MultiMinds at MAHED 2025: Multimodal and Multitask Approaches for Detecting Emotional, Hate, and Offensive Speech in Arabic Content

Riddhiman Swanan Debnath
Shahjalal University of Science and Technology, Sylhet

Abdul Wadud Shakib
Metropolitan University, Sylhet

Md Saiful Islam
Athabasca University, Canada

Abstract

This paper describes the MultiMinds team's participation in the MAHED 2025 shared task at ArabicNLP 2025, which targets the detection of hate speech, hope speech, and emotional expression in Arabic content. We addressed two subtasks. For the text-based subtask (Task 2), we experimented with multiple models, including Support Vector Machines with TF-IDF and AraBERT embeddings, XGBoost with fused AraBERT and XLM-RoBERTa embeddings optimized via Optuna, and a fine-tuned AraBERT model and GPT-5 (gpt-oss-20b). The fine-tuned AraBERT achieved the best performance with an F1 score of 0.68. For the multimodal subtask (Task 3), we proposed an architecture combining DistilBERT for text representation with a lightweight ELU-Net enhanced by a cross-attention mechanism, reaching 75% accuracy. Major challenges included dataset imbalance and noisy text, which we mitigated through preprocessing, class-weighted optimization, and feature fusion. Our results demonstrate the benefits of combining multiple embedding layers for text classification and leveraging lightweight multimodal architectures for robust hate speech detection in Arabic.

1 Introduction

Online media has become an important avenue for the consumption and distribution of information, and many people now rely on it as their primary source of news (Perrin, 2015). These have enabled individuals to share their views effortlessly through images and texts (multimodal and/or unimodal), reaching a broad and diverse audience (Fortuna and Nunes, 2018). With the rapid increase in media posts, manual detection of emotion, hate, and offensive (EHO) content becomes impractical. Consequently, there is a growing interest in developing automated methods for EHO detection.

MAHED 2025 (Zaghouani et al., 2025) is a shared task at ArabicNLP 2025 Co-located with EMNLP 2025, focusing on the detection of hate speech, hope speech, and emotional expression in Arabic content. Participants may choose to participate in one or more of the following three subtasks:(i) Text-based Hate and Hope Speech Classification, (ii) Emotion, Offensive, and Hate Detection (Multitask), and (iii) Multimodal Hateful Meme Detection. We, MultiMinds, participated in MAHED 2025, with particular interest in tasks (ii), (iii) and ranked 10th and 7th, respectively.

For Task (ii), three methods were tested for Arabic emotion, offensive, and hate-speech classification: Support Vector Machines (SVM) as a Baseline model with TF-IDF (best macro F1: 0.517); XGBoost with TF-IDF, AraBERT embeddings, and fused AraBERT and XLM-RoBERTa embeddings, which were optimized via Optuna (best F1: 0.57); and a deep learning approach fine-tuning AraBERT, which achieved the highest performance score. As the dataset was imbalanced and contained unnecessary information, the key challenge was to extract the correct information from the text. In our experiment for Task (iii), we used 1D-CNN model (Singh et al., 2021) as the Baseline model by extracting image and caption features by CLIP processor. Our enhanced ELU-Net architecture got the best results by incorporating a cross-attention mechanism to combine visual and textual features generated from the DistilBert (Sanh et al., 2019) tokenizer. Full Implementation here - Github. The main challenge of this task was that the classes were not equally distributed. Our key findings were as follows.

- Fusing multiple embedding layers from different textual models improves data representation.
- Using class weights enhances results.
- First-time use of a lightweight multimodal model to classify hateful and non-hateful memes.

Proceedings of The Third Arabic Natural Language Processing Conference, pages 677–682

2 Background

2.1 Emotion Detection

In recent years, research into developing state-of-the-art models for Arabic natural language processing tasks has gained momentum. Alswaidan and Menai (2020) proposed three models for emotion recognition in Arabic text. Abdullah et al. (2018) described their system - SEDAT, and showed substantial improvements in Spearman correlation scores over the baseline models. Alsmearat et al. (2015) explored the Gender Identification(GI) problem for Arabic text as a supervised learning problem and compared the Bag-Of-Words (BOW) approach with computing features related to sentiments and emotions. Biswas and Zaghouani (2025b) introduces a bilingual dataset comprising 23,456 entries for Arabic and 10,036 entries for English, annotated for emotions and hope speech, addressing the scarcity of multi-emotion (Emotion and hope) datasets. Al-Henaki et al. (2025) introduced MultiProSE, an open-source extension of the existing Arabic propaganda dataset, ArPro, with the addition of sentiment and emotion annotations for each text.

2.2 Offensive And Hate Speech Detection

While social media promotes free expression, it also fosters environments where hate speech spreads, making its detection a key research priority. Alsafari et al. (2020) built a reliable Arabic textual corpus by crawling data from Twitter. Mubarak et al. (2023) introduced a generic, language-independent method to collect a large percentage of offensive and hate tweets. Aldjanabi et al. (2021) developed a classification system for determining offensive and hate speech using a pre-trained Arabic language model. Biswas and Zaghouani (2025a) introduces multilabel hate speech dataset with offesnive content in the Arabic language. Zaghouani et al. (2024) analyzes 70,000 Arabic tweets, from which 15,965 tweets were selected and annotated, to identify hate speech patterns and train classification models.

2.3 MultiModal Hate Speech Detection

The usage of social media has enabled individuals to disseminate hateful messages through the use of memes. Chhabra and Vishwakarma (2023) highlighted handcrafted feature-based and deep learning-based algorithms by considering multimodal and multilingual inputs. Alam et al. (2024a) explored the intersection between propaganda and hate in memes using a multi-agent LLM-based approach. El-Sayed and Nasr (2024) described an approach to hateful meme classification for the Multimodal Hate Speech Shared Task at CASE 2024. Arya et al. (2024) introduced a novel approach by leveraging the CLIP model, fine-tuned through the incorporation of prompt engineering. Alam et al. (2024b) focused on developing an Arabic memes dataset with manual annotations of propagandistic content. AlDahoul and Zaki (2025) explores the potential of large language models to effectively identify hope, hate speech, offensive language, and emotional expressions. Kmainasi et al. (2025) introduced MemeIntel, an explanation-enhanced dataset for propaganda memes in Arabic and hateful memes in English. However, multimodal hate speech detection lacks the use of lightweight architectures.

3 System Overview

Before tackling Task 2, we observed that the dataset (Zaghouani et al., 2024), (Biswas and Zaghouani, 2025b), (Biswas and Zaghouani, 2025a) was both imbalanced and noisy. To address the noise, we performed text cleaning and preprocessing, converting the text into TF–IDF features and tokenizing it using the AraBERT tokenizer. We then fused the embedding layers of XLM-RoBERTa (Conneau et al., 2019) and AraBERT (Antoun et al., 2020). Furthermore, to mitigate the impact of class imbalance, we incorporated class distribution-based weighting. For preprocessing, we compiled Arabic and English punctuation, removed Arabic diacritics via regex [1], eliminated repeated characters, English words, and numbers, and collapsed multiple spaces into one for clean tokenization. Arabic characters were standardized to reduce variations, ensuring a consistent representation of letters that look or sound similar; for example, different forms of Alif (ا ،آ ،أ ،إ) were replaced with the standard form ا (U+0627).

For feature extraction, we used TF-IDF (Jalilifard et al., 2021) with the top 5,000 terms (unigrams and bigrams). AraBERT and XLM-RoBERTa embeddings were integrated with a 128-token limit, applying padding and truncation, and extracting the [CLS] token from the final hidden state. To fine-tune GPT-5 (Daniel Han and team, 2023), we employ LoRA adapters within the PEFT

[1]https://docs.python.org/3/howto/regex.html

framework, incorporating a curated set of few-shot examples.

For Task 3, we employed the CLIP via Radford et al. (2021) processor for feature extraction, utilizing the ViT-B/32 [2] transformer architecture as the image encoder and a masked self-attention transformer as the text encoder. The extracted multimodal features were fed into a Support Vector Machine for classification; it failed to identify hateful memes accurately. The main challenge was dataset (Alam et al., 2024a), (Alam et al., 2024b) imbalance, which could be mitigated by collecting more hateful memes for a balanced distribution. Additionally, as non-Arabic speakers, understanding the language and cultural context was difficult, so we relied on a CNN-based neural network for better performance.

To achieve our objective of developing a lightweight model, we employed the ELUNet architecture via Deng et al. (2022). Since all captions in the dataset are in the Arabic language, textual features were extracted using the DistilBERT tokenizer via Devlin et al. (2018). In the case of preprocessing and cleaning, the same procedure as Task 2 was followed. Another challenge we faced was that the tokenizers' lengths were not equal for all memes, as they hold different sizes of text. So we fixed the tokenizer size to 256. If the tokenizer length is smaller than the value, the previous value will repeat; otherwise larger size tokenizer will be shrunk using the PCA algorithm (Drikvandi and Lawal, 2023). The corresponding images were processed through the encoder component of the ELUNet architecture. Inspired by Li et al. (2024), a cross-attention mechanism was then applied, integrating the encoded image features from the encoder with the textual embeddings generated by the tokenizer, positioned at the intermediate layers of ELUNet. The cross-attention outputs were subsequently passed through the decoder component of ELUNet. The proposed model (Figure 1) produces two outputs.

4 Experimental Setup

4.1 Emotion, Offensive Language, and Hate Detection

The whole dataset was split into Train(70%), Test(15%), and Validation(15%) via stratified sampling across emotion, hate, and offensive tasks,

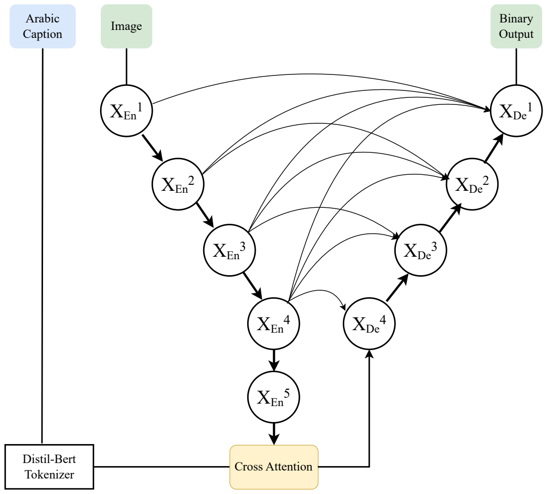

Figure 1: The architecture of Attention-based ELUNet

with exception for GPT-5 (80-10-10). Table 1 provides a brief overview of various emotions in the dataset, including its size and distribution of various emotions, as well as there are offensive (yes - 1744, no - 4216) and hate (yes - 303, no - 1441). Table 2 reveals the Task 2 dataset contains the most non-Arabic characters (see Figure 2).

Name	Amount
Anger	1551
Disgust	777
Neutral	661
Love	593
Joy	533
Anticipation	491
Optimism	419
Sadness	335
Confidence	210
Pessimism	194
Surprise	143
Fear	53

Table 1: Emotion Proportions in Training Data – Task 2

We used Optuna with a class-weighted objective to optimize XGBoost hyperparameters for the highest macro F1-score. We incorporated a deep learning approach using AraBERTv2 [3] for multitask classification across emotion, offensive language, and hate speech tasks. Three task-specific linear layers mapped the 768-dimensional hidden representation to class logits, with dropout applied to improve generalization. For fine-tuning GPT-5, we

[2] https://huggingface.co/openai/clip-vit-base-patch32

[3] https://huggingface.co/aubmindlab/bert-base-arabertv2

Name	Non-Arabic Chars Count
Train (Task 2)	157138
Test	32968
Validation	32075
Train (Task 3)	4737
Test	1340
Validation	1310

Table 2: Non-Arabic Characters in Tasks 2 & 3

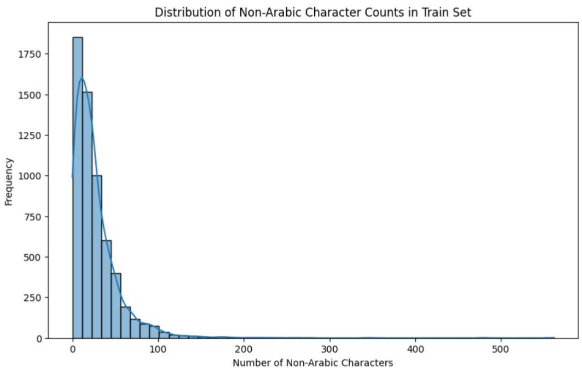

Figure 2: Non-Arabic Character Distribution – Train Set (Task 2)

configured the rank, selected specific transformer layers, and applied an appropriate scaling factor, while enabling gradient checkpointing to optimize memory usage. Furthermore, no bias parameters were introduced to ensure that the fine-tuning process remained lightweight.

	Emo	**Offn**	**Hate**
learning rate	0.0060	0.0011	0.0037
max depth	10	7	10
num. estimator	50	282	182
subsample	0.9453	0.8231	0.7524
colsample_bytree	0.7366	0.6489	0.8440
scale_pos_weight	**x**	2.4179	0.0535

Table 3: Best parameter value from trial run

4.2 Multimodal Hate Speech Detection in Memes

Table 4 presents the distribution of hateful content in training, development, and test sets. We processed each meme (text + image) using CLIP to create joint features. Text was tokenized and images scaled to RGB via CLIPProcessor, producing tensors for both modalities. Features were concatenated and fed to a 1D-CNN. Then we evaluated our enhaced ELUNet model with AraBert, DistilBert tokenizers. Our best model, ELUNet with the DistilBert tokenizer gave the accuracy of 75%. In our experiment, we chose batch size 16, epoch 5, and learning rate 10^{-3}. This model was trained in Google Colab and consumed 6.2 GB of GPU.

Name	**Hate**	**Not Hate**
Train	213	1930
Dev	31	281
Test	154	452
Total	398	2663

Table 4: Dataset Size – Task 3 (Initial)

5 Results

Table 5 summarizes our model's performance on the task 2 dataset. The results indicate that applying class weights improves performance based on the average F1 score, while incorporating deep learning approaches yields even higher results. For instance, in our experiments with AraBERT, using a batch size of 8, 5 epochs, a dropout rate of 0.3, and a learning rate of 10^{-5} with the exception (10^{-4}) for Gpt-5, we achieved an F1 score of 0.67. Reducing dropout to 0.1, while doubling both batch size and epochs, increased the score to 0.68, matching the performance of DistilBERT. However, with respect to accuracy, GPT-5 and AraBERT achieved comparable performance on the offensive and hate detection tasks, while exhibiting notable differences in the emotion classification task.

App.	**Model**	**Emo**	**Offn**	**Hate**	**Avg**
Without Weight	XGB	0.172	0.416	0.344	0.312
	XGB-AraBERT	0.241	0.712	0.541	0.484
	XGB-AraBERT+XLMRoBERTa	0.244	0.414	0.500	0.384
	SVM(Baseline)	0.284	0.702	0.564	**0.513**
With Weight	XGB	0.212	0.712	0.400	0.393
	XGB-AraBERT+XLMRoBERTa	0.264	0.723	0.500	0.493
	XGB-AraBERT+XLMRoBERTa Trial	0.324	0.775	0.624	**0.574**
DL	AraBERT	0.267	0.834	0.954	**0.684**
	DistilBERT	**0.373**	0.774	0.924	**0.683**
	Gpt-oss-20b (PC)	0.014	0.412	0.483	0.300

Table 5: Performance of the models on the Task 2 dataset. Here, PC, Emo, Offn, Hate, and Avg denote the post-competition, emotion, offensive, hate, and average macro F1 scores, respectively.

The model performances in Task 3 are described in Table 6. For adding class weight, the result has been improved. Finally, we get an accuracy of 75%. For each testing section test dataset was utilized. Despite fixing the epoch to 20, the best-fitting model took only 5 epochs by using the early stopping concept.

Model	Acc	MacroAvg-f1	Hateful(f1)	Non-Hateful(f1)
1D-CNN(Baseline)	0.745	0.431	0	0.851
ELUNet-DistilBert	0.746	0.421	0	0.852
ELUNet-AraBert	0.744	0.422	0	0.853
ELUNet-AraBert (WW)	0.746	0.372	0	0.855
ELUNet-DistilBert(WW)	0.754	0.500	0.165	0.858

Table 6: Performance of the models on the Task 3 dataset. Here, WW represents 'with weight'.

6 limitations

Both subtasks (Task 2: Emotion, Offensive, and Hate Detection; Task 3: Multimodal Hateful Meme Detection) suffered from severe class imbalance. This led to biased models, poor performance on minority classes, and necessitated mitigations such as class weighting, which still did not fully resolve the issue. Fine-tuning was limited (e.g., 5 epochs with early stopping, a fixed tokenizer length of 256, and PCA for shrinkage), which may have led to underfitting. GPT-5 experiments were constrained by few-shot examples and memory optimizations (e.g., LoRA adapters), resulting in lower emotion detection scores (F1=0.014).

7 Conclusion

Our participation in MAHED 2025 highlights the effectiveness of advanced NLP and multimodal methods for detecting hate speech, hope speech, and emotions in Arabic. For Task 2, our fine-tuned AraBERT scored 0.68 macro F1, surpassing SVM and XGBoost baselines through class-weighted optimization and fused embeddings to address imbalance and noise. For Task 3, our lightweight ELU-Net, cross-attention with tokenizer generated from DistilBert, achieved 75 % accuracy on hateful meme classification despite imbalance. Challenges included limited Arabic meme data, non-Arabic characters, and noisy text affecting preprocessing and features. Future work will explore data augmentation, advanced multimodal fusion, and improved preprocessing and fine-tuning to boost robustness and generalization.

References

Malak Abdullah, Mirsad Hadzikadicy, and Samira Shaikhz. 2018. Sedat: sentiment and emotion detection in arabic text using cnn-lstm deep learning. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)*, pages 835–840. IEEE.

Lubna Al-Henaki, Hend Al-Khalifa, Abdulmalik Al-Salman, Hajar Alqubayshi, Hind Al-Twailay, Gheeda Alghamdi, and Hawra Aljasim. 2025. Multiprose: A multi-label arabic dataset for propaganda, sentiment, and emotion detection. In *International Conference on Applications of Natural Language to Information Systems*, pages 156–172. Springer.

Firoj Alam, Md Rafiul Biswas, Uzair Shah, Wajdi Zaghouani, and Georgios Mikros. 2024a. Propaganda to hate: A multimodal analysis of arabic memes with multi-agent llms. In *International Conference on Web Information Systems Engineering*, pages 380–390. Springer.

Firoj Alam, Abul Hasnat, Fatema Ahmad, Md. Arid Hasan, and Maram Hasanain. 2024b. ArMeme: Propagandistic content in Arabic memes. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21071–21090, Miami, Florida, USA. Association for Computational Linguistics.

Nouar AlDahoul and Yasir Zaki. 2025. Detecting hope, hate, and emotion in arabic textual speech and multimodal memes using large language models. *arXiv preprint arXiv:2508.15810*.

Wassen Aldjanabi, Abdelghani Dahou, Mohammed AA Al-Qaness, Mohamed Abd Elaziz, Ahmed Mohamed Helmi, and Robertas Damaševičius. 2021. Arabic offensive and hate speech detection using a cross-corpora multi-task learning model. In *Informatics*, volume 8, page 69. MDPI.

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

Kholoud Alsmearat, Mohammed Shehab, Mahmoud Al-Ayyoub, Riyad Al-Shalabi, and Ghassan Kanaan. 2015. Emotion analysis of arabic articles and its impact on identifying the author's gender. In *2015 IEEE /ACS 12th International Conference of Computer Systems and Applications (AICCSA)*, pages 1–6. IEEE.

Nourah Alswaidan and Mohamed El Bachir Menai. 2020. Hybrid feature model for emotion recognition in arabic text. *IEEE Access*, 8:37843–37854.

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

Greeshma Arya, Mohammad Kamrul Hasan, Ashish Bagwari, Nurhizam Safie, Shayla Islam, Fatima Rayan Awad Ahmed, Aaishani De, Muhammad Attique Khan, and Taher M Ghazal. 2024. Multimodal hate speech detection in memes using contrastive language-image pre-training. *IEEE Access*, 12:22359–22375.

Md. Rafiul Biswas and Wajdi Zaghouani. 2025a. An annotated corpus of arabic tweets for hate speech analysis. *CoRR*, abs/2505.11969.

Md. Rafiul Biswas and Wajdi Zaghouani. 2025b. Emohopespeech: An annotated dataset of emotions and hope speech in english and arabic. *CoRR*, abs/2505.11959.

Anusha Chhabra and Dinesh Kumar Vishwakarma. 2023. A literature survey on multimodal and multilingual automatic hate speech identification. *Multimedia Systems*, 29(3):1203–1230.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Michael Han Daniel Han and Unsloth team. 2023. Unsloth.

Yunjiao Deng, Yulei Hou, Jiangtao Yan, and Daxing Zeng. 2022. Elu-net: An efficient and lightweight u-net for medical image segmentation. *IEEE Access*, 10:35932–35941.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Reza Drikvandi and Olamide Lawal. 2023. Sparse principal component analysis for natural language processing. *Annals of data science*, 10(1):25–41.

Ahmed El-Sayed and Omar Nasr. 2024. Aast-nlp at multimodal hate speech event detection 2024: A multimodal approach for classification of text-embedded images based on clip and bert-based models. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 139–144.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *Acm Computing Surveys (Csur)*, 51(4):1–30.

Amir Jalilifard, Vinicius Fernandes Caridá, Alex Fernandes Mansano, Rogers S Cristo, and Felipe Penhorate Carvalho da Fonseca. 2021. Semantic sensitive tf-idf to determine word relevance in documents. In *Advances in Computing and Network Communications: Proceedings of CoCoNet 2020, Volume 2*, pages 327–337. Springer.

Mohamed Bayan Kmainasi, Abul Hasnat, Md Arid Hasan, Ali Ezzat Shahroor, and Firoj Alam. 2025. Memeintel: Explainable detection of propagandistic and hateful memes. *arXiv preprint arXiv:2502.16612*.

Hongchan Li, Yantong Lu, and Haodong Zhu. 2024. Multi-modal sentiment analysis based on image and text fusion based on cross-attention mechanism. *Electronics*, 13(11):2069.

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.

Andrew Perrin. 2015. Social media usage. *Pew research center*, 125:52–68.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Kuljeet Singh, Amit Mahajan, and Vibhakar Mansotra. 2021. 1d-cnn based model for classification and analysis of network attacks. *International Journal of Advanced Computer Science and Applications*, 12(11):604–613.

Wajdi Zaghouani, Md Rafiul Biswas, Mabrouka Bessghaier, Shimaa Ibrahim, Georgio Mikros, and Firoj Alam. 2025. Overview on mahed 2025 shared task: Multimodal detection of hope and hate emotions in arabic content. In *The Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou. Association for Computational Linguistics.

Wajdi Zaghouani, Hamdy Mubarak, and Md. Rafiul Biswas. 2024. So hateful! building a multi-label hate speech annotated Arabic dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15044–15055, Torino, Italia. ELRA and ICCL.