Offensive Content Identification in Arabizi Tweets

Majd Al Aawar, Ramzi Haddad, Yazan Hajj Diab, Ali Ismail

***Abstract*—Due to having versatile functionality, social media platforms have become more popular. With the increasing traffic on social media platforms comes an increase in negative and destructive behavior by users across the global multilingual landscape. Removing these actions or posts and punishing the perpetrators has been one of the main problems companies such Facebook, Twitter, and Instagram are trying to tackle. Due to the global nature of these platforms, detecting hostile languages in every language has not been accomplished yet. In this paper, we focus on detecting offensive posts on Twitter written in Arabizi. Arabizi is a new and popular Arabic writing system, mainly used as an easier form of communication on the internet, that is a combination of Latin alphabets and Roman digits. We aim to build a deep learning model based on Arabizi that can detect offensive Tweets in a more efficient manner. Since there are no available datasets on Arabizi, we will be using a transformed version of the Arabic L-HSAB dataset. Finally, we aim to compare the results we obtain with the results of the state-of-the-art paper, which uses an Arabic-based model, on the same dataset.**

# Introduction

Social media has seen its influence increase tremendously in the 21st century. Although social media is mainly used to connect people together, platforms such as Twitter, Facebook, and Instagram have been involved in elections [1], personal data scandals [2] and have become a prominent source of news [3] and a powerful marketing and advertising tool [4].

Unfortunately, the ability to communicate with people instantaneously is not always beneficial. According to a poll by UNICEF and the UN Special Representative of the Secretary-General on Violence against Children, 33% of children in 30 countries have been the target of offensive posts on social media platforms. This has led 20% of these children to deliberately absent from school as they are being cyberbullied [5]. In addition, social media allows people endorsing hate speech to reach an audience and recruit each other to form racist, sexist, or xenophobic groups that aim to hurt other groups of people [6].

Due to the harmful effects of these posts, social media platforms are implementing measures to block them [7]. However, it is challenging for these platforms to moderate posts in various languages. Facebook finds it difficult to keep up with the flood of new languages being used on its platform [8]. The difficulty lies in the training process typically being language specific, meaning that for each language to be processed, there is a need to collect a separate, large set of data which is comparable to solving the problem from scratch [9]. Even though Facebook is available in 111 languages, its community standards, which prevent users from posting offensive content, are only available in 41 languages. This impedes the job of Facebook in making it a friendly social media platform. In addition, some countries such as Australia, Singapore, and the United Kingdom are threatening harsh new regulations, fines, and even the incarceration of executives if Facebook fails to remove such offensive content.

Arabic is one of the most commonly used languages on social media especially in the MENA region. It is also the fastest growing language on Twitter [10]. Arabic is often regarded by the natural language processing (NLP) society as a low-resource language offering its own set of challenges due to its complex morphology and the large variety of arabic dialects used which introduce ambiguity in word meanings since the same word can have different meanings in different dialects (for example “solb” means “solid” in Egyptian but in Sudanese it has a profane meaning) [11][12]. Furthermore, Arabizi has emerged as a way to type arabic with latin characters and numbers. It is widely adopted by the young generation because it’s “easy” to type it on keyboards [13].

Offensive post detection in social media has been the object of a plethora of works due to its problematic nature and the challenges it presents. Several have undertaken the task of detecting whether a social media post is offensive or not [14][15]. Many have focused on distinguishing between specific types of offense such as hate/general offense [16][17][18], racism/sexism [19] and offensive posts targeting specific people or groups [20]. Because of the lack of uniformity in depicting offensive social media content (hate speech, cyberbullying, abusive language, toxicity, racism, sexism…), the Offensive Language Identification Dataset (OLID) [21] was compiled from Twitter. Each tweet is annotated in a hierarchical manner: **offensive** or **not**, **targeted** insult or **untargeted**, and targeting an **individual** or a **group**. The release of this dataset has spurred a large number of works to tackle this classification using different approaches [22 - 29]. There is also a developing interest in accomplishing offensive content identification in different languages such as German, Hindi, Marathi, Italian and Indonesian [11][31] [32][33].

Since social media has a prominent role in the Arab World, Arabic has attracted the attention of researchers in the field of NLP for offensive content detection [11][12][34][35]. Yet the ability to do this in Arabizi text has not been explored at the time of writing. In fact, many works consider Arabizi text as noise and therefore remove it [36]. A few papers have targeted the Arabizi language for other tasks such as sentiment analysis and yet many of them transliterate it into arabic characters in the process [37][38]. Therefore, a clear gap in the processing of Arabic text written in Arabizi is identified and needs to be addressed as this writing system is becoming a norm among arab-speaking users.

In this work, we aim to detect offensive content in Arabizi tweets. We will then compare the classification results on Arabizi to classification on regular Arabic text. This will inform on which writing system is more easily learned by machine learning algorithms and whether it is a better practice to work with Arabizi writing or transliterate to standard Arabic as others have previously done. Fig. 1 shows the inputs and outputs of the intended model.

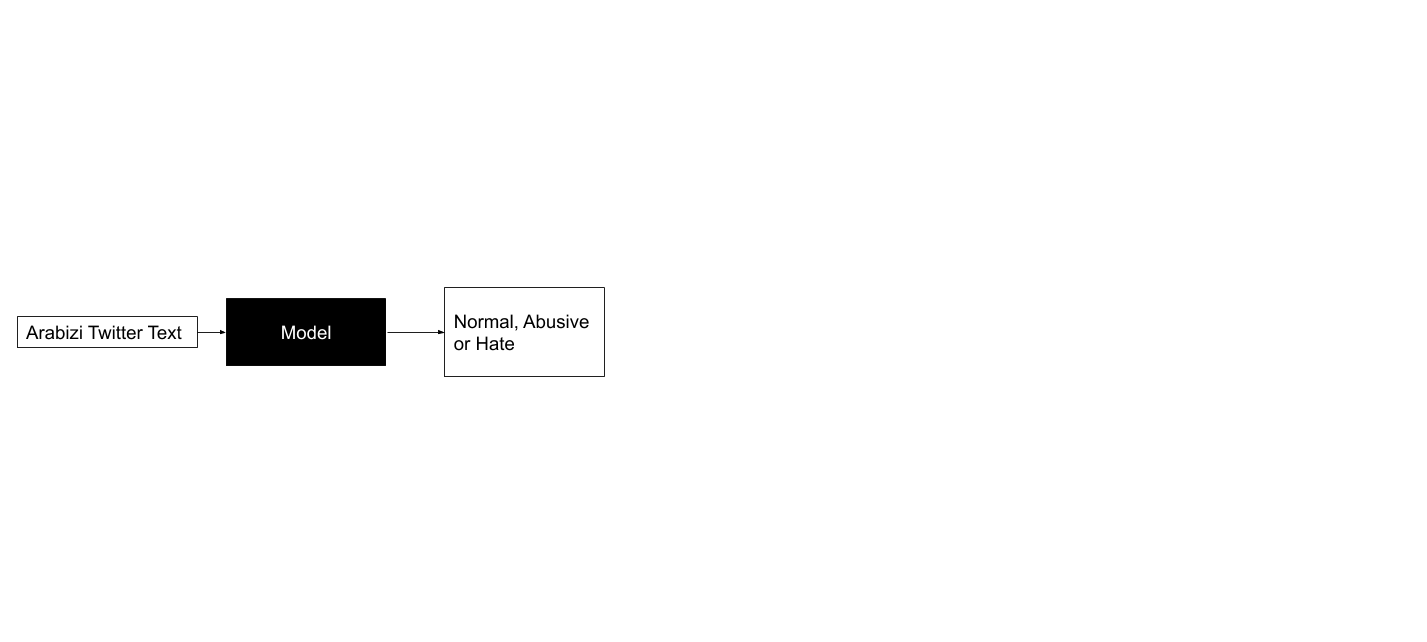


Fig. 1. Expected model inputs and outputs

The L-HSAB dataset[[1]](#footnote-0) will be used in this work. It consists of 5846 Arabic tweets annotated as **normal**, **abusive** or **hate** (targeted offense). It is the most similar description of offensive content to that of the OLID in the Arabic language. This dataset will be transliterated into Arabizi (since at the time of writing there is no native Arabizi dataset) and used to train and test the Arabizi model. It will also be used in its original form to compare classifications in Arabic to Arabizi.

# Related Works

There have been several works done on offensive text identification in various languages. This section will cover a comprehensive summary of previous approaches and more detailed descriptions on papers that were found to be seminal and state of the art.

## Offensive Language Identification

Most of the works that have tackled the challenge of offensive text identification did so in English.

Noteworthy works done on English offensive language identification include Davidson et al. (2017) work on attempting to separate hate speech from other offensive language [16]. In this work, a crowdsourced dataset of tweets categorized as hate speech, offensive language, and neither was used. Using basic models like Logistic Regression and Linear SVM, it was found that:

1. Lexical methods are effective ways to identify potentially offensive terms but are inaccurate at identifying hate speech.
2. Certain terms are particularly useful for distinguishing between hate speech and offensive language.
3. People tended to misclassify some of the data.

Although they intended to address the problem of distinguishing between offensive language and hate speech content, their results yielded a 40% misclassification rate of the hate speech content. This limitation can be expected when working with labels which have a somewhat overlapping definition.

The various works submitted by competitors for SemEval’s OffensEval 2019 task are shown by Zampieri et al. (2019) [22]. Results showed that among the top-10 teams, seven used BERT while the top non-BERT model (ranked sixth) used an ensemble of CNN and BLSTM+BGRU, together with Twitter Word2vec embeddings and token/hashtag normalization.

For a comprehensive survey of other works done in this field the reader is referred to the survey by Al-Hassan and Al-Dossari [11].

## Offensive Language Identification in Arabic

Arabic is considered to be a challenging language because it has many linguistic properties that set it apart from Indo-European languages. For instance, Arabic is written and read from right to left and has no capitalization. Also, short vowels are expressed using special punctuation marks called Diacritics which are often omitted for ease of typing. Three variations of the Arabic language are recognized [12]: Traditional arabic: this is the form used in ancient Arabic literature and in religious scripts. Modern Standard Arabic (MSA): this is the formal language officially taught in schools and used in books and formal documents. It is known to all arabic speakers. And Arabic Dialects: these are commonly practiced between people and vary between countries and even groups of people within a single country. A survey of the literature reveals several publications that have worked on the identification of offensive posts in arabic social media.

Haidar et al. (2017) [12] were the first to tackle this problem in Arabic. First, they provide a detailed background on cyberbullying detection in other languages on other NLP work done with Arabic such as sentiment analysis and named entity recognition. To develop their model, they scraped posts from Twitter and Facebook while restricting their search to the Middle East and Gulf area. They preprocessed and manually labelled the samples as “bullying” or “not bullying”. They experimented with the Naive Bayes (NB) classifier and with a support vector machine (SVM). For the SVM, they adapted TweetToSentiStrengthFeatureVector for Arabic which assigns sentiment weights to tweets. The SVM model outperformed NB but their results in general suffered from class imbalance.

Haidar et al. (2018) [39] continue their previous work by updating the dataset they collected by removing URLs, emoticons and all non-arabic characters. They use a Feed Forward Neural network to do the classification. Their experiments centered around tuning hyperparameters such as number of epochs and number of hidden layers. They show that their final results are an improvement over their previous work.

In a more recent publication, Haidar et al. (2019) [40] introduce ensemble methods to further improve the performance of their system. They use the same dataset as before and experiment with different models (KNN, random forest, SVM, logistic regression) and different ensemble strategies (stacking, bagging and boosting). The optimal results were obtained with the stacking model and they are an improvement over their previous results.

Mohaouchane et al. [41] have worked on binary offensive classification of Youtube comments on popular Arab celebrities. They preprocess the data by removing non-Arabic characters, diatritics and punctuation. They oversample the offensive class (by duplication) to balance classes. The AraVec model is used and fine tuned to generate word embeddings. Four architectures are trained and compared: CNN, BiLSTM, BiLSTM with attention unit and CNN-LSTM. The CNN model achieves the highest overall performance. However, combining CNN and LSTM and including an attention unit with the BiLSTM increase the recall of the model but at the cost of precision. The authors argue that, in this context, it is better to have a higher recall since it means that offensive samples are more likely to be classified accurately.

To classify Youtube and Twitter comments as cyberbullying or not, Mouheb et al. [42] rely on a lexicon based approach as opposed to the more prevalent machine learning approaches. They manually compile a list of bullying words from a sample of comments and each word is assigned a weight ranging from slight to strong bullying. Comments are classified as bullying if they contain one word from the lexicon. Each comment is also assigned a degree of strength (mild, medium or strong) based on a function that takes into account the weight, number and repetition of offensive words. This approach is not context aware as it cannot detect offense unless explicitly expressed with certain vocabulary. Also, no validation criteria is proposed.

## The Arabizi Language

Arabizi (or Arabish) is a non-standardized writing system for the Arabic language using Latin characters. Digits are also used to express letters that are not present in Latin based languages (for example “5” is used to represent the voiceless uvular fricative “Khāʾ”). This way of writing is popular in informal situations and among the younger generation. The literature search performed during this work as well as the survey provided by Guellil et al. [43] show that works on Arabizi are scarce and that it is an emerging field.

Duwairi et al. [37] were the first to perform sentiment analysis on tweets written in Arabizi and classify them as positive, negative or neutral. The dataset for this paper was extracted using Twitter API and labelled using a crowdsourcing tool. As a preprocessing step, they tokenized every tweet into words, transformed every emoticon into its correlated word, removed stopwords, weighted every token using the Binary Model, and converted the tweet from Arabizi to Arabic using their built-in converter, which was one of the main contributions of this paper. Finally two classifiers were used, which were NB and SVM. The predictions, Recall values, and Precision values were better using SVM rather than NB. Also, removing the Neutral class from the dataset improved the performance and the Precision values in predicting the Positive and Negative classes. In this paper, their tool, which translates Arabizi to Arabic, wasn’t very efficient because usually filtering would increase the accuracy significantly while here the accuracy increased slightly when filtering was applied.

Duwari et al. [38] used a tool called Rapidminer which is a platform that helps in data preprocessing and contains machine learning tools. They collected data using a PHP script and crowdsourcing. In Rapidminer, they built their own dictionary for stopwords and added it for topical classification. Also, they added their own packages to convert emoticons to their corresponding words, removed repeated letters, detect negation, map dialect words to MSA, remove links and mentions, and convert the Arabizi text to Arabic. The only used classifiers were the ones built in Rapidminer, which were KNN, NB and SVM. Since Rapidminer has memory issues the dataset used was small (1000 tweets) which makes this paper slightly reliable, but needs further investigation with more data.

Darwish [44] presented a method to detect Arabizi and also to translate Arabizi to Arabic. Since in our paper we will transform one dataset from Arabic to Arabizi, which is one of the contributions of our paper, we can use the method presented to detect if our new dataset is transformed properly to Arabizi.

## Seminal and State-of-the-Art Work

In this section, we introduce works that are considered seminal and state-of-the-art to this work. The seminal paper is considered to be a previous related work that introduced a significant change in the field of offensive content detection. The state-of-the-art paper is the work that is most similar to this work and to which we can compare our results.

Although SemEval releases tasks frequently, and in many languages, to address the issue of offensive text identification, they recently turned their focus to English and used the Offensive Language Identification Dataset (OLID). Zampieri et al. compiled this three-level hierarchical annotated dataset from a large dataset of English tweets so that they could satisfy the general need for a typology to characterize the different types of offenses [21]. This is considered the new seminal work in this taks since their scheme became thereafter the gold standard for annotation of offensive language and it was carried to different languages and used in competitions. The data was collected from Twitter by using its API and searching for keywords related to offensive messages. A key challenge which was not resolved was obtaining a well-balanced dataset [21]. They also performed experiments with different machine learning models for each level so that they could set important baselines for future works to compare to. Models included a linear SVM trained on word unigrams, BiLSTM, and CNN (an input embedding layer (both using pre-trained FastText as well as updatable embeddings). Results were compared using macro-averaged F1-score to account for the data imbalance. It was found that for level A,B, and C, the CNN had the highest macro-F1 score of 0.80, 0.69, and 0.47 respectively.

Mulki et al. [34] focus on the identification of hate speech specifically in the levantine dialect (Lebanese and Syrian). We consider this paper the most relevant state-of-the-art work since it is one of the more recent works on the problem of offensive text in Arabic and because we are using their dataset, which means that our results will naturally be compared to theirs. They collected the **L**evantine **H**ate **S**peech and **AB**usive dataset (L-HSAB) from the Twitter timelines of politicians, activists and news sources and by querying certain keywords. They manually filtered symbols, digits, non-arabic characters and non-levantine tweets. They labelled the tweets according to the following scheme:

* **Normal**: no offensive content of any type.
* **Abusive**: tweets with offensive, aggressive or profane content.
* **Hate**: tweets that are **abusive** and also direct that offense towards a person or a group of people based on identity (race, religion, nationality…).

The annotators were made aware of the context and background of these tweets (names of ethnic groups, political parties…). Samples that were labeled differently by each of the three annotators were discarded. With the dataset ready, they did some data exploration to discover the most frequent words occurring under each class and which words are significant to each class (by comparing them to their occurrence in the Normal class). They also evaluated the quality of their labelling using statistical annotator agreement measures (PRAM, Cohen’s kappa and Krippendorff’s alpha). It is shown that more disagreement is encountered in the Hate class since it relies on the annotator’s background and prior knowledge. They performed two classification experiments: a binary classification (Abusive and Hate merged into one class) and a multiclass classification. They removed stopwords and experimented with several n-gram schemes and used term frequency (TF) weighing. The classifiers used are SVM and NB. The latter model performed better than the former in all experiments.

## Conclusion and Comparison Against Previous Work

In summary, the literature is rich in works that have attempted to classify offensive text in social media. For arabic, many works have tackled this task targeting different arabic dialects using different annotation layouts and classification methodologies. However, none have done so for text in Arabizi. In fact, research in Arabizi is emerging and many works rely on converting the text to Arabic rather than processing the text in its Arabizi form. Table I shows a comparison of the most relevant works to the work being done in this paper.

Table I. Comparison with related work

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Ref** | **Problem** | **Target classes** | **Language** | **Models** |
| 21 | Offensive content detection | OFF/NOT,  TIN/UNT,  IND/GRP/OTH | English | SVM,  BiLSTM,  CNN |
| 34 | Offensive content detection | Normal, abuse, hate | Levantine Arabic | SVM, NB |
| 37 | Sentiment analysis | Positive/negative/neutral | Arabizi transliterated to Arabic | SVM, NB |
| **This work** | **Offensive content detection** | **Normal, abuse, hate** | **Arabizi** | **BERT, RNN-CNN** |

# Proposed Solution

We propose a pipeline whose goal is to ultimately classify tweets written in Arabizi as **Normal**, **Abuse** or **Hate**.

As a first step the L-HSAB dataset will be preprocessed and cleaned using the most appropriate Arabic text preprocessing methodologies. Mohaouchane et al. [41] provide good details on the technical aspects of these steps. These namely include removing special characters, symbols, diacritics and punctuation. Though sometimes contradicted [11], stemming of arabic words has been found by several works to improve performance [12] and will also be considered.

Next, the dataset will be transliterated into Arabizi. This can be done by writing a set of rule-based transformations to map arabic characters into corresponding latin ones. A plausible approach would be to do the inverse of what was done in previous works [37][44].

The L-HSAB dataset has a significant class imbalance. Therefore, strategies will be considered for treating this imbalance such as class weighing [25] and oversampling (random [41] or SMOTE [27]).

Afterwards, word embeddings will be generated as input features to the classification models. Transfer learning of embedding models is a common approach and has been successful in many situations. In multilingual contexts, fastText has been shown to generate good word embeddings for different languages [33]. AraVec is also an embedding model dedicated to Arabic [41]. However, since the input text in this work will be in Arabizi, and since there is no model that has been pre-trained to generate embeddings for Arabizi, it might be preferable to train an embedding layer from scratch. Indeed, in some cases a custom embedding layer outperformed a pre-trained layer even in a supported language [24] since these pre-trained models might be trained on a different type of document (such as Wikipedia articles).

For classification, the multilingual BERT model will be used since it has yielded state-of-the-art performance in many NLP tasks and specifically in offense detection [22]. Also, combinations of RNN networks (LSTM or GRU) with CNN networks have been reported to be successful in several occasions for offense detection [27][41] and will be tested for this task as well. The outputs of any of these representation learners will be fed to a fully connected layer followed by a dense classification layer. To compare with classification of regular Arabic text, the results of Mulki et al. [34] will be first considered. Then the model used for Arabizi will be retrained for Arabic and used to generate predictions for the same test set used in Arabizi but in its Arabic form. Fig. 2 shows the high level architecture of the proposed system.

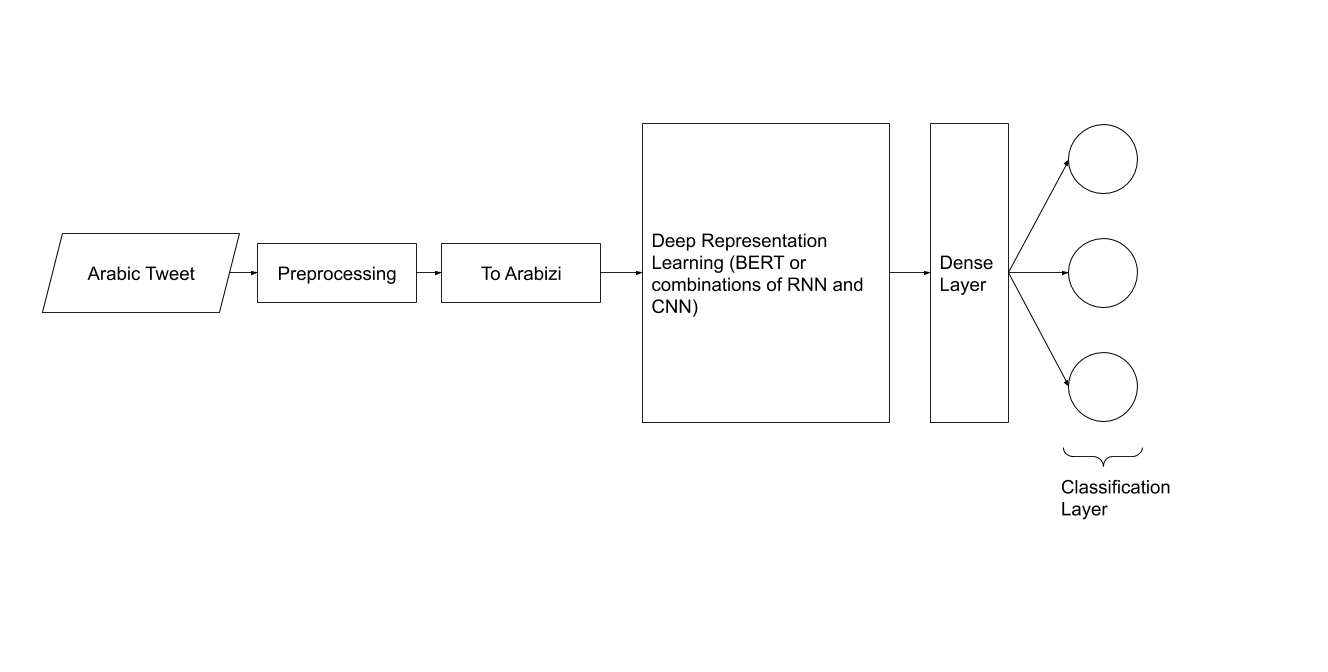


Fig. 2. System architecture of the proposed solution

Model performance will be evaluated based on the F1 score, precision rate and recall.

##### References

1. "NPR Choice page", Npr.org, 2020. [Online]. Available: https://www.npr.org/2019/10/08/768319934/senate-report-russians-used-used-social-media-mostly-to-target-race-in-2016.
2. "Facebook Security Breach Exposes Accounts of 50 Million Users", Nytimes.com, 2020. [Online]. Available: https://www.nytimes.com/2018/09/28/technology/facebook-hack-data-breach.html.
3. "Social media outpaces print newspapers in the U.S. as a news source", Pew Research Center,2020. [Online]. Available: https://www.pewresearch.org/fact-tank/2018/12/10/social-media-outpaces-print-newspapers-in-the-u-s-as-a-news-source/.
4. "Is Social Media Replacing Traditional Advertising?", *Cobizmag.com*, 2020. [Online]. Available: <https://www.cobizmag.com/Business-Insights/Is-Social-Media-Replacing-Traditional-Advertising/>.
5. "Cyberbullying Statistics", 2019. [Online]. Available: https://enough.org/stats\_cyberbullying.
6. "Council on Foreign Relations", 2019. [Online]. Available: <https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons>.
7. "Instagram is using AI to stop people from posting abusive comments", MIT Technology Review, 2020. [Online]. Available: https://www.technologyreview.com/f/613930/instagram-is-using-ai-to-stop-people-posting-abusive-comments/.
8. U.S. (2020). Facebook’s flood of languages leave it struggling to monitor content. [online] Available at: <https://www.reuters.com/article/us-facebook-languages-insight/facebooks-flood-of-languages-leave-it-struggling-to-monitor-content-idUSKCN1RZ0DW>
9. V. Stoyanov and N. Fazil Ayan, "Under the hood: Multilingual embeddings", Ai.facebook.com, 2018. [Online]. Available: <https://ai.facebook.com/blog/under-the-hood-multilingual-embeddings/>.
10. "The importance of the Arabic language in social media | Seven Media", Sevenmedia.ae. [Online]. Available: <https://sevenmedia.ae/the-importance-of-the-arabic-language-in-social-media/>.
11. A. Al-Hassan and H. Al-Dossari, “DETECTION OF HATE SPEECH IN SOCIAL NETWORKS: A SURVEY ON MULTILINGUAL CORPUS,” in Computer Science & Information Technology(CS & IT), 2019, pp. 83–100, doi: 10.5121/csit.2019.90208.
12. B. Haidar, M. Chamoun, and A. Serhrouchni, “A Multilingual System for Cyberbullying Detection: Arabic Content Detection using Machine Learning,” Adv. sci. technol. eng. syst. j., vol. 2, no. 6, pp. 275–284, Dec. 2017, doi: 10.25046/aj020634.
13. "Arabic in Latin Script: Lost in Arabic - Qantara.de", Qantara.de - Dialogue with the Islamic World. [Online]. Available: https://en.qantara.de/content/arabic-in-latin-script-lost-in-arabic.
14. H. Yenala, A. Jhanwar, M. K. Chinnakotla, and J. Goyal, “Deep learning for detecting inappropriate content in text,” Int J Data Sci Anal, vol. 6, no. 4, pp. 273–286, Dec. 2018, doi: 10.1007/s41060-017-0088-4.
15. G. Xiang, B. Fan, L. Wang, J. Hong and C. Rose, "Detecting offensive tweets via topical feature discovery over a large scale twitter corpus", Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM '12, 2012. Available: 10.1145/2396761.2398556.
16. T. Davidson, D. Warmsley, M. Macy, and I. Weber, “Automated Hate Speech Detection and the Problem of Offensive Language,” arXiv:1703.04009 [cs], Mar. 2017.
17. Z. Zhang, D. Robinson, and J. Tepper, “Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network,” in The Semantic Web, vol. 10843, A. Gangemi, R. Navigli, M.-E. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai, and M. Alam, Eds. Cham: Springer International Publishing, 2018, pp. 745–760.
18. A. Gaydhani, V. Doma, S. Kendre, and L. Bhagwat, “Detecting Hate Speech and Offensive Language on Twitter using Machine Learning: An N-gram and TFIDF based Approach,” arXiv:1809.08651 [cs], Sep. 2018.
19. G. K. Pitsilis, H. Ramampiaro, and H. Langseth, “Detecting Offensive Language in Tweets Using Deep Learning,” Appl Int/242-y.
20. V. S. Chavan and Shylaja S S, “Machine learning approach for detection of cyber-aggressive comments by peers on social media network,” in 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Kochi, India, 2015, pp. 2354–2358.
21. M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Darra and R. Kumar, "Predicting the Type and Target of Offensive Posts in Social Media". [Online]. Available: <https://www.researchgate.net/publication/331370899_Predicting_the_Type_and_Target_of_Offensive_Posts_in_Social_Media>.
22. M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, “SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval),” arXiv:1903.08983 [cs], Apr. 2019.
23. S. Sapora, B. Lazarescu, and C. Lolov, “Absit invidia verbo: Comparing Deep Learning methods for offensive language,” arXiv:1903.05929 [cs], Mar. 2019.
24. H. Uglow, M. Zlocha, and S. Zmyślony, “An Exploration of State-of-the-art Methods for Offensive Language Detection,” arXiv:1903.07445 [cs, stat], Mar. 2019.
25. A. Cambray and N. Podsadowski, “Bidirectional Recurrent Models for Offensive Tweet Classification,” arXiv:1903.08808 [cs], Mar. 2019.
26. N. Frisiani, A. Laignelet, and B. Güler, “Combination of multiple Deep Learning architectures for Offensive Language Detection in Tweets,” arXiv:1903.08734 [cs, stat], Mar. 2019.
27. R. Ong, “Offensive Language Analysis using Deep Learning Architecture,” arXiv:1903.05280 [cs], Mar. 2019.
28. A.-B. Puiu and A.-O. Brabete, “Towards NLP with Deep Learning: Convolutional Neural Networks and Recurrent Neural Networks for Offensive Language Identification in Social Media,” arXiv:1903.00665 [cs], Apr. 2019.
29. H. Bansal, D. Nagel, and A. Soloveva, “HAD-Tübingen at SemEval-2019 Task 6: Deep Learning Analysis of Offensive Language on Twitter: Identification and Categorization,” in Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, Minnesota, USA, 2019, pp. 622–627, doi: 10.18653/v1/S19-2111.
30. J. Pavlopoulos, N. Thain, L. Dixon, and I. Androutsopoulos, “ConvAI at SemEval-2019 Task 6: Offensive Language Identification and Categorization with Perspective and BERT,” in Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, Minnesota, USA, 2019, pp. 571–576, doi: 10.18653/v1/S19-2102.
31. R. Pawar and R. R. Raje, “Multilingual Cyberbullying Detection System,” in 2019 IEEE International Conference on Electro Information Technology (EIT), Brookings, SD, USA, 2019, pp. 040–044, doi: 10.1109/EIT.2019.8833846.
32. P. Saha, B. Mathew, P. Goyal, and A. Mukherjee, “HateMonitors: Language Agnostic Abuse Detection in Social Media,” arXiv:1909.12642 [cs], Sep. 2019.
33. T. Ranasinghe, M. Zampieri, and H. Hettiarachchi, “BRUMS at HASOC 2019: Deep Learning Models for Multilingual Hate Speech and Oﬀensive Language Identiﬁcation,” p. 9, 2019.
34. H. Mulki, H. Haddad, C. Bechikh Ali, and H. Alshabani, “L-HSAB: A Levantine Twitter Dataset for Hate Speech and Abusive Language,” in Proceedings of the Third Workshop on Abusive Language Online, Florence, Italy, 2019, pp. 111–118, doi: 10.18653/v1/W19-3512.
35. H. Mubarak, K. Darwish, and W. Magdy, “Abusive Language Detection on Arabic Social Media,” in Proceedings of the First Workshop on Abusive Language Online, Vancouver, BC, Canada, 2017, pp. 52–56, doi: 10.18653/v1/W17-3008.
36. N. Ousidhoum, Z. Lin, H. Zhang, Y. Song, and D.-Y. Yeung, “Multilingual and Multi-Aspect Hate Speech Analysis,” arXiv:1908.11049 [cs], Aug. 2019.
37. R. M. Duwairi, M. Alfaqeh, M. Wardat, and A. Alrabadi, “Sentiment analysis for Arabizi text,” in 2016 7th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 2016, pp. 127–132, doi: 10.1109/IACS.2016.7476098.
38. R. M. Duwairi, R. Marji, N. Sha’ban, and S. Rushaidat, “Sentiment Analysis in Arabic tweets,” in 2014 5th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 2014, pp. 1–6, doi: 10.1109/IACS.2014.6841964.
39. B. Haidar, M. Chamoun, and A. Serhrouchni, “Arabic Cyberbullying Detection: Using Deep Learning,” in 2018 7th International Conference on Computer and Communication Engineering (ICCCE), Kuala Lumpur, 2018, pp. 284–289, doi: 10.1109/ICCCE.2018.8539303.
40. B. Haidar, M. Chamoun, and A. Serhrouchni, “Arabic Cyberbullying Detection: Enhancing Performance by Using Ensemble Machine Learning,” in 2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), Atlanta, GA, USA, 2019, pp. 323–327, doi: 10.1109/iThings/GreenCom/CPSCom/SmartData.2019.00074.
41. H. Mohaouchane, A. Mourhir, and N. S. Nikolov, “Detecting Offensive Language on Arabic Social Media Using Deep Learning,” in 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), Granada, Spain, 2019, pp. 466–471, doi: 10.1109/SNAMS.2019.8931839.
42. D. Mouheb, R. Ismail, S. A. Qaraghuli, Z. A. Aghbari, and I. Kamel, “Detection of Offensive Messages in Arabic Social Media Communications,” in 2018 International Conference on Innovations in Information Technology (IIT), Al Ain, 2018, pp. 24–29, doi: 10.1109/INNOVATIONS.2018.8606030.
43. I. Guellil, H. Saâdane, F. Azouaou, B. Gueni, and D. Nouvel, “Arabic natural language processing: An overview,” Journal of King Saud University - Computer and Information Sciences, p. S1319157818310553, Feb. 2019, doi: 10.1016/j.jksuci.2019.02.006.
44. Darwish, K., 2013. Arabizi detection and conversion to Arabic. arXiv preprint arXiv:1306.6755.

1. https://github.com/Hala-Mulki/L-HSAB-First-Arabic-Levantine-HateSpeech-Dataset [↑](#footnote-ref-0)