Correction for Hate Speech Classification. 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*, pages 186–193, Marseille, France. European Language Resources Association.

Afaf Alotaibi and Mozaherul Hoque Abul Hasanat. 2020. Racism Detection in Twitter Using Deep Learning and Text Mining Techniques for the Arabic Language. In *Proceedings of the 1st International Conference of Smart Systems and Emerging Technologies (SMARTTECH)*, pages 161–164.

Rawan Abdullah Alraddadi and Moulay Ibrahim El-Khalil Ghembaza. 2021. Anti-Islamic Arabic Text Categorization using Text Mining and Sentiment Analysis Techniques. *International Journal of Advanced Computer Science and Applications*, 12(8).

Safa Alsafari and Samira Sadaoui. 2021. Ensemble-based semi-supervised learning for hate speech detection. In *The International FLAIRS Conference Proceedings*, volume 34.

Safa Alsafari, Samira Sadaoui, and Malek Mouhoub. 2020a. Deep Learning Ensembles for Hate Speech Detection. In *Proceedings of the 32nd IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 526–531.

Safa Alsafari, Samira Sadaoui, and Malek Mouhoub. 2020b. Hate and offensive speech detection on Arabic social media. *Online Social Networks and Media*, 19:100096.

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).

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. 2020. Deep learning models for multilingual hate speech detection. *arXiv preprint arXiv:2004.06465*.

Faisal Yousif Al Anezi. 2022. Arabic Hate Speech Detection Using Deep Recurrent Neural Networks. *Applied Sciences*, 12(12):6010.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based Model for Arabic Language Understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Abdullah Aref, Rana Husni Al Mahmoud, Khaled Taha, and Mahmoud Al-Sharif. 2020. Hate Speech Detection of Arabic Shorttext. In *Proceedings of the 9th International Conference on Information Technology Convergence and Services (ITCSE 2020)*, pages 81–94. AIRCC Publishing Corporation.

Salma Azzi and Chiraz Zribi. 2022. Comparing Deep Learning Models for Multi-label Classification of Arabic Abusive Texts in Social Media. In *Proceedings of the 17th International Conference on Software Technologies*, pages 374–381, Lisbon, Portugal. SCITEPRESS - Science and Technology Publications.

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 *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*, pages 176–180, Marseille, France. European Language Resources Association.

Catherine Blaya. 2019. Cyberhate: A review and content analysis of intervention strategies. *Aggression and Violent Behavior*, 45:163–172.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Rehab Duwairi, Amena Hayajneh, and Muhannad Quwaider. 2021. A Deep Learning Framework for Automatic Detection of Hate Speech Embedded in Arabic Tweets. *Arabian Journal for Science and Engineering*, 46(4):4001–4014.

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate Lingo: A Target-Based Linguistic Analysis of Hate Speech in Social Media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Imane Guellil, Ahsan Adeel, Faical Azouaou, Mohamed Boubred, Yousra Houichi, and Akram Abdelhaq Moumna. 2022. Ara-Women-Hate: An Annotated Corpus Dedicated to Hate Speech Detection against Women in the Arabic Community. In *Proceedings of the Workshop on Dataset Creation for Lower-Resourced Languages within the 13th Language Resources and Evaluation Conference*, pages 68–75, Marseille, France. European Language Resources Association.

George Hripcsak and Adam S. Rothschild. 2005. Agreement, the F-Measure, and Reliability in Information Retrieval. *Journal of the American Medical Informatics Association: JAMIA*, 12(3):296–298.

Klaus Krippendorff. 1970. Bivariate agreement coefficients for reliability of data. *Sociological Methodology*, 2:139–150.

Klaus Krippendorff. 2004. Measuring the reliability of qualitative text analysis data. *Quality and quantity*, 38:787–800.

Angel Felipe Magnossão de Paula, Paolo Rosso, Imene Bensalem, and Wajdi Zaghouani. 2022. UPV at the Arabic Hate Speech 2022 Shared Task: Offensive Language and Hate Speech Detection using Transformers and Ensemble Models. 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*, pages 181–185, Marseille, France. European Language Resources Association.

Abdelkader El Mahdaouy, Abdellah El Mekki, Ahmed Oumar, Hajar Mousannif, and Ismail Berrada. 2022. Deep Multi-Task Models for Misogyny Identification and Categorization on Arabic Social Media. *arXiv preprint arXiv:2206.08407*.

Abir Messaoudi, Chayma Fourati, Mayssa Kchaou, and Hatem Haddad. 2021. iCompass Working Notes for Arabic Misogyny Identification. In *Working Notes of the 2021 Forum for Information Retrieval Evaluation (FIRE 2021)*, page 5.

Djamila Mohdeb, Meriem Laifa, Fayssal Zerargui, and Omar Benzaoui. 2022. Evaluating transfer learning approach for detecting Arabic anti-refugee/migrant speech on social media. *Aslib Journal of Information Management*, 74(6):1070–1088.

Hamdy Mubarak, Hend Al-Khalifa, and Abdulmohsen Al-Thubaity. 2022. 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*, pages 162–166, Marseille, France. European Language Resources Association.

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 and Bilal Ghanem. 2021. ArMI at FIRE 2021: Overview of the First Shared Task on Arabic Misogyny Identification. In *Working Notes of the 2021 Forum for Information Retrieval Evaluation (FIRE 2021)*, pages 820–830.

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 3rd Workshop on Abusive Language Online*, pages 111–118, Florence, Italy. Association for Computational Linguistics.

Abdusalam Nwesri, Stephen Wu, and Harmain Harmain. 2021. Detecting Misogyny in Arabic Tweets. In *Working Notes of the 2021 Forum for Information Retrieval Evaluation (FIRE 2021)*, page 6.

Ahmad Shapiro, Ayman Khalafallah, and Marwan Torki. 2022. AlexU-AIC at Arabic Hate Speech 2022: Contrast to Classify. 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*, pages 200–208, Marseille, France. European Language Resources Association.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21.

Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Ankit Yadav, Shubham Chandel, Sushant Chatufale, and Anil Bandhakavi. 2023. LAHM: Large Annotated Dataset for Multi-Domain and Multilingual Hate Speech Identification. *arXiv preprint arXiv:2304.00913*.

Zhuoyi Yang, Ming Ding, Yanhui Guo, Qingsong Lv, and Jie Tang. 2022. Parameter-Efficient Tuning Makes a Good Classification Head. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7576–7586. Association for Computational Linguistics.

Appendix

A. Annotation Guidelines

The annotation guidelines can be downloaded from our Github repository.[5]

B. Implementation Details

B.1. Hardware

For both Tasks 1 and 2, we ran the SVM and LSTM experiments on a single Tesla V100 GPU with 51 GB RAM using the Google Colab Pro+ platform.[6] Also, we used a single NVIDIA A100 with 84 GB RAM to run the fine-tuning experiments with MARBERT, AraBERT-b and AraBERT-l.

B.2. Software Frameworks

Python 3.10.12 was used in implementing all models and experiments. Different machine learning frameworks were used. Firstly, the scikit-learn toolkit[7] was used in developing the SVM model. Additionally, we employed the skmultilearn library[8] which applies the binary relevance technique to a multi-label classification problem. For our LSTM model, Keras[9] was used. Lastly, we utilised Hugging Face's Transformers library[10] to fine-tune the pre-trained MARBERTv2 and AraBERT-twitter (base and large) language models for our multi-label classification tasks. Specifically, we loaded them and built our models using the `AutoModelForSequenceClassification` class, leveraging Hugging Face's Trainer API.

For evaluation, we used the metrics implemented in the scikit-learn toolkit.

For reproducibility, we set the seed parameter to 42 in all AraTar experiments.

[5] https://github.com/SehamAlghamdi/AraTar
[6] https://colab.research.google.com/
[7] https://scikit-learn.org/stable/
[8] http://scikit.ml/
[9] https://keras.io/
[10] https://huggingface.co/docs/transformers/index

C. Detailed Results

For both Tasks 1 and 2, we report the results of a single run trained for 50 epochs with early stopping based on validation loss. Tables 6, 7 and 8 present detailed results, including precision and recall scores, to complement Tables 3, 4 and 5 in the paper.

	SVM			**LSTM**			**MARBERT**			**AraBERT-b**			**AraBERT-l**		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
RH	**86.05**	68.94	76.55	82.21	83.23	82.72	82.56	**88.20**	85.29	85.37	86.96	**86.15**	84.87	80.12	82.43
EH	85.71	50.00	63.16	**95.74**	53.57	68.70	79.76	79.76	**79.76**	78.16	80.95	79.53	72.73	**85.71**	78.69
NH	83.33	53.49	65.16	74.86	60.93	67.18	**84.39**	67.91	75.26	79.07	**79.07**	79.07	80.95	**79.07**	**80.00**
GH	82.29	65.56	72.98	72.96	**81.74**	77.10	78.93	79.25	**79.09**	**86.77**	68.05	76.28	85.44	73.03	78.75
UDH	**75.00**	16.67	27.27	46.38	29.63	36.16	58.67	40.74	48.09	49.58	**54.63**	**51.98**	53.12	31.48	39.53
CL	85.81	86.36	86.08	**95.31**	80.77	87.44	92.33	87.97	90.10	91.46	92.05	**91.76**	87.06	**93.23**	90.04
Micro	84.96	71.72	77.78	85.33	74.20	79.37	**86.28**	80.98	83.55	85.21	**83.79**	**84.50**	83.85	83.28	83.56
Macro	**83.03**	56.84	65.20	77.91	64.98	69.88	79.44	73.97	76.26	78.40	**76.95**	**77.46**	77.36	73.78	74.91

Table 6: Complete Results for Hate Type Detection (Task 1).

	SVM			**LSTM**			**MARBERT**			**AraBERT-b**			**AraBERT-l**		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Islam	90.83	75.00	82.16	90.91	**90.91**	90.91	92.80	87.88	90.27	**93.50**	87.12	90.20	91.60	**90.91**	**91.25**
Judaism	75.00	35.29	48.00	75.00	70.59	72.73	75.00	70.59	72.73	91.67	64.71	**75.86**	**1.00**	58.82	74.07
Christian.	**1.00**	12.50	22.22	**1.00**	**50.00**	**66.67**	00.00	00.00	00.00	**1.00**	**50.00**	**66.67**	**1.00**	37.50	54.55
UD-RH	66.67	14.29	23.53	**76.92**	71.43	74.07	75.00	64.29	69.23	66.67	**85.71**	**75.00**	71.43	71.43	71.43
Arab	85.71	53.33	65.75	**86.84**	73.33	79.52	84.09	**82.22**	**83.15**	79.55	77.78	78.65	81.40	77.78	79.55
African	**1.00**	47.06	64.00	**1.00**	52.94	69.23	**1.00**	64.71	78.57	**1.00**	**82.35**	**90.32**	**1.00**	76.47	86.67
UD-EH	88.89	33.33	48.48	**1.00**	66.67	80.00	90.48	**79.17**	84.44	**1.00**	70.83	82.93	**1.00**	**79.17**	**88.37**
Iranian	55.56	25.00	34.48	78.57	55.00	64.71	66.67	**90.00**	76.60	77.27	85.00	**80.95**	**78.95**	75.00	76.92
Israeli	00.00	00.00	00.00	75.00	37.50	50.00	**80.00**	50.00	**61.54**	50.00	**62.50**	55.56	75.00	37.50	50.00
Saudi	75.00	15.79	26.09	79.31	60.53	68.66	72.22	**68.42**	70.27	**80.65**	65.79	**72.46**	76.47	**68.42**	72.22
Turkish	00.00	00.00	00.00	**1.00**	55.56	71.43	**1.00**	44.44	61.54	88.89	**88.89**	**88.89**	**1.00**	77.78	87.50
Qatari	**89.66**	56.52	69.33	89.19	71.74	79.52	87.80	78.26	82.76	88.00	**95.65**	**91.67**	89.13	89.13	89.13
Amer.	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	50.00	33.33	**40.00**	00.00	00.00	00.00
UD-NH	87.30	50.00	63.58	88.35	**82.73**	**85.45**	87.88	79.09	83.25	85.19	83.64	84.40	**88.78**	79.09	83.65
Females	89.22	80.11	84.42	**91.28**	**95.70**	**93.44**	90.58	93.01	91.78	87.82	93.01	90.34	90.77	95.16	92.91
Males	**96.55**	45.90	62.22	93.18	67.21	78.10	78.33	77.05	77.69	79.03	**80.33**	79.67	94.12	78.69	**85.71**
UD-GH	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00
Micro	88.54	56.35	68.87	**89.38**	79.59	84.20	86.56	80.95	83.66	86.03	**84.05**	85.03	89.37	82.97	**86.05**
Macro	64.73	32.01	40.84	77.92	58.93	66.14	69.46	60.54	63.75	77.54	70.98	**73.15**	**78.68**	64.29	69.64

Table 7: Complete Results for Hate Target Identification (Task 2).

AraTar				**Arabic Twitter**				**OSACT5**			
AraBERT-b				AraBERT-b				AraBERT-b			
	P	R	F1		P	R	F1		P	R	F1
RH	85.37	86.96	86.15	RH	86.05	77.08	81.32	Race-HS1	72.73	30.77	43.24
EH	78.16	80.95	79.53	EH	87.25	77.39	82.03	Religion-HS2	0.00	0.00	0.00
NH	79.07	79.07	79.07	NH	81.20	86.36	83.70	Ideology-HS3	80.00	12.50	21.62
GH	86.77	68.05	76.28	GH	81.05	72.64	76.62	Social Class-HS5	0.00	0.00	0.00
UDH	49.58	54.63	51.98	OFF	84.75	76.34	80.32	Gender-HS6	70.91	58.65	64.20
CL	91.46	92.05	91.76	Clean	93.08	96.65	94.83	NOT_HS	93.73	98.72	96.16
Micro	85.21	83.79	84.50	Micro			91.14	Micro			92.44
Macro	78.40	76.95	77.46	Macro	85.56	81.08	83.14	Macro	52.89	33.44	37.54

Table 8: Complete Results of AraBERT-b on the AraTar, Arabic Twitter and OSCAT5 corpora.