EasyChair Preprint

№ 15865

Offensive Language Detection in Arabic Social Media Using Machine Learning With TF-IDF Technique

Saleem Abu Lehyeh, Mahmoud Omari, Fatima Shannag and Ghaith Jaradat

February 24, 2025

Offensive Language Detection in Arabic Social Media Using Machine Learning With TF-IDF Technique

Saleem Abu Lehyeh [1, a)], Mahmoud Omari [2, b)], Fatima Shannag [3, c)], and Ghaith M. Jaradat [4, d)]

[1,4] College of Computer Sciences and Informatics, Amman Arab University
[2] Department of Computer Science, Engineering, and Mathematics
University of South Carolina Aiken
[3] College of Computer & Information Sciences, Prince Sultan University

a) *Saleemn880@gmail.com*

b) momari@usca.edu

c) falshannag@psu.edu.sa

d)g.jaradat@aau.edu.jo

Abstract. Our lives now revolve around social communication, and because Arabic text is so complicated and includes so many dialects, it can be difficult to identify offensive language in Arabic social media. This paper examines the implementation of machine learning models. A chosen classifier is used, including a decision tree, support vector machine, random forest, and logistic regression. The "ArCybC" dataset, which contains 4505 tweets, was used in experiments to evaluate how well the machine learning model performed. According to the results of the experiments, using more runs enhances machine learning models' performance, particularly regarding precision and recall rate. With more runs, the Decision tree (DT) and Random Forest (RF) classifiers showed better recall and precision, but the DT classifiers showed better precision.

Keywords: ArCybC, Machine Learning, offensive language

INTRODUCTION

Social media platforms are crucial for the dissemination of knowledge and participation in academics in the communications of science. With improvements to web resources. Considering the development of several digital technologies over the past ten years, it has become simpler to produce online content and connect with academics worldwide. On networking sites created specifically for this purpose, various digital materials, such as books, articles, photos, and videos, have become commonplace tools for professional discourse and information sharing (Zimba & Gasparyan, 2021).

(Khan et al., 2014) Public relations professionals can communicate with the public on social media in a variety of ways. Social media platforms, for example, can reveal details about the beliefs, feelings, plans, actions, and traits of citizens.

While social media use in the public sector has numerous advantages, there are certain disadvantages as well, which raise questions and suspicion. Social media interaction for example presents additional privacy, security, data management, accessibility, and governance issues.

Although there isn't a specific definition of offensive language can be hate speech, abusive language, and describe (Davidson et al., 2017)

Machine language is a process of teaching a machine to maximize a performance metric by utilizing sample data or historical information.

Creating techniques that can automatically identify patterns in data and utilize those patterns to forecast future data or other interesting results is the aim of machine learning (Kauchak, 2016; Tarawneh et al., 2023).

RELATED PREVIOUS STUDIES

(Husain & Uzuner, 2021). Researchers employed many machine-learning models as classifiers in a study to identify offensive language on YouTube in Arabic comments. Different feature extensions and selections are also used. The authors experiment with a wide range of techniques, such as singular value, ExtraTreesclassifiers, feature ranking with recursive feature elimination, logistic regression with L1 regularization, and tree-based ensemble methods. The trained ML classifiers (SVM, NB, DT, and LR) ranked an accuracy of 84%, a precision of 89%, a recall of 76% as well as an F1-score of 81%.

(Al-Badri et al.2022). was utilized to support vector machine the feature of cross-lingual transfer learning. The goal was to find features that work well across different languages. The classifier for this position is deep learning models based on transformers. A multilingual dataset containing offensive tweets written in Spanish, Arabic, and English was used to show that technique. The dataset was labeled with the type of offensive expressions that were generated. the cross-lingual offensive language detection system performed better after incorporating ACO-based characteristics

Al-Harbi et al. (2022). Using chi-square, mutual information, and PSO statistics to choose features from a collection of Arabic tweets classified as offensive or not. The Naïve Bayes classifier served as the foundation to evaluate the model's performance. Based on the findings, the model's accuracy was 95.5%. However, the model outperformed the baseline technique that ignored feature selection and reduced the number of features to 60, possibly retaining only the most useful feature and increasing the model's efficiency.

Shannag et al. (2022) created and evaluated ArCybC- a multi-parlance Arabic Social Media Cyberbullying Corpus for cyberbullying detection and study in the Arabic language. Four groups of 'X' social media (formerly called Twitter) accounts which are more prone to abuses, gaming, sports, news, and celebrity accounts, were determined through their tweets and records were created and they were made into a corpus. These tweets were then filtered with the learned list of Arabic profane phrases. The tweets were divided into 5 groups and each of them was run through a similar annotation procedure by the annotators who labeled the tweets as offensive or non-offensive and as cyberbullying or non-cyberbullying. The effectiveness of the ArCybC was assessed by retraining it with two distinct machine learning techniques which were bag-of-words and word embedding. A Support Vector Machine with a word embedding approach was shown to be the most precise with an overall accuracy of 86.3% and an F1 score of 85%. While the process of producing the ArCybC was such a feat, there were also a lot of difficulties that were encountered. Such difficulties included the rare availability of the required Arabic cyberbullying and the complexity of the annotation of the texts.

ElSherif et al. (2023) Researchers proposed a GA that comprises a contradiction - classifier accuracy (measured by F1-score) and sparsity measure of features (number of selected features) - to accomplish the trade-off balance between model accuracy and explainability. In particular, they used the Support Vector Machine classifier to assess the performance of their model on the dataset comprising more than 40,000 Arabic social media messages classified as either offensive or not. Within the posts, I collected the content from social media platforms, such as Facebook, Instagram, and 'X'. The F1-score of the researchers reached 0.85% which was 2-3% better than information gain and chi-square algorithm-based feature selection techniques. Furthermore, the number of features was reduced to only 25%, a trend that made more comprehensible models available.

PROCEDURE and METHODOLOGY

This research studies the challenges of identifying the offensive language in Arabic social media using a four-step analysis, shown in Figure 1

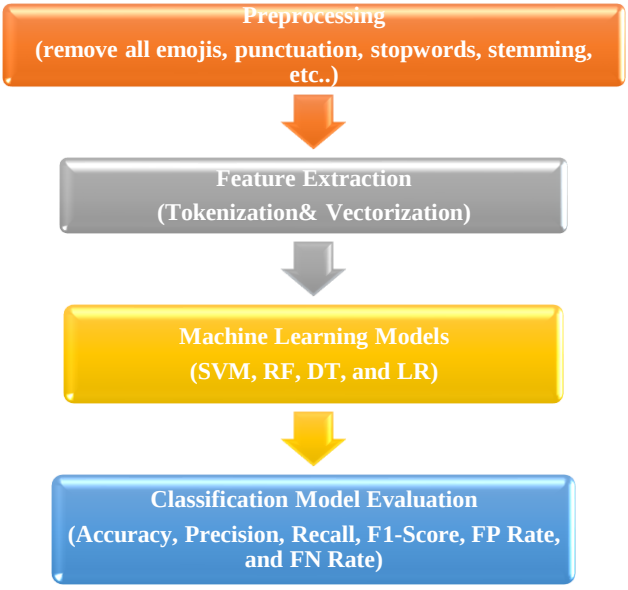

FIGURE 1. Proposed Methodology

The initial step includes a comprehensive cleaning of tweets' raw text data. This means removing all characters that are not Arabic and punctuation, emoticons, emojis, numerical characters, and diacritics.

The process includes applying stemming and removing Arabic stopwords. The preparation of this data provides the foundation for any additional study. The cleaned raw data is then put into a format that machine learning models can use to process. TF-IDF (Term Frequency – Inverse Document Frequency) technique is used in this step

TF-IDF is a method to determine a word's importance based on its frequency inside the document and its overall frequency throughout the data collection. Through this process of feature extraction machine machine-learning models learn to identify the essential linguistic patterns associated with offensive language. Here, the vector to be used as the input of machine-learning models is the TF-IDF score of each word in each document or tweet. We further modify the properties we have already obtained in this step.

The final step is to provide the machine learning classifiers with all previously acquired features to complete this process. In this step, the dataset generated by classifying them (Offensive and non-offensive tweets) is used to train the machine-learning models such as support vector machine (SVM), random forest (RF), decision tree (DT), and logistic regression (LR).

To confirm that the models can accurately identify the offensive language in Arabic social media and that the training process is continually evaluated using performance metrics such as accuracy, precision, recall, f1-score, and misclassification error. This four-step process is established by the study and shows how well the implementation of offensive language detection in Arabic social media is applied.

Dataset Description

The Arabic Cyberbullying Corpus (ArCybC) is available to academics seeking datasets to use in the field of Arabic cyberbullying detection According to (Shannag, F.;2023), this dataset consists of 4505 Arabic language tweets from various domains, categories, Final OFF, and Final CB. The tweets are all publically accessible. By letting users classify tweets based on these distinctions, it enables them to delve deeper into understanding different facets surrounding online harassment within the Arabic language context. This paper also details how there are 1887 offensive tweets within this collection representing approximately 42 %.%. As displayed in Table 1 below, three such examples can be found.

TABLE 1.Sample content of the dataset

Id	text	domain	categories	Final Off	Final CB
1.320000e+18	ما الذي اختلف...	Celebrities	Appearance	yes	Yes
1.310000e+18	تحية من ارض الكنانة ... لحبيب @USERNAME	Celebrities	Appearance	no	No
1.290000e+18	كل العراق يحبك الله... يحفظك. انت اس @USERNAME	Celebrities	Appearance	no	No

Dataset Preprocessing

The tweet text in the dataset is processed by Removing All Emojis, Diacritics and Normalization, URLs, Hashtag, Mention, English Alphabets, Numbers, Punctuations, Arabic Stop Words, Arabic Stemming, Duplicates and Unnecessary Attributes. As shown in Figure 2

	text	label
0	ختلف يوم حتا قوم شعب لبنان باعطاء ثقة اذا بعد ...	1
1	اي هجوم ايرن باي شكل واج برد اقو بالف مرة	0
2	ابو علي حبيب دع قفل سير فقد ملك قلوب عرب جميع ...	1
3	تحي عظيم ارض كنان حبيب مصر حيط علم بان قلوب قش...	0
4	عراق يحب الل حفظ انت اسطور فتخر كعرب اكو هيج ف...	0
...	...	...
4500	رجع قول جسد رحم ابن مفتر	1
4501	طول حلم فهذ ناد طموح ايتعد حدود حلم فقط الا عل...	0
4502	ياشم غمض عين عني فما زال ماض يشف مني	0
4503	رحم امك مدرب مرريض طلع بزر راشفورد نزل دي بيك...	1
4504	اخر مبارا كثير ميس مو عاد حق نيك طيز الي تجن	1

4444 rows × 2 columns

FIGURE 2. Sample contents of the dataset after preprocessing

EXPERIMENTS and RESULTS

The dataset was divided into 80% training and 20% testing datasets. The Python sci-kit learn library is utilized for research experiments. A well-known, easy-to-use tool for predictive data analysis is a ski-kit-learn library. To generate a feature matrix, we utilize TF-IDF together with the TfidfVectorizer method to generate a numerical presentation of the text input. By specifying the n-gram_range option (1,1) a unigram model is applied in this instance. It follows that the TF-IDF approach considers the frequency of each word in a document as well as its overall significance for the entire dataset.

By considering every feature in the feature matrix, numerous experiments were carried out on the Arabic ArCybC corpus dataset. Based on the previously specified parameter choices, the TfidfVectorizer function generates

1909 features in total. The classifiers Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), and Logistic Regression (LR) are used in the dataset studies.

Table 2 shows a comparison of the performance results for the four classifiers

Table 2. The performance of classifiers with all features

Classifier	AVG Accuracy	AVG Precision	AVG Recall	AVG F1-Score	FN_rate	FP_rate
SVM	0.825	0.828	0.736	0.778	0.265	0.11
RF	0.83	0.872	0.694	0.773	0.305	0.073
DT	0.737	0.943	0.396	0.557	0.603	0.017
LR	0.822	0.816	0.741	0.777	0.258	0.119

As shown in Table 2, the "AVG Accuracy" indicates the average accuracy across multiple runs. The table shows that most classifiers exhibit similar average accuracies indicating consistent performance except for the DT, which achieved (0.737). Precision, which represents the true positive ratio, measures the proportion of true positive predictions among all positive predictions. DT achieved the highest precision (0.943), meaning it has a low rate of false positives (mistakenly classifying offensive content). However, this comes at a cost - DT's average recall is the lowest (0.557).

Recall (sensitivity) represents the proportion of true positive predictions among all correctly predicted instances, which reflects the ability to catch true positives (actual offensive content). LR exhibits the highest average recall (0.741), signifying it can effectively identify most offensive language contents. This strength is balanced with a reasonable precision of 0.816.

F1-Score combines precision and recall, providing a more balanced view. Here, SVM takes the lead with the highest average F1-score (0.778), suggesting a good overall performance in capturing both offensive and non-offensive content accurately.

For the FN_rate which reflects the proportion of actual offensive instances misclassified as non-offensive (false negatives), the DT has the highest FN_rate (0.603), indicating it misses many offensive instances, despite its high precision, while the other classifiers perform better in minimizing false negatives. FP_rate represents the proportion of actual non-offensive instances misclassified as offensive (false positives). The LR has the highest FP rate (0.119), still, indicating few false alarms, while the DT achieved the lowest FP rate (0.017), but at the cost of very low recall (0.396).

In summary, understanding these performance measures of the machine-learning models helps select an appropriate classifier based on the consequences of false positives and false negatives in offensive language detection systems. Accordingly, if the detection system needs to capture the most offensive content, SVM might be suitable, which has the lowest false rate. On the other hand, if minimizing false positives (flagging harmless content) is crucial, DT might be preferable despite missing some offensive language.

DISCUSSION

This section compares our experiment results with another study in the literature review and shows that all classifiers achieved high accuracy rates in the TF-IDF model. Regarding Table 2, the accuracy rates ranged from 0.737 to 0.83. The RF models were observed to be slightly better in predicting offensive language with an equal accuracy rate of 0.83. However, the SVM model showed a higher F1-score rate of 0.778, while the DT exhibited a lower F1-score rate of 0.557. While the recall for SVM is 0.736, the recall for DT is lower at 0.396. On the other hand, the precision for DT is 0.943.

Table 3 compares our model and related work on the ArCybC dataset based on the precision measure.

Table 3. Compare our model with other related work based on precision

Precision	SVM	RF	DT	LR
Shannag et al. (2022a)	0.818	0.805	0.781	0.816
Our model	0.828	0.872	0.943	0.816

CONCLUSION

In this research, an investigation is performed to evaluate the impact of increased runs on the performance of various machine learning classifiers, Support Vector Machine, Random Forest, Decision Trees, and Logistic Regression in detecting offensive Arabic-language content on social media.

The results of this research prove that the application of increase runs generally leads to improvements in several key performance metrics, in particular, average precision and recall, across most classifiers. For example, the DT classifier, when utilized on precision, showed an increase in average precision. which demonstrates a more balanced performance. Likewise, the Random Forest classifier, which already had a strong precision, benefited from a remarkable boost in recall and a slight increase in overall accuracy.

FUTURE WORK

In the future, combining Arabic language-specific feature extraction methods and word embedding techniques. Word embedding techniques like AraVec and AraBERT, for example, can greatly enhance the dataset's word representation. Word language interactions and semantic similarity can be represented by these embeddings.

REFERENCES

1. Al-Badri, Y., ElSherief, M., & Diab, M. T. (2022). Towards fair and robust offensive language detection: A cross-cultural perspective. arXiv preprint arXiv:2204.00050.
2. Davidson, T., Warmsley, D., & Macy, M. (2017b). Automated Hate Speech Detection and the Problem of Offensive Language.
3. ElSherief, M., Al-Badri, Y., & Diab, M. T. (2023). Challenges and opportunities for offensive language detection in Arabic social media. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 7544-7564).
4. Husain, F., & Uzuner, O. (2021). A Survey of Offensive Language Detection for the Arabic Language. In *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* (Vols. 20–20, Issue 1, pp. 12–12). https://doi.org/10.1145/3421504.
5. Kauchak, D. (2016). Neural Networks. In *Spring 2016*. https://cs.pomona.edu/~dkauchak/classes/s16/cs30-s16/lectures/lecture12-NN-basics.pdf.
6. Khan, G. F., Swar, B., & Lee, S. K. (2014). Social Media Risks and Benefits. Social Science Computer Review, 32(5), 606–627. https://doi.org/10.1177/0894439314524701 .
7. Shannag, F., Hammo, B. H., & Faris, H. (2022a). The design, construction, and evaluation of annotated Arabic cyberbullying corpus. *Education and Information Technologies*, 1-47.
8. Shannag, Fatima (2023), "ArCyC: A Fully Annotated Arabic Cyberbullying Corpus", Mendeley Data, v1, http://dx.doi.org/10.17632/z2dfgrzx47.1, DOI is reserved

but not active, https://data.mendeley.com/v1/datasets/z2dfgrzx47/draft?a=12a9ff5d-6c5c-4b2e-8990-7d044d7c12e2.

9. Tarawneh, O., Tarawneh, M., Sharrab, Y., & Husni, M. (2023, October). Mushroom classification using machine-learning techniques. In *AIP Conference Proceedings* (Vol. 2979, No. 1). AIP Publishing.
10. Zimba, O., & Gasparyan, A. Y. (2021). Social media platforms: a primer for researchers. *Reumatologia*, *59*(2), 68–72. https://doi.org/10.5114/reum.2021.102707.