A Deep Learning Approach for Identification of Arabic Misogyny from Tweets

Abhinav Kumar[1], Pradeep Kumar Roy[2] and Jyoti Prakash Singh[3]

[1]Department of Computer Science & Engineering,
Siksha 'O' Anusandhan Deemed to be University, Bhubaneswar, India
[2]Department of Computer Science & Engineering,
Indian Institute of Information Technology Surat, India
[3]Department of Computer Science & Engineering,
National Institute of Technology Patna, India

Abstract

Online misogyny has become a major cause of concern for Arab women who face gender-based online abuse regularly. Misogyny is a form of hate speech that denigrates a person or a group that identifies as feminine; it is generally described as hatred or contempt towards women. Arab women exposed to many forms of online misogyny, which sadly reinforces and justifies gender inequality, inferior social standing, sexual assault, violence, maltreatment, and underestimating. This paper proposes three methods to identify and classify misogyny behavior from Arabic tweets: (i) BERT, (ii) Ensemble-based model, and (iii) Dense Neural Network-based model. The suggested approach performs admirably on both tasks. The BERT model outperformed the other two suggested methods for misogyny identification, with an accuracy of 0.883, while the ensemble-based approach outperformed the other two suggested methods for misogyny behavior classification task, with an accuracy of 0.764.

Keywords

Misogyny, Arabic tweets, BERT, Ensemble model, Deep learning

1. Introduction

Online social platforms like Facebook, Twitter, and Instagram are among the popular platforms for spreading the news, sharing the achievement, and connecting with the worldwide community [1, 2, 3, 4, 5]. However, due to freedom of post, many negative contents are floating in high volume every day. Abusive content, Hate Speech [6, 7, 8, 9, 10], Rumour [11], False news [12] are a few of the negative news categories which online users mostly post [13, 14]. Misogyny is a type of hate speech that is used for the female gender [15].

On the internet, misogyny has evolved into a worldwide issue that has spread across several social media platforms [16, 17]. Women in the Arab world, like their counterparts around the globe, are subjected to various types of online misogyny, which unfortunately promotes and excuses gender inequality and violence against women [16, 18, 19]. In the past few years, online misogynistic language flooded on Twitter, Facebook, and other social platforms, and this language consists of sexual abuse, violence, hate speech, and bully content. Online misogyny

Forum for Information Retrieval Evaluation, December 13-17, 2021, India
abhinavanand05@gmail.com (A. Kumar); pradeep.roy@iiitsurat.ac.in (P. K. Roy); jps@nitp.ac.in (J. P. Singh)

may evolve to target a particular female personality to threaten or bullying by launching some campaign on social platforms [17]. Identifying such misogyny content is very much needed to prohibit misogynistic Arabic content on online social platforms. Consequently, it enables the Arab female to use the social platform to express their opinion freely [20, 21, 22].

To identify misogyny and its behavior from the tweets, this paper proposes three different models: (i) Ensemble-based model (Support Vector Machine (SVM) + Logistic Regression (LR)), (ii) Dense Neural Network (DNN) model, and (iii) BERT (bert-base-arabic) [23]. The proposed models are validated with the dataset released in ArMI 2021 (a subtrack of HASOC FIRE2021)[1]. Two sub-tasks were shared by the organizer: (i) Misogyny Content Identification and (ii) Misogyny Behavior Identification. The dataset consists of 7,866 Misogyny tweets labeled into various categories like dominance, damning, harassment, and others. The dataset was developed by collecting the tweets from Twitter during January 2019-2021 and annotated into eight classes. The organizer has proposed two tasks [24] on this topic: (i) Task 1 is a binary classification task where tweets have categorized either misogyny or not, and (ii) Task 2 is a multi-class classification where tweets are labeled into eight categories- seven misogyny categories and the last one is from none misogyny category. The categories are (i) Damning (Damn): Tweets containing cursing content fall under this category, (ii) Derailing (Der): This category includes tweets that justify women's violence or mistreatment, (iii) Discredit (Disc): Slurs and insulting words directed towards women can be found in tweets in this category. (iv) Dominance (Dom): Tweets in this category imply that men are superior to women, (v) Sexual Harassment (Harass): Sexual approaches and sexual nature abuse are discussed in tweets in this category, (vi) Stereotyping Objectification (Obj): Tweets in this category promote a stereotypical picture of women or describe their physical attractiveness, (vii) Threat of Violence (Vio): The content of tweets in this category is frightening, with threats of physical violence, (viii) None: if no misogynistic behaviors exist.

The rest of the paper is organized as follows: Section 2 discusses the proposed methodology in detail, the findings of the proposed system are listed in Section 3 and finally, the paper is concluded in Section 4.

2. Methodology

The systematic diagram for the proposed model for misogyny comment and behavior identification can be seen in Figure 1. Three different models were proposed: (i) Ensemble-based model (Support Vector Machine (SVM) + Logistic Regression (LR)), (ii) Dense Neural Network (DNN) model, and (iii) BERT (bert-base-arabic) [23]. The models were trained with the dataset published on the HASOC-ArMI 2021 track[2]. The overall statistic of the dataset can be seen in Table 1.

[1]http://fire.irsi.res.in/fire/2021/hasoc
[2]https://sites.google.com/view/armi2021/

Table 1
Data statistic used to validate the proposed system

Class	Class	Number of samples	Total
None	None	3,061	3,061
Misogyny	Discredit	2,868	4,805
	Damning	669	
	Stereotyping & objectification	653	
	Threat of violence	230	
	Dominance	219	
	Derailing	105	
	Sexual harassment	61	

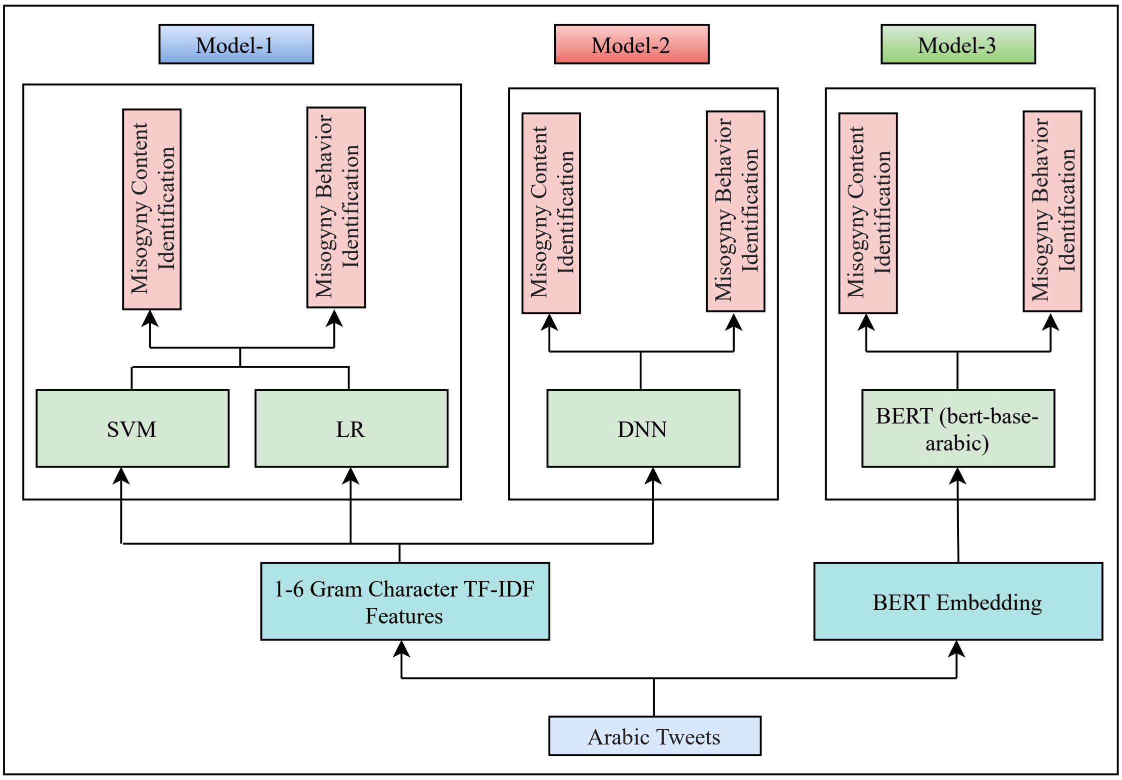

Figure 1: Proposed model for the misogyny identification and behaviour classification

2.1. Ensemble-based model (SVM + LR):

Extensive experiments were carried out to determine the best-suited n-gram range for the ensemble-based model by varying the character n-gram range from one to six. Among different combinations of the n-gram range, we found that one to six character N-grams TF-IDF features performed best. Therefore, in the proposed ensemble-based model, one to six character N-grams

Table 2
Results for the different models for task1 with validation data

	Class	Precision	Recall	F_1-score	Accuracy
SVM	Misogyny	0.88	0.91	0.89	0.87
	None	0.85	0.81	0.83	
	Macro Avg.	0.87	0.86	0.86	
LR	Misogyny	0.86	0.91	0.88	0.86
	None	0.85	0.77	0.81	
	Macro Avg.	0.85	0.84	0.85	
DNN	Misogyny	0.88	0.88	0.88	0.86
	None	0.82	0.82	0.82	
	Macro Avg.	0.85	0.85	0.85	
BERT (bert-base-arabic)	Misogyny	0.90	0.91	0.91	0.88
	None	0.86	0.84	0.85	
	Macro Avg.	0.88	0.88	0.88	

TF-IDF features were used by SVM and LR classifiers to predict the probabilities for each of the output classes. The class-wise output probabilities of both the models were averaged. Finally, the higher average probability is used to decide the final class label. The overall flow diagram of the ensemble-based model can be seen in Figure 1.

2.2. Dense Neural Network (DNN) model:

The proposed dense neural network-based model uses one to six character N-grams TF-IDF features as an input to the network. The TF-IDF features are then passed through a four-layered dense network containing 4,096, 512, 64, and 2-neurons. As the performance of deep learning models is very sensitive to the chosen hyper-parameters, extensive experiments were performed by varying learning rate, batch size, dropout rate, optimizer, loss function, and the number of epochs. The best-suited hyper-parameters were found with a learning rate of 0.001, batch size of 20, the dropout rate of 0.2, Adam as the optimizer, binary cross-entropy for misogyny identification, and categorical cross-entropy for misogyny behavior classification. The model was trained for 50 epochs for the final prediction. The detailed flow diagram of the proposed dense neural network-based model can be seen in Figure 1.

2.3. BERT (bert-base-arabic):

The BERT (bert-base-arabic) pretrained on approx 8.2 Billion Arabic words [23]. The corpus and vocabulary set are not limited to Modern Standard Arabic; dialectical Arabic is included as well. To train BERT (bert-base-arabic), the pretraining method is similar to that of BERT, with the following exceptions: Instead of 1M training steps with a batch size of 256, 3M training steps with a batch size of 128 were trained. To fine-tune the BERT (bert-base-arabic) on our dataset, we set the maximum length of tweets to 30 because we observed that most tweets are less than 30 words in length. It indicates that tweets with less than 30 words were padded with zero, while tweets with more than 30 words were curtailed out. The experiment was performed

Table 3
Results for the different models for task2 with validation data

	Class	Precision	Recall	F_1-score	Accuracy
SVM	Damning	0.86	0.55	0.67	0.73
	Derailing	0.00	0.00	0.00	
	Discredit	0.70	0.77	0.73	
	Dominance	0.78	0.29	0.42	
	None	0.74	0.91	0.82	
	Sexual harassment	0.00	0.00	0.00	
	Stereotyping & objectification	0.82	0.49	0.61	
	Threat of violence	0.50	0.08	0.14	
	Macro Avg.	0.55	0.39	0.42	
LR	Damning	0.85	0.55	0.67	0.73
	Derailing	0.00	0.00	0.00	
	Discredit	0.71	0.79	0.75	
	Dominance	0.67	0.17	0.27	
	None	0.72	0.91	0.81	
	Sexual harassment	0.00	0.00	0.00	
	Stereotyping & objectification	0.81	0.46	0.59	
	Threat of violence	1.00	0.08	0.15	
	Macro Avg.	0.59	0.37	0.40	
DNN	Damning	0.90	0.56	0.69	0.74
	Derailing	0.00	0.00	0.00	
	Discredit	0.73	0.77	0.75	
	Dominance	0.53	0.42	0.47	
	None	0.78	0.88	0.82	
	Sexual harassment	0.00	0.00	0.00	
	Stereotyping & objectification	0.71	0.62	0.66	
	Threat of violence	0.50	0.40	0.44	
	Macro Avg.	0.52	0.46	0.48	
BERT (bert-base-arabic)	Damning	0.80	0.75	0.77	0.77
	Derailing	0.07	0.10	0.08	
	Discredit	0.77	0.79	0.78	
	Dominance	0.57	0.54	0.55	
	None	0.83	0.87	0.85	
	Sexual harassment	0.00	0.00	0.00	
	Stereotyping & objectification	0.77	0.65	0.71	
	Threat of violence	0.60	0.36	0.45	
	Macro Avg.	0.55	0.51	0.52	

with a learning rate of 2e-5, a batch size of 32, and it is trained for 50 epochs.

3. Results

The performance of the proposed models is measured in terms of accuracy, macro precision, recall, and F_1-score. The experimentation was first performed by taking 10% data samples from

Table 4
Results on the testing dataset for the submitted models

Task	Models	Run	Accuracy	Macro Precision	Macro Recall	Macro F_1-score
Task-1	BERT (bert-base-arabic)	1	0.883	0.878	0.876	0.877
	Ensemble (SVM+LR)	2	0.873	0.868	0.865	0.866
	DNN	3	0.854	0.846	0.850	0.848
Task-2	Ensemble (SVM+LR)	2	0.764	0.676	0.480	0.531
	DNN	3	0.745	0.559	0.508	0.526
	BERT (bert-base-arabic)	1	0.780	0.549	0.502	0.519

the provided training dataset. The results of different models with the validation dataset for misogyny identification are listed in Table 2. For the validation dataset, BERT (bert-base-arabic) model performed best with an accuracy of 0.88, macro precision, recall, and F_1-score of 0.88 (as can be seen in Table 2). The results of the different models for misogyny behavior classification are listed in Table 3. Again, the proposed BERT(bert-base-arabic) model performed best with an accuracy of 0.77, macro precision of 0.55, recall of 0.51, and F_1-score of 0.52, respectively.

In the HASOC-ArMI-2021 track, we had submitted three models (i) Ensemble-based model (SVM + LR), (ii) Dense Neural Network (DNN), and (iii) BERT (bert-base-arabic) for the final evaluation. The result of these models for misogyny identification and behavior classification are listed in Table 4. The submitted models performed significantly well for both tasks. The BERT (bert-base-arabic) performed best for misogyny identification among all the three submitted models with an accuracy of 0.883, macro precision of 0.878, recall of 0.876, and F_1-score of 0.877. The ensemble-based model achieved an accuracy of 0.873, macro precision of 0.868, recall of 0.865, and F_1-score of 0.866. The Dense neural network-based model achieved an accuracy of 0.854, macro precision of 0.846, recall of 0.850, and F_1-score of 0.848.

For misogyny behavior classification, the proposed ensemble-based (SVM + LR) performed best among all the submitted models with an accuracy of 0.764, macro precision of 0.676, recall of 0.480, and F_1-score of 0.531. The dense neural network-based model achieved an accuracy of 0.745, macro precision of 0.559, recall of 0.508, F_1-score of 0.526. The BERT (bert-base-arabic) model achieved an accuracy of 0.780, macro precision of 0.549, recall of 0.503, and F_1-score of 0.519 (as can be seen in Table 4).

4. Conclusion

Misogynistic language on social media sites such as Facebook and Twitter has been highlighted as a global issue that has increased over the last decade. In this paper, we proposed three different models such as BERT, ensemble-based model, and dense neural network-based model for the identification of misogyny and classification of misogyny behavior. We explored the role of character-level features and found that the use of character-level features from Arabic tweets shows promising performance in the misogyny identification and misogyny behavior classification from the tweets. The proposed fine-tuned BERT model performed best with an accuracy of 0.883 for the misogyny identification task whereas the proposed ensemble-based model form best with an accuracy of 0.764 for the misogyny behavior classification task.

References

[1] A. Kumar, N. C. Rathore, Relationship strength based access control in online social networks, in: Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 2, Springer, 2016, pp. 197–206.

[2] R. Pandey, A. Kumar, J. P. Singh, S. Tripathi, Hybrid attention-based long short-term memory network for sarcasm identification, Applied Soft Computing 106 (2021) 107348.

[3] A. Kumar, J. P. Singh, Location reference identification from tweets during emergencies: A deep learning approach, International journal of disaster risk reduction 33 (2019) 365–375.

[4] A. Kumar, J. P. Singh, Y. K. Dwivedi, N. P. Rana, A deep multi-modal neural network for informative twitter content classification during emergencies, Annals of Operations Research (2020) 1–32.

[5] A. Kumar, J. P. Singh, S. Saumya, A comparative analysis of machine learning techniques for disaster-related tweet classification, in: 2019 IEEE R10 Humanitarian Technology Conference (R10-HTC)(47129), IEEE, 2019, pp. 222–227.

[6] A. Kumar, S. Saumya, J. P. Singh, Nitp-ai-nlp@ hasoc-dravidian-codemix-fire2020: A machine learning approach to identify offensive languages from dravidian code-mixed text., in: FIRE (Working Notes), 2020, pp. 384–390.

[7] A. Kumar, S. Saumya, J. P. Singh, Nitp-ai-nlp@ hasoc-fire2020: Fine tuned bert for the hate speech and offensive content identification from social media., in: FIRE (Working Notes), 2020, pp. 266–273.

[8] R. Jain, D. Goel, P. Sahu, A. Kumar, J. P. Singh, Profiling hate speech spreaders on twitter, in: CLEF, 2021.

[9] S. Saumya, A. Kumar, J. P. Singh, Offensive language identification in dravidian code mixed social media text, in: Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, 2021, pp. 36–45.

[10] G. Kumar, J. P. Singh, A. Kumar, A deep multi-modal neural network for the identification of hate speech from social media, in: Conference on e-Business, e-Services and e-Society, Springer, 2021, pp. 670–680.

[11] J. P. Singh, A. Kumar, N. P. Rana, Y. K. Dwivedi, Attention-based lstm network for rumor veracity estimation of tweets, Information Systems Frontiers (2020) 1–16.

[12] A. Kumar, S. Saumya, J. P. Singh, Nitp-ai-nlp@ urdufake-fire2020: Multi-layer dense neural network for fake news detection in urdu news articles., in: FIRE (Working Notes), 2020, pp. 458–463.

[13] P. K. Roy, A. K. Tripathy, T. K. Das, X.-Z. Gao, A framework for hate speech detection using deep convolutional neural network, IEEE Access 8 (2020) 204951–204962.

[14] P. K. Roy, S. Chahar, Fake profile detection on social networking websites: A comprehensive review, IEEE Transactions on Artificial Intelligence 1 (2020) 271–285.

[15] M. E. Moloney, T. P. Love, Assessing online misogyny: Perspectives from sociology and feminist media studies, Sociology Compass 12 (2018) e12577.

[16] B. Poland, Haters: Harassment, abuse, and violence online, U of Nebraska Press, 2016.

[17] M. Ferrier, N. Garud-Patkar, Trollbusters: Fighting online harassment of women journalists, in: Mediating Misogyny, Springer, 2018, pp. 311–332.
[18] M. Amjad, N. Ashraf, A. Zhila, G. Sidorov, A. Zubiaga, A. Gelbukh, Threatening language detection and target identification in Urdu tweets, IEEE Access 9 (2021) 128302–128313.
[19] M. Amjad, G. Sidorov, A. Zhila, A. F. Gelbukh, P. Rosso, Overview of the shared task on fake news detection in Urdu at FIRE 2020., in: FIRE (Working Notes), 2020, pp. 434–446.
[20] H. Mulki, B. Ghanem, Let-mi: An arabic levantine twitter dataset for misogynistic language, in: Proceedings of the Sixth Arabic Natural Language Processing Workshop, 2021, pp. 154–163.
[21] M. Amjad, N. Ashraf, G. Sidorov, A. Zhila, L. Chanona-Hernandez, A. Gelbukh, Automatic abusive language detection in Urdu tweets, ACTA POLYTECHNICA HUNGARICA (2021).
[22] M. Amjad, A. Zhila, G. Sidorov, A. Labunets, S. Butt, H. I. Amjad, O. Vitman, A. Gelbukh, UrduThreat@ FIRE2021: Shared track on abusive threat identification in Urdu, in: FIRE 2021: Forum for Information Retrieval Evaluation, December 13-17, 2021, India, ACM, 2021.
[23] A. Safaya, M. Abdullatif, D. Yuret, KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 2054–2059. URL: https://www.aclweb.org/anthology/2020.semeval-1.271.
[24] H. Mulki, B. Ghanem, ArMI at FIRE2021: Overview of the First Shared Task on Arabic Misogyny Identification, in: Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation, CEUR, 2021.