2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM)

2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM) | 978-1-6654-1407-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICSECS52883.2021.00051

A Data-oriented Approach for Detecting offensive Language in Arabic Tweets

Eshrag A. Refaee

College of Computer Sciences and Information Technology

Jazan University

Jazan, Saudi Arabia

erefaie@jazanu.edu.sa

***Abstract*— The growing popularity of social media (SM) platforms has made these platforms a crucial part of modern societies. Users from different cultures, backgrounds, demographics get aboard in an increasing manner to express their views, stances, and opinions on a varied range of topics. Since users on SM can easily hide their real identity, a closer look at daily posts on social medial platforms shows that users do not seem to reflect only their stances and views, but also, they get an opportunity for revealing their behaviors, which could be negative towards the others. Although only a small population of SM users can show negative behavior towards other individuals, groups, and society in general, the impact could be catastrophic. This has resulted in the emerge of terms like cyberbullying, online extremism/hatred/threatening, online trolling, online political-polarity discourse. To ensure safe social networking, the domain of automatic detection of offensive/hatred language has lately grown notably. This work focuses on utilizing a publicly available dataset of Arabic tweets labeled for offensive/non-offensive language. Unlike previous work which focuses merely on developing and tuning machine learning models to be as accurate as possible on the benchmark dataset used, we turn to focus on the characteristics of the offensive language used in SM. The purpose is to have an in-depth look into the dataset to disclose what seems to be hidden patterns in offensive language expressed daily online. Our findings reveal the benefit of using larger training dataset that covers a wide range of offensive language patterns to build robust machine learning classifiers with a better ability to generalize well on highly sparse data used in SM.**

Keywords—offensive language, Arabic NLP, machine learning, classification, twitter.

I. Introduction

The growing popularity of SM to become an essential part of the daily lives of millions of people globally has correlated with the importance of maintaining safe social networking. Cyberbullying and online hatred/polarity discourse are emerging terms that essentially refer to the use of a personal electronic device (such as, mobile, PC, etc.) to send hurtful content to/about an entity (mostly an individual) that results in causing harm. According to the United Nations (UN), hate speech can be defined as "any kind of communication in speech, writing or behavior, that attacks or uses discriminatory language concerning a person or a group based on their religion, ethnicity, nationality, race, color, gender, or other identity factors. This often generates intolerance and can be demeaning and divisive" [22].

Research on the automatic detection of offensive/hatred language is motivated by its real-world applications. This domain of research can help to prevent a growing phenomenon of cyberbullying. Cyberbullying has been recognized to have a negative impact on individuals (e.g., an increase in the number of suicide cases) as well as groups and society in general (e.g., causing hate crimes and/or terrorism and atrocity crimes) [22]. To automatically detect cyberbullying, machine learning (ML) approaches have been frequently employed to train classification models by feeding them with a large number of training examples [20]. The resulted ML-based systems can help to serve several real-world applications. For instance, automatic detection of offensive language can help in automatic blocking and reporting of such content, which can significantly help to reduce cyberbullying cases quickly and efficiently. This can result in preventing the dangerous consequences of cyberbullying (e.g., suicide cases, thoughts, and attempts) [17,19]. In addition, automatic and accurate detection of abusive/negative/hatred language can play a major role in the early recognition of extremism discourse [16][15]. Another crucial potential application of offensive/abusive language detection task is in spotting pornographic advertisements which have spread noticeably recently. Cyberbullying can also contribute to an increase in the cases of depression, decreased self-worth, hopelessness, and loneliness [19]. Such an observation stresses the fact that mechanisms currently utilised by SM platforms like Twitter blocking mechanisms are far from being sufficient.

The automatic detection of offensive/hatred content on SM is a recently emerging domain with growing interest in building robust and accurate machine learning models for this task. The problem is mostly addressed as a binary classification task in which a trained ML model needs to decide if a given data instance is offensive or not offensive.

In this work, we utilize a benchmark dataset to explore characteristics of the offensive language, aiming ultimately to further improve the accuracy of automatic offensive language detection. The more accurate ML models become in detecting offensive/hatred language, the more useful the deployed models/systems become for SM platforms. In this work, we explore the following research questions:

RQ1: How would manipulating data parameters affect ML models, e.g., data size, class balance, etc. How would an ML-based model react to a data-centric approach?

RQ2: Would offensive language hold to the same characteristics across cultures and languages? An example of English vs. Arabic.

978-1-6654-1407-4/21/$31.00 ©2021 Crown
DOI 10.1109/ICSECS52883.2021.00051

Unlike previous work on detecting offensive language, as those reported during the international OffensEval-2020 shared task, which has focused primarily on a fixed benchmark dataset while tuning ML model parameters, in this work we have the model fixed and experiment with different data settings. The aim is to help better understanding the dataset. The idea is inspired by [18] who has proposed adopting a data-centric approach as an alternative to the algorithm/model-centric approach, which is currently predominant [4]. By adopting a data-centric approach, we aim to help to gain an insight into the main features defining the syntax and semantic of offensive/hatred language in Arabic content of social media. The results of such investigation can help to build more robust ML-based models for detecting offensive/hatred content, which will ultimately contribute to having a safer online environment.

II. RELATED WORK

Due to its importance and its large number of real-world applications, the Natural Language Processing (NLP) task of utilizing an ML-based approach for automatically detecting offensive language has attracted a considerable volume of research recently. The ability to spot hatred/offensive language in social media was addressed using multiple approaches, namely utilizing lexical-based and/or ML-based methodologies. The authors in [5] have employed a small seed list of offensive Arabic words which was later used to compile a larger list of 3.4k unigrams and bigrams. After that, they employed the generated list for spotting Twitter authors who tend to be active and use abusive language more often. To test their approach, they experimented on a test set of nearly 1k tweets and attained an F-score of 0.60. Although the lexical-based approach has shown to performs reasonably well on the same task for languages like English [22], the high level of data sparsity resulting from the morphologically rich nature of Arabic making this task harder on Arabic using merely lexical-based methods. In [13], the authors investigate the possibility of recognizing Arabic Twitter accounts with abusive content. They built a dataset of 1.3 M Arabic tweets and employed an off-the-shelf tool to detect misspelled words followed by a simple n-gram analysis with an SVM model. They reported a significant correlation between the use of misspelled words and abusive content of Twitter, as people posting abusive content tend to use misspelled words to avoid being spotted by the platform administration.

The growing interest in tackling the issue of cyberbullying has led to launching large-scale efforts among researchers. For instance, in [4] and [11] the authors have led international shared tasks, namely SemEval and OSACT, for detecting offensive/hatred language in Twitter across multiple languages, including English and Arabic. A gold standard data-set of 10k Arabic tweets has been built for this purpose. The results of both competitions have shown a considerable level of difficulty in detecting offensive language in Arabic tweets. Many reasons have been pointed out for why the task of automatic detection of offensive language seems to be highly challenging for ML models. Amongst these reasons, few stood out to be unique for Arabic, such as being a morphologically rich language (i. e. large number of possible word forms variations) and a low-resourced language (i.e. as compared to English, for instance). Another reason for considering this task highly challenging is the extensive use of sarcasm, especially when conveying a negative stance [7][10]. The top-performing system in the shared task was reported by [12]. The authors in [12] use a pre-processing method for Arabic tweets that mask emoticons to their Arabic meaning. It is interesting to mention here that emoticons have previously shown to be effective in detecting emotional stances conveyed in Arabic tweets [2]. A full summary of automatic offensive language detection experiments on the benchmark data set we use in this work can be found in [11].

Previous work shows a considerable interest in employing ML-based methods for the automatic detection of offensive and hateful language in social media [4] [11]. In this work, we utilize the dataset built in the aforementioned shared task to explore characteristics of the offensive language, aiming to further improve the accuracy of this NLP task.

III. METHODOLOGY

In this work, we employ an ML-based approach with Support Vector Machines (SVM) and utilize a publicly available benchmark dataset of 10k Arabic tweets manually annotated for offensive vs. non-offensive language. The experimental setup has focused on a data-centric approach, i.e., a fixed ML model with variable settings for the dataset. The experiments followed by a manual inspection aiming for gaining a broad understanding of the nature and characteristics of offensive/hatred language used in the Arabic content of SM.

A. Dataset

We use an existing and publicly available dataset of 10k Arabic tweets [4]. The dataset was manually annotated for offensive/hatred language by native speakers of Arabic with a reported Kappa coefficient inter-annotator agreement at 0.92, showing an acceptable level of the quality of the data annotation. The dataset is divided into 8k training and 2k testing, using the same data split in [11]. It is important to note that the dataset is highly unbalanced with non-offensive tweets representing 79.94%. Although it is naturally expected to have more non-offensive content in a random sample of tweets, with a highly unbalanced dataset the performance of ML models can likely skew towards the majority class, i.e., non-offensive. As such, we experiment with two settings, 1) original class distribution of the dataset 2) corrected data distribution. For the latter, we utilize a popular technique, namely Synthetic Minority Oversampling Technique (SMOTE) [21]. This method involves oversampling the minority class by creating synthetic minority class examples.

To compare the style of offensive language spotted in Arabic tweets with another language, we consider looking into an English dataset. Specifically, we use the English dataset from the same shared task of detecting offensive language in SemEval competition [11,20]. We do not run any experiments on the English dataset, instead, we use it only to explore the main differences in a language style that distinguish Arabic users from other users, especially when it comes to cyberbullying.

B. Data pre-processing

Arabic twitter datasets are extremely noisy, e.g., with spelling mistakes, elongation, Arabizi, usernames, hashtags, etc. As such, we utilize similar pre-processing settings as those reported in [4]. These settings include:

- **Character normalization**: For instance, normalizing all different forms of alef, hamza, and ta into a single unified form.
- **Word stemming**: Arabic is a morphologically rich language. Meaning that a single word can have hundreds of different variant word forms. The application of stemming tools helps reduce the feature vector representation of the dataset. ML-based models we use in this work rely on the feature vector to learn. The reduction of the feature vector size has two main benefits 1) reducing the computational power required and 2) reducing the number of noisy features and help the trained model in avoiding noise created by irrelevant features. W use a light Arabic stemmer which has shown to outperform root stemmer on a similar text classification work [10].
- **Removing stop words**.
- **Masking usernames, URLs and hashtags with placeholders**.

C. Experimental setup

We use a LibLinear implementation of SVM, as it has shown to outperform other ML-based algorithms on text classification tasks [10]. For features, we employ word n-grams, specifically unigrams and bigrams, as they have shown to outperform other feature sets in our preliminary experiments.

Similar to [4,11], we report the results using the macro F1-score (1). In addition, we report the accuracy, per-class precision (p), and recall (R) scores in order to give a broader insight into the performance of the trained models.

$$(2 * Precision * Recall)/ Precision + Recall \quad (1)$$

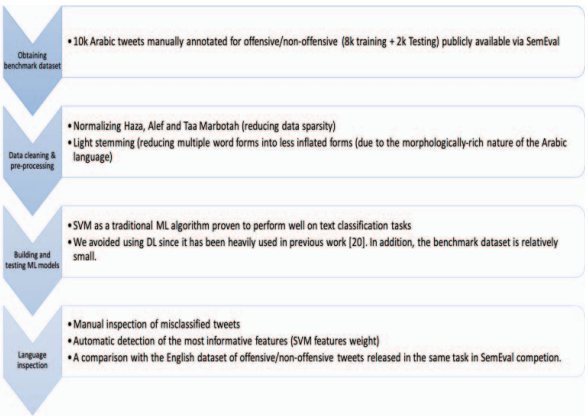

FIGURE 1. METHODOLOGY FRAMEWORK

IV. RESULTS AND DISCUSSION

Table 1 summarises the results of our experiments. Note that we use OFF as a short for offensive and NOT as a short for not offensive. It is interesting to see that the results of the SVM model which was trained on a feature vector of nearly 10k features of word n-grams have attained an F-score of 0.898, which is a nearly identical performance to that obtained by a Neural Network (NN) model at 0.901[11]. It is worth mentioning that NN used was trained using millions of tokens. This probably suggests the competitive performance and robustness of traditional ML-based models like SVM as compared to NN, especially in classification tasks. A further improvement of performance could be obtained by employing an ensemble of different models [11].

A. Highly Unbalanced Dataset

As for corrected class balancing experiments with SMOTE, the results show a significant improvement in the performance of up to 7% in F-score and attaining an accuracy score of 94.2%. However, we believe this improvement might be a direct result of model overfitting. This means that since SMOTE approach synthesis new data instances out of the existing ones to correct a given highly skewed dataset, in this case, the generated data instances have a considerable degree of similarity to the instances of the original dataset. As such, there is a high chance that the model memorizes the labels rather than learning to predict them on new and previously unseen instances [18]. It is very important to avoid this case from happening while learning a new ML model. This is due to the possibility of deploying an overfitted model with high performance on a specific dataset, whilst the performance is likely to drop significantly as soon as the model is exposed to a new dataset. In such a case, we still believe that increasing the size of the training set is crucial for enhancing the performance of an ML model on detecting offensive language task, yet we recommend more practical approaches for obtaining a larger training set while maintaining reasonable data quality. Data quality can involve many aspects like consistency, diversity, etc. One possible approach is the use of a semi-supervised approach that can be fed with the existing gold-standard manually annotated dataset and allow the partially trained model to contribute to expanding the dataset. This approach has shown to be successful with English [20].

B. Size of the Dataset vs. Learning Curve

To investigate the impact of the size of the training dataset, we experiment with multiple random samples of the dataset. The dataset sizes were set to 1k, 5k, 7k, and the full 8k of the obtained dataset. The results in table 1 indicate a clear steady learning curve of the model, demonstrating the effectiveness of increasing the size of the training set. However, it is important to be careful to have the training instances collected from different samples of data [18]. This can increase the possibility of increasing data coverage, and hence, exposing the trained models to new examples and avoid overfitting.

TABLE 1. RESULTS OF DETECTING OFFENSIVE/HATRED LANGUAGE EXPERIMENTS

	Task: offensive language detection			
	F score	***Precision***	***Recall***	***Accuracy***
1k tweets	0.798	P-OFF 0.434 P-NOT 0.880	R-OFF 0.458 R-NOT 0.870	79.60%
5k tweets	0.833	P-OFF 0.566 P-NOT 0.900	R-OFF 0.593 R-NOT 0.890	83.17%

	Task: offensive language detection			
	F score	***Precision***	***Recall***	***Accuracy***
7k tweets	0.859	P-OFF 0.899 P-NOT 0.687	R-OFF 0.577 R-NOT 0.935	86.36%
Full original dataset	0.898	P-OFF 0.806 P-NOT 0.915	R-OFF 0.646 R-NOT 0.961	89.80%

The list of keywords in Table 2 shows several features that were selected by the trained SVM model as informative features, meaning that the model found them useful for deciding whether a given data instance is offensive or not.

TABLE 2. SAMPLE OF SVM FEATURES

Keyword/symbol	***English translation***	***SVM weight/coefficient***
❤️	Heart icon	0.39
يلعن	Damn	0.32
يهود	Jews	0.33
يا حمار	You donkey	0.24
يامعرص	Hypocrite	0.80
مرتزق	Mercenary	0.58
حقير	Dispicable	0.36

A closer look at the informative features that were used by the model (Table 1) indicates what type of language is being used by Arabic users on social media while expressing offensive/negative stances. For instance, the heart icon is declared as a strong clue, which is probably because people when using this type of icon are likely to express non-offensive language. Note that the model always seeks clues that can help improve its performance. As such, models tend to try improving not only precision but also the recall rate, i.e., the model's ability to fetch the correct instance for a specific class. This is particularly important in the offensive language detection task as you do not want the deployed model to keep missing offensive instances just because they did not have sufficient clues. As such, the features coverage and the size of the training dataset play an important role (as discussed in section B). In addition, we noticed that the trained models rely heavily on words describing race, ethnicity, or sexual groups. For instance, the model use words like *Jews, She'ah* (a Muslim sect), *Farsey* (Persian), *pervert*. In general, we can observe a clear vulgar language while examining a random set of offensive tweets, featuring keywords like *shit* and private parts, which was also observed by [14].

When it comes to English, it is interesting to see that amongst the most informative keywords are those indicating political polarities, such as *liberals* and *conservatives*. It seems that English users tend more to use offensive language cunningly or indirectly, while Arabic users tend more towards using straightforward, but vulgar language when conveying offensive content on SM.

V. CONCLUSION

The research on detecting hate speech in the new media is motivated by the need to find the relationship between the misuse of SM platforms to spread hate speech and the factors that drive individuals to commit violence [22]. Studying and analyzing hate speech data can play a major factor in gaining a better understanding of hate/abusive/offensive speech trends, actors, and drives, and as a result, have a better ability to detect and block them efficiently.

This work explores a publicly available dataset of 10k Arabic tweets manually annotated for offensive vs. non-offensive. The dataset was manually annotated by native speakers of Arabic. Unlike previous work which has focused on enhancing and improving the model performance by tuning its parameters, in this work we adopted a data-centric approach. The aim is to study the characteristics of hate speech discourse. We studied the learning curve on the given benchmark dataset by gradually increasing the size of the training dataset. Our experiments have shown that traditional ML-based algorithms perform as well as neural networks (NN)-based algorithms, although the latter was trained on a much larger dataset. We strongly believe that using an ensemble of models compromising both traditional and NN approaches will significantly improve the deployed model's ability to detect offensive language. Improving the quality of the model's performance is crucial in order to be able to automatically detect, and hence, block the SM content with offensive/hatred language. The importance of this task is highlighted by its possible contribution to reducing cyberbullying and allow a safer online environment. Research on other languages like English has already attained remarkable progress due to many factors, out of which the size of the training dataset stands out as it exceeds 9M English tweets. Our results show a steady learning curve while increasing the size of the training dataset, which is only 10k Arabic tweets. This opens the possibilities for future research that involves building and experimenting with larger datasets. The larger dataset used; the better language coverage attained to address the complexity of the language used in offensive/hatred discourse. As such, our findings reveal a strong need for building a larger training dataset that covers wide range of offensive language patterns. Future work can also explore more in-depth characteristics of offensive data. For instance, research can investigate which type of hate speech (e.g., gender-based, religion-based, etc.) is more common or conduct a cross-language pattern comparison.

VI. REFERENCES

[1] A. Alshehri, E.M.B. Nagoudi and M. Abdul-Mageed, "Understanding and Detecting Dangerous Speech in Social Media," May 04,. 2020.

[2] E. Refaee and V. Rieser, "Can We Read Emotions from a Smiley Face? Emoticon-based Distant Supervision for Subjectivity and Sentiment Analysis of Arabic Twitter Feeds," Proceedings of the 5th International Workshop on Emotion, Social Signals, Sentiment and Linked Open Data, lREC, .

[3] S. Alsafari, S. Sadaoui and M. Mouhoub, "Hate and offensive speech detection on Arabic social media," Online Social Networks and Media, vol. 19, pp. 100096, Sep. 2020.

[4] H. Mubarak, K. Darwish, W. Magdy and H. Al-Khalifa, "Overview of OSACT4 Arabic Offensive Language Detection Shared Task," With a Shared Task on Offensive Language Detection. Language Resources and Evaluation Conference (LREC 2020), pp. 11, -05. 2020.

[5] H. Mubarak, K. Darwish and W. Magdy, "Abusive Language Detection on Arabic Social Media," Proceedings of the First Workshop on Abusive Language Online, pp. 52, Aug 04,. 2017.

[6] F. Husain and O. Uzuner, "Transfer Learning Approach for Arabic Offensive Language Detection System -- BERT-Based Model," Feb 08,. 2021.

[7] F. Husain and O. Uzuner, "Leveraging Offensive Language for Sarcasm and Sentiment Detection in Arabic," Proceedings of the Sixth Arabic Natural Language Processing Workshop, pp. 364, -04-19. 2021.

[8] A. Alakrot, L. Murray and N.S. Nikolov, "Towards Accurate Detection of Offensive Language in Online Communication in Arabic," Procedia Computer Science, vol. 142, pp. 315-320, 2018.

[9] S.A. Chowdhury, H. Mubarak, A. Abdelali, S. Jung, B.J. Jansen and J. Salminen, "c European Language Resources Association (ELRA), licensed under CC," Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pp. 11, -05. 2020.

[10] E. Refaee, "Sentiment Analysis for Micro-blogging Platforms in Arabic ," in Social Computing and Social Media. Applications and Analytics, 2017.

[11] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis and Ç Çöltekin, "SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020)," Jun 12,. 2020.

[12] H. Alami, S. Ouatik, E. Alaoui, A. Benlahbib and N. En-Nahnahi, "LISAC FSDM-USMBA Team at SemEval-2020 Task 12: Overcoming AraBERT's pretrain-finetune discrepancy for Arabic offensive language identification," Proceedings of the 14th International Workshop on Semantic Evaluation, pp. 2080, -12-12. 2020.

[13] E. A. Abozinadah and J. H. Jones James, "Improved Micro-Blog Classification for Detecting Abusive Arabic Twitter Accounts," International Journal of Data Mining & Knowledge Management Process, vol. 6, pp. 17-28, Nov 30,. 2016.

[14] H. Mubarak, A. Rashed, K. Darwish, Y. Samih and A. Abdelali, "Arabic Offensive Language on Twitter: Analysis and Experiments," Apr 05,. 2020.

[15] A. ALDayel and W. Magdy, "Stance detection on social media: State of the art and trends," Information Processing & Management, vol. 58, pp. 102597, Jul. 2021.

[16] S. Hamidi. “Mining Ideological Discourse on Twitter: the Case of Extremism in Arabic”. the Discourse and Communication journal. (in press)

[17] S. Almutiry, M.A. Fattah, A. Saudi and Almunawarah, "Arabic CyberBullying Detection Using Arabic Sentiment Analysis," . The Egyptian journal of language enguneering. vol. 8, pp. 39-45. 2021.

[18] A. NG, Machine Learning Yearing, deeplearning.ai. 2018.

[19] S. McLeod, “Cyberbullying: A special feature presentation”. The Centre for Suicide Prevention workshop Straight Talk: Youth Suicide Prevention.

[20] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal’ N. Farra, R. Kumar. Predicting the Type and Target of Offensive Posts in Social Media”. Proceedings of NAACL-HLT 2019, pages 1415–1420 Minneapolis, Minnesota, June 2 - June 7, 2019. Association for Computational Linguistics.

[21] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: synthetic minority over-sampling technique. Journal of arti_cial intelligenceresearch, pages 321-357, 2002.

[22] S. Sood, J. Antin, and E. Churchill. 2012b. Using crowdsourcing to improve profanity detection. In In AAAI Spring Symposium: Wisdom of the Crowd, volume SS-12-06 of AAAI Technical Report. AAAI.

[23] United Nations. The United Nations Strategy and Plan of Action on Hate Speeach. Available at : https://www.un.org/en/genocideprevention/documents/advising-and-mobilizing/Action_plan_on_hate_speech_EN.pdf. Accessed on 05.06.2021.

[24] C. Shammur Absar H. Mubarak, H. A. Abdelali, S. Jung,J.Bernard J. and J. Salmine. A Multi-Platform {A}rabic News Comment Dataset for Offensive Language Detection. Proceedings of the 12th Language Resources and Evaluation Conference. 6203--6212,may 2020, Marseille, France. European Language Resources Association.