Comparison of Machine Learning Techniques for Cyberbullying Detection on YouTube Arabic Comments

Tahani Alsubait[1†] and Danyah Alfageh [2††],

College of Computer and Information Systems,
Umm Al-Qura University, Makkah, Saudi Arabia

Summary
Cyberbullying is a problem that is faced in many cultures. Due to their popularity and interactive nature, social media platforms have also been affected by cyberbullying. Social media users from Arab countries have also reported being a target of cyberbullying. Machine learning techniques have been a prominent approach used by scientists to detect and battle this phenomenon. In this paper, we compare different machine learning algorithms for their performance in cyberbullying detection based on a labeled dataset of Arabic YouTube comments. Three machine learning models are considered, namely: Multinomial Naïve Bayes (MNB), Complement Naïve Bayes (CNB), and Linear Regression (LR). In addition, we experiment with two feature extraction methods, namely: Count Vectorizer and Tfidf Vectorizer. Our results show that, using count vectroizer feature extraction, the Logistic Regression model can outperform both Multinomial and Complement Naïve Bayes models. However, when using Tfidf vectorizer feature extraction, Complement Naive Bayes model can outperform the other two models.

Key words:
Cyberbullying; Arabic dataset; Machine Learning; YouTube.

1. Introduction

As defined by UNICEF [1], cyberbullying is harassment carried out using digital technology. Indeed, cyberbullying, as well as traditional bullying, has negative effects on the bullied ones, requiring parents and educators to intervene and take protective measures. It can be seen that many countries have designed policies and law protocols to deal with cyberbullying acts, aiming to restrict them. Nevertheless, it is not easy to detect cyberbullying as it can take place on various digital outlets such as: social media, messaging sites, gaming platforms and mobile phones. Cyberbullies actions aim at frightening, angering or humiliating their victims. Cyberbullying includes: spreading lies about or sharing embarrassing images of someone on social media, sending hurtful messages or threats via messaging channels, impersonating someone and sending mean messages to others on their behalf.

Many researchers have attempted to study cyberbullying from different angles. In one of the earliest research studies against cyberbullying, Feinberg et al. [2] have urged schools to intervene even in cases of cyberbullying. Smith et al. [3] found that effects of website and text message bullying were equal to traditional bullying. Notably, Abaido [4] conducted a survey on a sample of 200 university students aged between 19-25 and reported that 91% of them strongly agreed that cyberbullying is widely spread amongst Arabic social media communities.

In this paper, we contribute to the cyberbullying research field by suggesting and comparing models that can be used to automate the detection of cyberbullying. In particular, we compare the performance of machine learning classifiers in analyzing Arabic cyberbullying content. More precisely, we examine how the three selected machine learning models perform under two feature extraction settings which is, to the best of our knowledge, has not been done before on an Arabic cyberbullying dataset. Since the first ever video shared over YouTube on April 23 2005, YouTube has been growing to become the most popular video sharing platform nowadays. The easiness of both sharing and watching a YouTube video can be considered a key reason for its popularity across different countries, user interests, and age groups. In addition to sharing videos, users can share textual comments to express their feelings and opinions on each video. These textual comments can be a great source for data analytics and natural language processing (NLP) researchers. It can help us get insights into various topics of interest, including the focus of this research which is cyberbullying acts.
The paper is structured as follows: The first section deals with background information about the data and the machine learning models. Then, a literature review of research on Arabic cyberbullying detection on YouTube is presented. Following that, we discuss the methodology and experimental setup of the research. Lastly, we discuss the results and conclude with the conclusion section.

2. Related Work

In this section, a literature survey in the area of Arabic cyberbullying detection using machine learning classifiers is discussed. For example, Mouheb et al. [5] have introduced a scheme that detects cyberbullying in Arabic text using the following steps: data cleaning and pre-

Manuscript received January 5, 2021
Manuscript revised January 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.1.1

processing, extracting bullying words and assigning weights, detecting cyberbullying comments, and calculating bullying strength and finally classifying the comments. They have used both YouTube and Twitter Arabic comments and they carried out the classification based on their own weighted function that sums the weights of the bullying word after multiplying it by two factors: a factor for the repetition of the bullying word and by the number of letters repetition within the word it-self. While in their latest paper [6], the researchers updated their approach by introducing real-time cyberbullying detection in Twitter. In both works, the researchers did not use machine learning techniques, but, more recently in [7], they have used Naive Bayes (NB) classifier algorithm to detect cyberbullying in both YouTube and Twitter data.

Haidar et al. [8] have introduced an approach for classifying cyberbullying text data using ensemble machine learning, which is a combination of predictive models into a single predictor. The paper is an improvement to a previous work by the authors [9] which used machine learning for Arabic cyberbullying detection where they used Nearest Neighbor, Naive Bayes and Support Vector Machine classifiers.

Rachid et al. [10] have used Convolutional and Recurrent Neural Networks coupled with Arabic pre-trained word embedding for classifying cyberbullying instances on Arabic comments dataset. Furthermore, they have compared the performance of deep learning models against machine learning models and noticed that they show competitive performance to deep learning.

3. Methods and Materials

In this section, an overview of the datasets and algorithms used in this research will be presented and discussed.

3.1 YouTube Annotated Cyberbullying Comments Dataset

For the empirical part of this research, we use a publicly available dataset which contains over 15,000 YouTube comments written by Arabic users. The dataset presented by Alakrot et al. [11, 12] is different and richer than other available datasets as most of cyberbullying datasets tend to be extracted from Twitter data. YouTube comments are different in nature than twitter comments as they allow for longer text which poses an extra challenge for cyberbullying detection. Furthermore, the dataset is the only readily available annotated and published Arabic comments dataset to date. It was collected from YouTube channels that upload videos about celebrities in the Arab world in the period from 2015 to 2017. Celebrities' videos usually attract a lot of engagement from a big audience and they tend to be a target for offensive comments, making such videos a good choice for the empirical work on cyberbullying detection.

The dataset is composed of 14 features for each YouTube comment: ids, user ids, timestamps, comment text, likes, replies and replies data. Additionally, each comment is annotated by three researchers with either P for positive or N for non-bullying or negative. Out of the 15,050 comments in the dataset, the percentage of positive comments is 39%, or 5,817 comments. A comment is labeled as positive if at least two out of the three annotators consider it offensive.

3.2 Machine Learning Models

We experiment with three machine learning algorithms for the purpose of training models for cyberbullying detection. The first one is the Multinomial Naïve Bayes Classifier [13], which determines the number of times the term appears in a document by taking into account that a term maybe significant to the decision on the sentiment of a document. Term frequency is also helpful in deciding whether or not this term is useful in our analysis.

The Complement Naïve Bayes classifier [14] is the second classifier we consider in this study. It is an adaptation of the Multinomial Naïve Bayes that uses statistics from the complement of each class to determine the model's weights.

The last algorithm we consider in this study is the Logistic Regression algorithm. It is a linear classification model [15] that takes a variable vector and evaluates the weights for each variable and predicts the given item class as a vector.

4. Empirical Work

In this section, we present the steps used to train the three machine learning algorithms for cyberbullying detection.

4.1 Preprocessing

For the data pre-processing stage, we remove all attributes of the dataset and only keep comment text and final annotated label. We also needed to clean the comments. As per the comment text we clean it in following steps:

- Cleaning the text: this is done by removing numbers, non-Arabic words, symbols, punctuation, URLs, and hashtags to clear the data from noise. then we remove Arabic diacritics and stop words.
- Normalization: we substitute different

representations of letters by their standard form.

- Stemming: all words are returned to their root form to reduce the features.

After cleaning and pre-processing and sampling we are left with 9500 comments for the next steps

4.2 Feature Extraction

Feature extraction is used in machine learning to reduce dimensionality. In this paper, we have chosen to use two feature extraction approaches: Count Vectorizer and $Tfidf$ Vectorizer. Both Count Vectorizer and $Tfidf$ Vectorizer convert text data into machine readable format.

Count Vectorizer returns an encoded vector that has the same length of the entire comment and an integer count for the number of times each word appeared in a comment.

$Tfidf$ Vectorizer, which stands for Term Frequency – Inverse Document Frequency ($TF - IDF$), is calculated by the following formulas, for a term t of a document d in a document set:

$$tf - idf(t,d) = tf(t,d) * idf(t) \qquad (1)$$

while idf is computed as:

$$idf(t) = log[n/df(t)] + 1 \qquad (2)$$

where n is the number of documents in the document set and $df(t)$ is the document frequency of t; the document frequency is the number of documents in the set that include the term t.

4.3 Model training

To prepare the models, we first split our data to 80% training and 20% testing subsets. Then three models are trained and tested on the data. Then we fit Multinomial Naïve Bayes, Complement Naïve Bayes and Linear regression models to the data and use them to predict the classification. Finally, we measure the model performance using the F1 score measure.

The F1 score measure [16] is the harmonic mean of precision and recall. F1 score values range from zero to one, and higher values suggest a more accurate model classification. The F1 score can be calculated as follows:

$$F1 = \frac{Precision \times Recall}{Precision + Recall} \qquad (3)$$

where recall represents the ratio of the positive correctly classified samples to the total number of positives. While, Precision is the ratio of correctly classified positive samples to the count of predicted positive samples.

5. Results and Discussion

As shown in Table 1, we have used the F1 score to get the results of the performance of three popular text machine learning classifiers on Arabic comments on the YouTube platform. We found that Complement Naive Bayes classifier did better with tfidf vectorizer feature extraction with the dataset. While, Logistic regression did the best with count vectorizer feature extraction. Overall, in regards to feature extraction, the models slightly give better results if tfidf vectorizer is used where the average of F1 scores for all models is 77.9% while count vectorizer's performance has an average of 77.5%.

Table 1: F1 Scores per Model

Feature Extraction Setting	Multinomial NB	Complement NB	Logistic Regression
Count Vectorizer	78.4%	76.6%	78.6%
Tfidf Vectorizer	77.0%	78.5%	76.8%

In count vectroizer feature extraction, as can be seen in Fig. 1, the Logistic Regression model has outperformed both Multinomial and Complement Naïve Bayes models. While, in $Tfidf$ vectorizer experiment, as can be seen in Fig. 2, Complement Naive Bayes model has outperformed Multinomial Naive Bayes and Logistic Regression models.

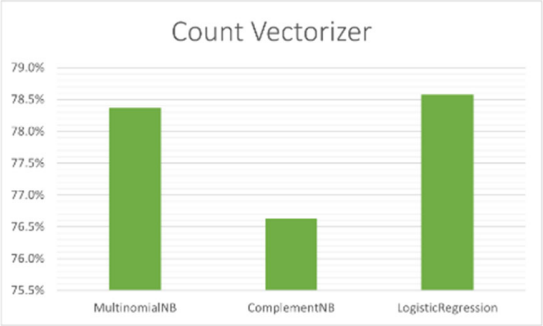

Fig. 1: F1 Scores of the models with Count Vectorizer

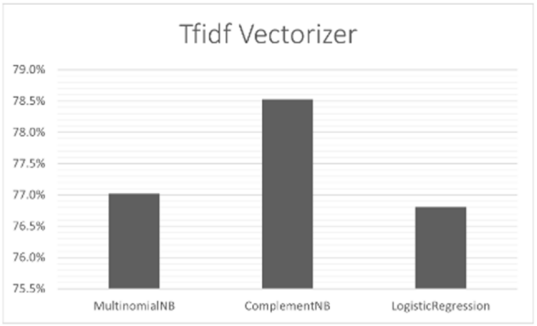

Fig. 2: F1 Scores of the models with Tfidf Vectorizer

.

6. Conclusion

Cyberbullying on social networking is a phenomenon that has been negatively affecting users in the Arab world. This paper compares three popular machine learning approaches in identifying cyberbullying in Arab YouTube comments. Furthermore, two feature extraction approaches were used and studied. The evaluation of the models' performance was carried out on an annotated dataset of Arabic comments. Lastly, results were compared using F1 score. We are hoping to provide guidance in this paper that can help researchers to use the best machine learning approaches to keep the social media safe and secure for the users.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] UNICEF. Cyberbullying: What is it and how to stop it. Feb. 2020.

[2] Ted Feinberg and Nicole Robey. "Cyberbullying". In: The education digest74.7 (2009), p. 26.

[3] Peter K Smith, Jess Mahdavi, Manuel Carvalho, Sonja Fisher, Shanette Russell, and Neil Tippett. "Cyberbullying: Its nature and impact in secondary school pupils". In: Journal of child psychology and psychiatry 49.4 (2008), pp. 376–385.

[4] Ghada M Abaido. "Cyberbullying on social media platforms among university students in the United Arab Emirates". In: International Journal of Adolescence and Youth 25.1 (2020), pp. 407–420.

[5] Djedjiga Mouheb, Rutana Ismail, Shaheen Al Qaraghuli, Zaher Al Aghbari, and Ibrahim Kamel. "Detection of Offensive Messages in Arabic Social Media Communications". In: 2018 International Conference on Innovations in Information Technology (IIT). IEEE. 2018, pp. 24–29.

[6] Djedjiga Mouheb, Masa Hilal Abushamleh, Maya Hilal Abushamleh, Zaher Al Aghbari, and Ibrahim Kamel. "Real-time detection of cyberbullying in Arabic twitter streams". In: 2019 10th IFIP International Conference on New Technologies, Mobility and Security (NTMS). IEEE. 2019, pp. 1–5.

[7] Djedjiga Mouheb, Raghad Albarghash, Mohamad Fouzi Mowakeh, ZaherAl Aghbari, and Ibrahim Kamel. "Detection of Arabic Cyberbullying on Social Networks using Machine Learning". In:2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA). IEEE. 2019, pp. 1–5.

[8] Batoul Haidar, Maroun Chamoun, and Ahmed 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 (Smart-Data). IEEE. 2019, pp. 323–327.

[9] Batoul Haidar, Maroun Chamoun, and Ahmed Serhrouchni. "A multilingual system for cyberbullying detection: Arabic content detection using machine learning". In: Advances in Science, Technology and Engineering Systems Journal 2.6 (2017), pp. 275–284.

[10] Benaissa Azzeddine Rachid, Harbaoui Azza, and Hajjami Henda Ben Ghezala."Classification of Cyberbullying Text in Arabic". In: 2020 International Joint Conference on Neural Networks (IJCNN). IEEE. 2020, pp. 1–7.

[11] Azalden Alakrot, Liam Murray, and Nikola S Nikolov. "Dataset Construction for the Detection of Anti-Social Behaviour in Online Communication in Arabic". In: Procedia computer science. vol.142 (2018), pp. 174-181.

[12] Azalden Alakrot, Liam Murray, and Nikola S Nikolov. "Towards accurate detection of offensive language in online communication in Arabic". In: Procedia computer science. vol142 (2018), pp. 315–320.

[13] G. Singh, B. Kumar, L. Gaur, and A. Tyagi. "Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification". In:2019 International Conference on Automation, Computational and Technology Management (ICACTM). 2019, pp. 593–596.doi:10.1109/ICACTM.2019.8776800.

[14] Jason D Rennie, Lawrence Shih, Jaime Teevan, and David R Karger. "Tack-ling the poor assumptions of

naive bayes text classifiers". In: Proceedings of the 20th international conference on machine learning (ICML-03). 2003, pp. 616–623.

[15] A. Prabhat and V. Khullar. "Sentiment classification on big data using Naïve bayes and logistic regression". In: 2017 International Conference on Computer Communication and Informatics (ICCCI). 2017, pp. 1–5.doi:10.1109/ICCCI.2017.8117734.

[16] Alaa Tharwat. "Classification assessment methods". In: Applied Computing and Informatics (2020).