Journal of Information Technology Management

Print ISSN: 2008-5893

Online ISSN: 2423-5059

Using Machine Learning Algorithms for Automatic Cyber Bullying Detection in Arabic Social Media

Bedoor Y. AlHarbi
Department of Information Technology, College of Computer, Qassim University, Saudi Arabia. ORCID: 0000-0002-3125-9248. E-mail: 362218730@qu.edu.sa

Mashael S. AlHarbi
Department of Information Technology, College of Computer, Qassim University, Saudi Arabia. ORCID: 0000-0002-8919-8900. E-mail: mashae1.017o@gmail.com

Nouf J. AlZahrani
Department of Information Technology, College of Computer, Qassim University, Saudi Arabia. ORCID: 0000-0001-9696-8662 E-mail: noufjamman0@gmail.com

Meshaiel M. Alsheail
Department of Information Technology, College of Computer, Qassim University, Saudi Arabia. ORCID: 0000-0002-2991-7299. E-mail: m.alsheail@qu.edu.sa

Dina M. Ibrahim
Department of Information Technology, College of Computer, Qassim University, Saudi Arabia. ORCID: 0000-0002-7775-0577. E-mail: d.hussein@qu.edu.sa

Abstract

Social media allows people interact to express their thoughts or feelings about different subjects. However, some of users may write offensive twits to other via social media which known as cyber bullying. Successful prevention depends on automatically detecting malicious messages. Automatic detection of bullying in the text of social media by analyzing the text "twits" via one of the machine learning algorithms. In this paper, we have reviewed algorithms for automatic cyberbullying detection in Arabic of machine learning, and after comparing the highest accuracy of these classifications we will propose the techniques Ridge Regression (RR) and Logistic Regression (LR), which achieved the highest accuracy between the various techniques applied in the automatic cyberbullying detection in English and between the techniques that was used in the sentiment analysis in Arabic text, The purpose of this work is applying these techniques for detecting cyberbullying in Arabic.

Keywords: Cyberbullying, Machine Learning (ML), Sentiment analysis, Cyberbullying Detection in Arabic.

DOI: 10.22059/jitm.2020.75796

Introduction

Cyber bullying is defined as the use of the Internet, cell phones, video game systems, or any other techniques to send messages, publish texts or photos prepared to harm or pose another person or group of people, or other deliberate action by one person or group of persons. Through digital means such as sending messages or posting comments against the victim. In recent years, social media are a place where people engage in social interaction (Zhao & Mao, 2016). However, given the current situation, some users are negative for social media. We are truly living in the information age where the data is generated by both humans and machines at an unprecedented rate, therefore it's nearly impossible to gain insights into such data for making intelligent decisions, manually (Alharbi et al., 2019). Sentiment analysis can help detect cyberbullying after processing a large amount of data by using sentiment classification. This classification targets to classify the text automatically. It can be classified into lexicon-based, machine-learning, and hybrid techniques (Feng et al., 2018).

Sentiment Analysis Background

Sentiment analysis, also called opinions and its related concepts such as sentiments, evaluations, attitudes, and emotions are the subjects of study of sentiment analysis and opinion mining, sentiment analysis applies to text that contains people's opinions. Some examples of analysis topics are opinions about products, services, food, education, etc. (Kotsiantis, Zaharakis & Pintelas, 2006). A new method of research involves where there are multiple areas such as natural language processing, computational linguistics, text analysis, automated learning, and artificial intelligence. The sentiment analysis uses many techniques in machine learning SVM, NB etc.

Sentiment Analysis is normally performed using one of two fundamental approaches:

a) Lexicon-based approach, in which rules separated from the linguistic study of a language are applied to the sentiment analysis.
b) Machine Learning (ML) approach, which depends on the well-known ML algorithms to understand sentiment analysis as a classification task (Gamal et al., 2019).

The algorithms used in Sentiment Analysis are many and the most important ones which we will focus on are the Ridge Regression (RR) and Logistic Regression (LR). RR is a technique for analyzing multiple regression data that suffer from multicollinearity (Number Cruncher Statistical Systems (NCSS), 2007).

The RR gives a better accuracy than other ML Algorithms for Sentiment classification on the twitter dataset (Gamal et al., 2019). Figure 1 visualizes the performance of Support Vector Machine (SVM), Naive Bayes (NB), K-nearest neighbors (KNN), Deep neural network (DNN), LR and C5.0 in terms of accuracy.

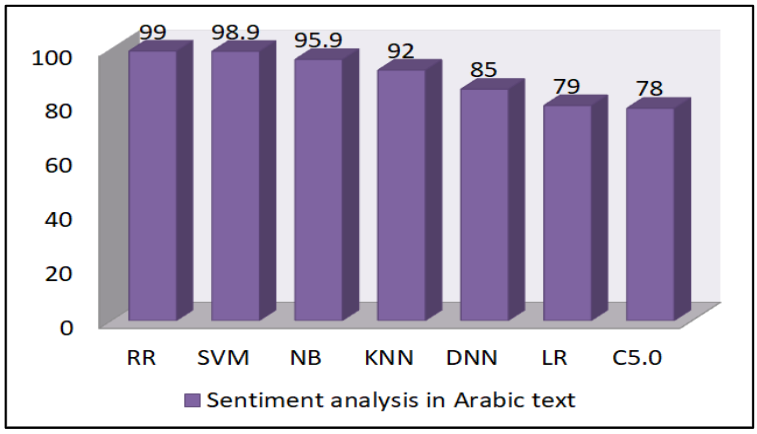

Figure 1. Sentiment Analysis in Arabic Text

Cyberbullying Detection Using Machine Learning

Machine learning uses algorithms to analyze data, learn from those data, and make decisions based on their learning, and the algorithms divided into: Supervised Machine learning, unsupervised learning.

A. Supervised Machine learning: Is the search for algorithms that cause external resource situations to provide general hypotheses, which make predictions about future situations: decision trees and rule-based classifiers (Kotsiantis, Zaharakis and Pintelas, 2006).

B. Unsupervised learning: Is where you only have input data (X) and no corresponding output variables. The aim of unsupervised learning is to model the basic structure or data distribution in order to learn more about the data, Unsupervised learning problems can be additionally classified into clustering and association problems.

Because the detection of cyberbullying is well-known, then research about supervised machine learning algorithm.

Cyberbullying Detection Techniques in English

The Authors in (Pradheep et al., 2018) using the NB workbook to classify cyberbullying data in physical, social and verbal bullying. The proposed system works on detection in images, video, and audio from social networks. (Van Hee et al..2018) Use of LSVM. Experiments on a hold-out test set reveal promising results for the detection of cyberbullying-related posts achieved good accuracy. (Mangaonkar, Hayrapetian, and Raje, 2015) They worked on equal representation groups that contained a balanced data set. Figure 2 shows the accuracy ratio and the comparison of most of the techniques used in English.

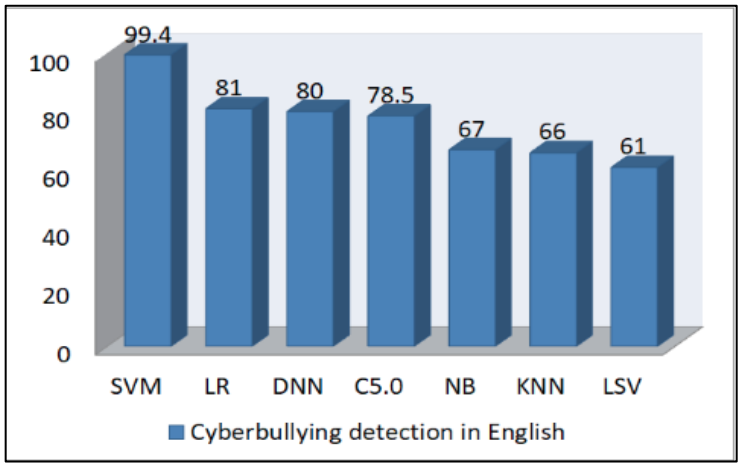

Figure 2. Cyberbullying Detection in English

The result confirms that the discriminant model LR works slightly better than the structural model NB, and the SVM is better than LR. Other research as in (Nurrahmi & Nurjanah, 2018) utilized SVM and KNN to detect and also learn more about cyberbullying texts. The results illustrated that SVM outcomes in the highest value of F1-score, that is approximately 68%. DNN is an artificial neural network (ANN) with multiple layers between the input and output layers and used to detection cyberbullying Motivated by neural networks documented success, three architectures are implemented from similar works: a simple Convolution neural network (CNN), a hybrid long short term memory (CNN-LSTM) and a mixed CNN-LSTM-DNN (Mangaonkar, Hayrapetian, & Raje, 2015). In Table 1, the latest studies and modern techniques used are presented.

Table 1. Techniques Used in English

Years	Techniques	Paper
2018	LSVM	Automatic detection of cyberbullying in social media text (Van Hee et al., 2018)
2018	SVM LR C-LSTM CNN CNN-LSTM	A “Deeper” Look at Detecting Cyberbullying in Social Networks (Rosa et al, 2018)
2017	SVM LR NB	Common discovery of Cyberbullying behavior in Twitter data
2017	SVM	Cyberbullying System Detection and Analysis
2016	SmSDA	Cyberbullying Detection based on Semantic-Enhanced Marginalized Denoising Auto-Encode
2016	NB	Twitter Bullying Detection
2015	CRF SVM	Understand and Fighting bullying with machine learning
2012	C4.5	Using Machine Learning to Detect Cyberbullying

Cyberbullying Detection Techniques in Arabic

With the development of modern technologies, it has become possible to discover cyberbullying. There is little research in Arabic to detect cyberbullying Here we will mention some studies of cyberbullying in Arabic:

In (Pradheep et al., 2018) the system is designed to prevent cyberbullying attacks, by detecting and stopping them. Use natural language processing (NLP) to identify and process the Arabic language. ML techniques are used to classify cyberbullying content Training and classification procedures were done several times for the purpose of reaching the best results. Two models were chosen in the first stage, NB and SVM. Researchers reached a conclusion that those are the best two algorithms for text classifications. A comparison between SVM and NB showed that SVM outperformed NB.

In (Haidar, Chamoun, & Serhrouchni, 2018) used Deep Learning as solution that employs Deep Learning methods in the process of Arabic cyberbullying detection. Specifically, a Feed Forward Neural Network (FFNN) is trained on an Arabic dataset for the purpose of cyberbullying detection. The dataset was originally labeled "1" was used for bullying content and "0" for non-bullying. At first the FFNN model was built with 4 hidden layers. The model was configured to split the dataset into 80% training and 20% testing and shuffle the dataset at each epoch. The model was trained on the "small" dataset, but the achieved results were not promising. The best validation accuracy achieved was 66.67%, which is not a considerable result concerning Deep Learning methods. The "large" dataset was used in training

A second model instead of the small dataset. The performance metrics enhanced instantly to 91.17% a better accuracy was achieved. The goal of this paper was to establish a better performance from classic ML methods presented previously, and this goal was achieved.

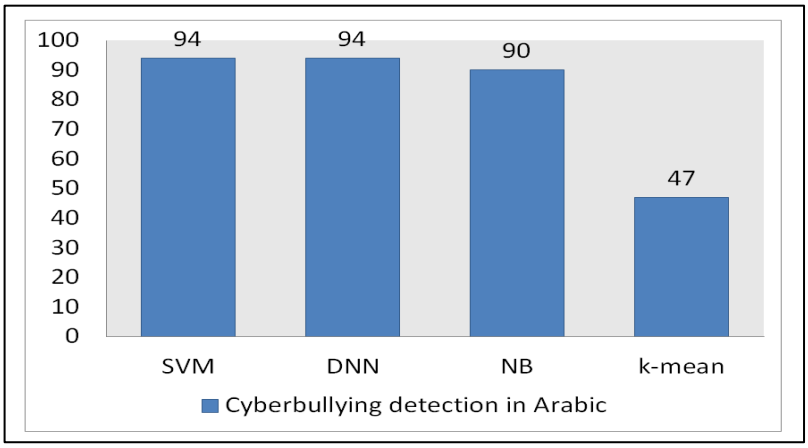

Figure 3. Cyberbullying Detection in Arabic

The following shows the accuracy ratio and the comparison of most of the techniques used in Arabic, we explained the comparison as in Figure 3 and in Table 1, the latest studies and modern techniques used are presented.

Table 2. Techniques Used in Arabic

Years	Techniques	Paper
2018	DNN	Arabic Cyberbullying Detection: Using Deep Learning
2017	NB SVM	A Multilingual System for Cyberbullying Detection: Arabic Content Detection Using Machine Learning
2017	K-mean	Uncensored Discovery of Violent Content in Arab Social Media

The Analysis Study of the Current Techniques

Through this research, there are many techniques that detect cyberbullying in English, and techniques for sentiment analysis in Arabic. On the other hand, there are few techniques in Arabic to detect cyberbullying, as shown Figure 4 which is compare cyberbullying detection algorithm in English, cyberbullying detection in Arabic, and sentiment analysis in Arabic text.

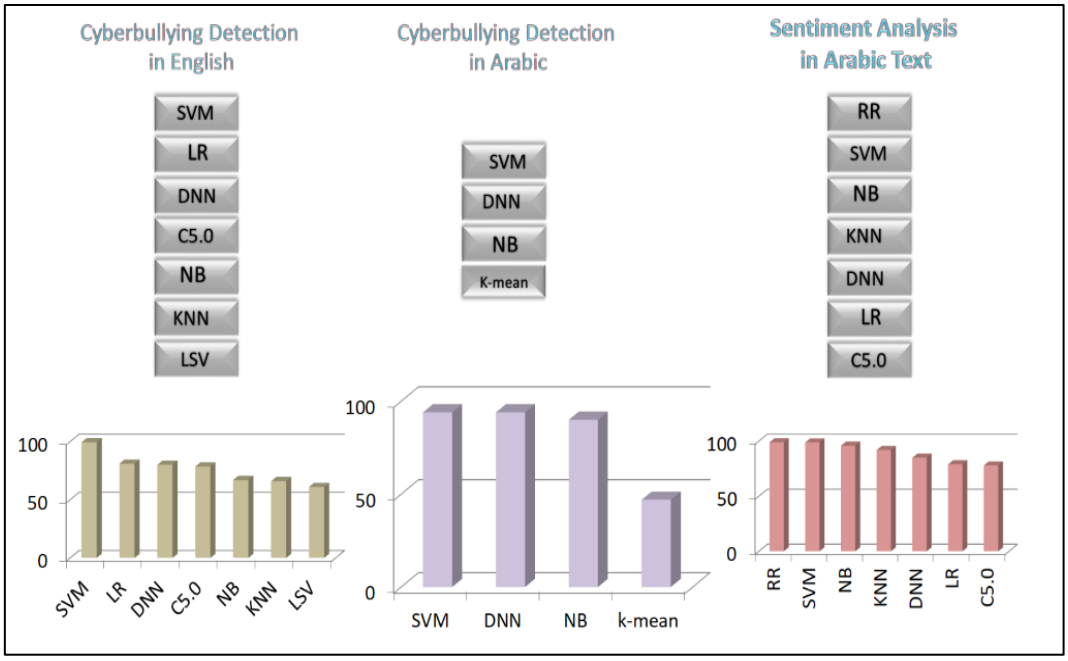

Figure 4. Analysis Study of the Current Techniques

Several measures are taken to contain and stop the phenomenon of cyberbullying. Among those measures are Sentiment Analysis solutions, it is more accurate for the words in

the Arabic language, which contains a lot of dialects and different from the English language. Initially, one of the techniques listed in the Fig 4 will be used, we have observation through our research that RR in Sentiment Analysis for Arabic text in the first order and RR is an extension for linear regression. It's basically a regularized linear regression model. RR is a technique for analyzing multiple regression data that suffer from multicollinearity, Ridge Regression reduces the standard errors. And the Logistic Regression used in detection cyberbullying in English is second in terms of accuracy in English and has also been applied in Sentiment Arabic. Following these techniques, we propose in this research a comparison between them and analyze the performance of the two techniques compared to current techniques then apply high accuracy the techniques to detect cyberbullying in Arabic.

Conclusion

After comparing the highest accuracy of these classifications we will use the techniques Ridge Regression (RR) and Logistic Regression (LR), which achieved the highest accuracy between the various techniques applied in the automatic cyberbullying detection in English and between the techniques that was used in the Sentiment techniques for automatic cyberbullying detection in Arabic. In future, will apply Ridge Regression (RR) and Logistic Regression (LR) and compare the accuracy with the existence techniques that used for automatic detection of cyberbullying in Arabic to reduce cyber bullying.

References

Alharbi, B. Y., Alharbi, M. S., Alzahrani, N. J., Alsheail, M. M., Alshobaili, J. F., & Ibrahim, D. M. (2019). Automatic Cyber Bullying Detection in Arabic Social Media. *Int J Engineering Research and Technology*, *12*(12), 2330-2335.

Feng, J., Gong, C., Li, X., & Lau, R. Y. (2018). Automatic Approach of Sentiment Lexicon Generation for Mobile Shopping Reviews. *Wireless Communications and Mobile Computing*, *2018*.

Gamal, D., Alfonse, M., El-Horbaty, E. S. M., & Salem, A. B. M. (2019). Twitter benchmark dataset for Arabic sentiment analysis. *Int J Mod Educ Comput Sci*, *11*(1), 33.

Haidar, B., Chamoun, M., & Serhrouchni, A. (2018, September). Arabic Cyberbullying Detection: Using Deep Learning. In *2018 7th International Conference on Computer and Communication Engineering (ICCCE)* (pp. 284-289). IEEE.

Mangaonkar, A., Hayrapetian, A., & Raje, R. (2015, May). Collaborative detection of cyberbullying behavior in Twitter data. In *2015 IEEE international conference on electro/information technology (EIT)* (pp. 611-616). IEEE.

Number Cruncher Statistical Systems (NCSS), (2007) “Statistical system for Windows,”.

Nurrahmi, H., & Nurjanah, D. (2018, March). Indonesian Twitter Cyberbullying Detection using Text Classification and User Credibility. In *2018 International Conference on Information and Communications Technology (ICOIACT)* (pp. 543-548). IEEE.

Pradheep, T., Sheeba, J. I., Yogeshwaran, T., & Pradeep Devaneyan, S. (2017, December). Automatic Multi Model Cyber Bullying Detection from Social Networks. In *Proceedings of the International Conference on Intelligent Computing Systems (ICICS 2017–Dec 15th-16th 2017) organized by Sona College of Technology, Salem, Tamilnadu, India*.

Rosa, H., Matos, D., Ribeiro, R., Coheur, L., & Carvalho, J. P. (2018, July). A "deeper" look at detecting cyberbullying in social networks. In *2018 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

Van Hee, C., Jacobs, G., Emmery, C., Desmet, B., Lefever, E., Verhoeven, B., ... & Hoste, V. (2018). Automatic detection of cyberbullying in social media text. *PloS one*, *13*(10).

Zhao, R., & Mao, K. (2016). Cyberbullying detection based on semantic-enhanced marginalized denoising auto-encoder. *IEEE Transactions on Affective Computing*, *8*(3), 328-339.

Bibliographic information of this paper for citing:

AlHarbi, B.Y., AlHarbi, M.S., Alzahrani, N.J., Alsheail, M.M., & Ibrahim, D.M. (2020). Using Machine Learning Algorithms for Automatic Cyber Bullying Detection in Arabic Social Media. *Journal of Information Technology Management*, 12(2), 123-130.