UTM UNIVERSITI TEKNOLOGI MALAYSIA
INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

Cyberbully Detection Using Term Weighting Scheme and Naïve Bayes Classifier

Rafeena Mohamad Rabii & Maheyzah Md Siraj
School of Computing Faculty of Engineering
Universiti Teknologi Malaysia
81310 UTM Johor Bahru, Johor, Malaysia
Email: rafeena1988@graduate.utm.my; maheyzah@utm.my

Submitted: 20/01/2020. Revised edition: 30/04/2020. Accepted: 1/05/2020. Published online: 20/05/2020
DOI: https://doi.org/10.11113/ijic.v10n1.254

***Abstract*—The internet especially social media has been a major platform where people interact with each other. We are able to interact with each other regardless of time and place because of the advancement of technology. Unfortunately, not all of the interaction that goes on are good or positive. One of the negative interaction that can happen online is cyberbullying which has rapidly increase throughout the years, whether it be through social media, emails or texting. Therefore, it is important to prevent cyberbullying from occurring which is why this research is done. Detecting the presence of cyberbullying is one if the main issue in avoiding it from happening. Cyberbullying detection can be challenging because the many languages used in the world, most of the time slangs and informal languages are used and special characters like emoji are also used during online conversation. The aim of this research is to detect the presence of text cyberbullying from online post. Two term weighting schemes and two classification algorithms are compared in this research. The weighting schemes used namely Entropy and Term Frequency-Inverse Document Frequency (TF-IDF) for feature selection and Naïve Bayes algorithm is used and compared with Support Vector Machine (SVM) algorithm. As a result, it shows that Naïve Bayes classifier yields a better accuracy when used with TF-IDF which is 97.60%. Hopefully this research is able give other researchers an insight, particularly to those who are interested in a similar area.**

***Keywords*—Cyberbullying, text mining, detection**

I. INTRODUCTION

Human communication methods has changed through the years, now we are mostly communicating online, whether it be emails or social media such as Facebook, Twitter and Instagram. These advancement has given us a wonderful ability to connect with each other no matter where we are, however they also increase the ability for us to harm others no matter where we are, one example is cyber bullying.

Social media sites has been a platform of spreading lies, anger and hateful comments. Day by day people are getting more comfortable expressing their emotions on the web, which can result to cyberbullying.

Cyberbullying can be defined as willful and repeated harm inflicted through the use of computers, cell phones, and other electronic devices [22]. Cyberbullying has become a major problem compared to conventional bullying, this is because it's easier, faster and can be done anonymously. People resort to cyberbullying because of their anger, frustration, need for revenge or they feel the need to be in control and powerful.

Cyberbullying can come in many forms, it includes rumours, inappropriate or embarrassing photos or videos and threatening, offensive or harassing massages posted on social networks or sent through emails.

Cyberbullying can have long term effect towards the victim. The victim will have lack of confidence, depression, and many more undesirable outcomes that will harm the victim mentally and might affect them their whole life.

There are websites that filter or block contents that are deemed inappropriate, at times this might not be enough. This is because sometimes the damage has been done where the victim has seen the inappropriate text, photos or video. The detection of cyberbully should be improved where one day maybe with the help of humans, the message that is intended for the victim will never be send and there will be a system that always monitor the website.

II. PROBLEM BACKGROUND

We can now see that countless people spend most of their time online whether shopping, making transaction or doing research, but mostly people would go online to go to social media websites. This is where most of the interaction with other people occurs. There are also other notable platforms where people interact, namely emails, blogs, chat rooms and instant messages applications.

Cyberbullying can be done using phones via texting or using computers where damaging inappropriate photos can be circulated, rumours can be spread and harsh and hateful words are exchanged. The regrettable thing is that cyberbullying can be done anywhere and anytime to anyone in the world and sometimes it can be done anonymously. Detection of cyberbully text can be quite complex, because there are slang or new terms which not all people can understand. Misspelling can occasionally occur within a text and more than one language can be used in a sentence.

When classifying documents that contains cyberbullying using a specific algorithm the problem that can occur is misclassifying, this is when documents that is non-cyberbully is classified as cyberbully also called false positive. Some documents can also be classified as cyberbully when they do not have cyberbully contents, this is call false negative. The low accuracy is effected by the false positive and false negative of the dataset.

III. STATEMENT OF PROBLEM

The detection of cyberbullying can be an interesting yet complicated undertaking, particularly if it is done on a group of posts. Usually, the main concern in classifying online post or comments in social media or chatroom is to accurately extract languages that are used and determine them as cyberbully or non-cyberbully.

Due to the many languages that is available it contributes to the difficulty to detect cyberbullying. It is possible to find multiple languages within a text. Since there are countless people with the ability to speak more than one language, they might use more than one language within their online post.

People online do not usually use formal language for their post of conversation. Slangs and short forms are typically used. Examples slangs are ‘noob’ and ‘meh’ of short forms are, ‘btw’ ‘lmao’ and ‘DM’. These words carry a meaning of their own however they are understood only by certain demographic and not considered as an official language. Therefore, these types of words might not be detected.

Special characters, emoticons and emoji are also used widely in the internet which also contributes to the noise of the data. Even though some of these emoticons in useful to convey various massages, these emoticons however are not contributing to increase the accuracy of detection.

The objectives of this research are;

i) To clean the dataset using pre-processing techniques of removal of special characters, tokenization, stop word removal, stemming and transform cases.
ii) To select features from the Formspring post using Entropy and TF-IDF.
iii) To evaluate and verify the effectiveness of the proposed approach in terms of accuracy, precision, recall and F score.

IV. RESEARCH METHODOLOGY

The research begins at the first phase. This phase includes data collection, pre-processing and resulting in a data that is ready for the next phase. Feature selection and classification is part of the second phase. During this phase term weighting schemes is applied to select relevant terms related to cyberbullying. Then the online comments are classified as bullying or non - bullying. Lastly is the third phase, during this phase the term weighting scheme and the classifiers performance used during this research are evaluated. Fig. 1 shows the steps that are carried out to achieved the target of the project.

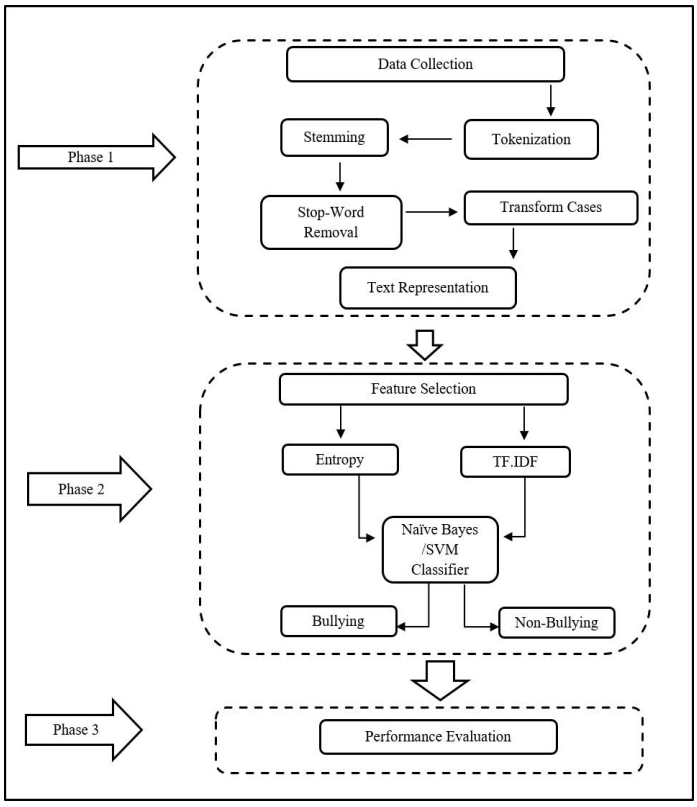


Fig. 1. Research Framework

V. DATASET

This study will use the data set obtained from Reynolds *et al*. (2011). They crawled the data from Formspring.me. The

information was extracted from 18,554 users that were randomly selected. The data contained around twelve thousand post. There are 500 post that is randomly selected for this research. The data was labelled by Amazons Mechanical Turk, which is a web service.

VI. RESEARCH DESIGN AND IMPLEMENTATION

A. Pre-Processing

This process is to ensure the raw data is transformed into a format that is understandable. Unwanted data will be removed and the wanted features will be extracted. This will improve the quality of the data and making sure it is ready for the next step The data that was obtain might have noises, they can be remove by using algorithm for pre-processing. Raw set will be presented as the result to make it easier to classify the data. Fig. 2 shows the process involved in pre-processing.

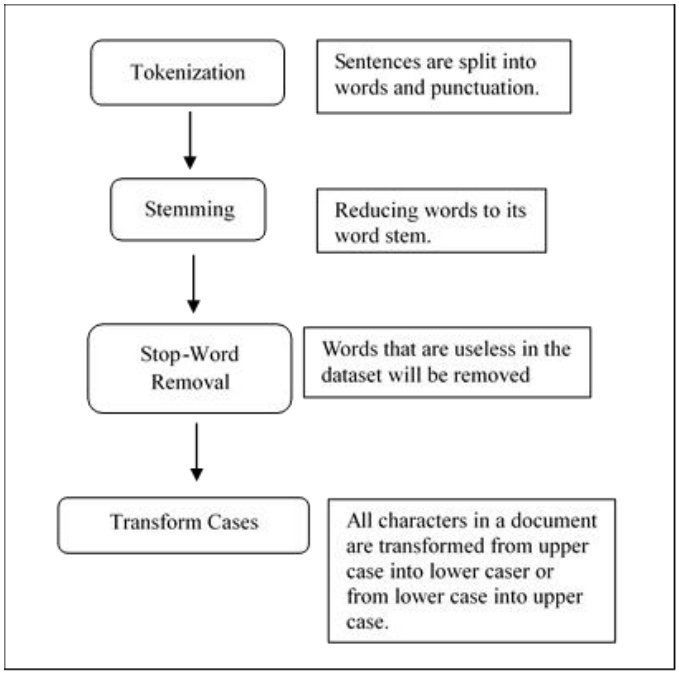


Fig. 2. Pre-Processing

B. Feature Selection

This step is important to eliminate features that are redundant. This is because not all not all features are necessary to be used for the research. The result from the previous phase will be used in this phase. During this research, the weighting schemes that will be used are Entropy and TFIDF.

This process is done to reduce the amount of features, only the best features are selected. Term feature ranking reduces dimension of term but not the documents. The most relevant features will be obtained for the purpose of data training and testing.

1. Entropy

This scheme will calculate the Local and Global weight. The length of the document will be taken into account during calculations. Entropy assumes a document is more significant, as compared to the term appeared in many documents in the collection.

2. Term Frequency Inverse Document Frequency (TFIDF)

Term frequency (TF) measures the frequency of a term appears in a document, in order to normalize the data, term frequency is divided by the length of document. Inverse document frequency (IDF) assumes that the importance of a term relative to a document is inversely proportional to the frequency of occurrence of this term in all the documents. According to this method the higher number of terms in a document the more significant the weight value. TFIDF is calculated on the product of TF and IDF. The TF and IDF is calculated separately for TFIDF.

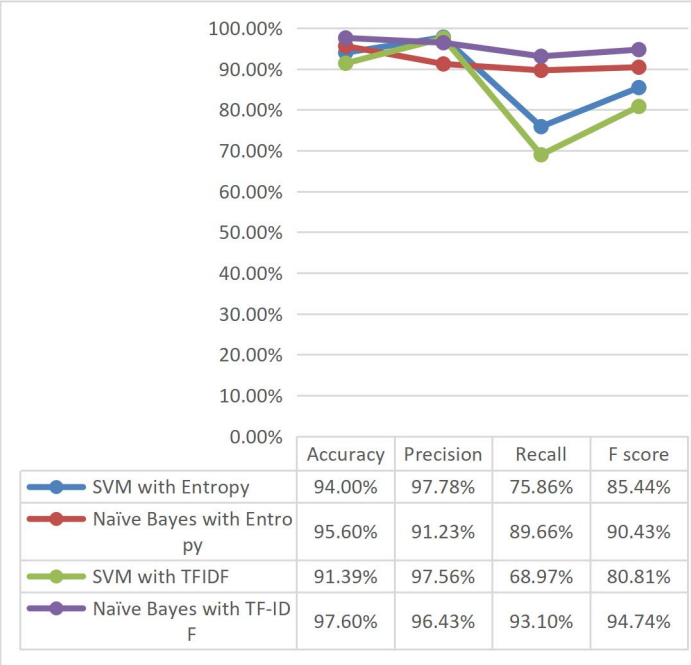
C. Classification

Rapidminer software is used for this research, from where classifiers such Naïve Bayes and SVM can be used. Rapidminer has included these classifiers in its software. The input of data into the classifier involves label and id settings to ensure that the classifier and Rapidminer is able to read and process the data. Rapidminer has an interface that is user friendly and that makes adding data into the software easy. Firstly, dataset is used as an input to train the classifiers. Then the dataset is tested. Lastly, the result will be evaluated where the Accuracy, Precision, Recall and F measure parameters are set.

VII. RESULT

This is the last step for this research. The weighting schemes and classification performance are measured. The accuracy of finding contents related to cyberbullying will be determined during this phase.

Results are evaluated for its accuracy, recall and F score. Fig. 3 below shows the result, which is the pattern of each of the classifier with each of the term weighting schemes based on their accuracy, precision, recall and F score.



	Accuracy	Precision	Recall	F score
SVM with Entropy	94.00%	97.78%	75.86%	85.44%
Naïve Bayes with Entropy	95.60%	91.23%	89.66%	90.43%
SVM with TFIDF	91.39%	97.56%	68.97%	80.81%
Naïve Bayes with TF-IDF	97.60%	96.43%	93.10%	94.74%

Fig. 3. Result of Performance Evaluation Measurement

From the result, it shows that Naïve Bayes accuracy is better than SVM, however SVM give better precision than Naïve Bayes. For Recall, Naïve Bayes have higher percentage compared to SVM. For F score, Naïve Bayes also give better performance than SVM. SVM give better result when used together with entropy as shown in the result that it gives higher percentages for accuracy, precision recall and F score. While Naïve Bayes resulted in higher percentages for accuracy, precision recall and F score when used together with TF-IDF. The best accuracy is Naïve Bayes with TF-IDF with 97.60%. The highest precision is shown by SVM with entropy with 97.78%. The best recall percentage is shown by Naïve Bayes with TF-IDF which resulted in 93.10%. Naïve Bayes with TF-IDF also have the best F score with 94.74%. In conclusion, from the result it shows that Naïve Bayes with TF-IDF give a good performance to classify whether a data has cyberbullying elements.

VIII. CONCLUSION

The aim of this research is to detect the presence of text cyberbullying from online post using term weighting scheme and Naïve Bayes Classifier. The term weighting scheme used were entropy and TF-IDF. The Naïve Bayes classifier were used and compared with SVM. Each of the classifiers are paired with each term weighting schemes. Their results are evaluated by looking at their accuracy, precision and F score. The schemes were proposed to improve to cyberbullying classification.

REFERENCES

[1] Bretschneider, U., Wöhner, T., and Peters, R. (2014). Detecting Online Harassment in Social Networks [online]. Available from http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1003&context=icis2014 [Accessed 10th March 2019].

[2] C. Chelmis, D. Zois and M. Yao. (2017). Mining Patterns of Cyberbullying on Twitter. 2017 *IEEE International Conference on Data Mining Workshops (ICDMW), New Orleans, LA*, 126-133.

[3] Cataldo Musto, Giovanni Semeraro, Marco Polignano. 2014. A Comparison of Lexicon-based Approaches for Sentiment Analysis of Microblog Posts, Department of Computer Science, University of Bari Aldo Moro, Italy.

[4] Chavan, V. S. and Shylaja, S. S. (2015). Machine Learning Approach for Detection of Cyber-aggressive Comments by Peers on Social Media Network. *International Conference on Advances in Computing, Communications and Informatics (ICACCI). Ker-ala. August 10-13, 2015. IEEE*, 2354-2358.

[5] Chen, Y., Zhou, Y., Zhu, S., and Xu, H. (2012). Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. *International Conference on Privacy, Security, Risk and Trust (PASSAT) and Social Computing (SocialCom). Amsterdam, September 3-5, 2012*. New York: IEEE.

[6] Cohen, W. (1995). Fast Effective Rule Induction. *Twelfth International Conference on Machine Learning,* 115-123.

[7] Cybersecurity Malaysia. (2018). MyCERT Incident Statistics. [Online]. Available from https://www.mycert.org.my/statistics/2018.php [Accessed 13th March 2019].

[8] Dadvar, M. and De Jong, F. (2012). Cyberbullying Detection: A Step Toward a Safer Internet Yard. *International Conference Companion on World Wide Web. 21st. Lyon, April 16-20, 2012.* London: ACM.

[9] Dadvar, M., Trieschnigg, D., and De Jong, F. (2014). Experts and Machines against Bullies: A Hybrid Approach to Detect Cyber-bullies. *Advances in Artificial Intelligence,* 8436, 275-281.

[10] Dadvar, M., Trieschnigg, D., Ordelman, R. and de Jong, F. (2013). Improving Cyberbullying Detection with User Context. *European Conference on Information Retrieval. 35th. Moscow. March 24th–7th, 2013.* Springer Berlin Heidelberg, 693-696.

[11] Dinakar, K., Reichart, R. and Lieberman, H. (2011). Modeling the Detection of Textual Cyberbullying. The Social Mobile Web [online] Available from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/download/3841Karthik/4384 [Accessed 10th March 2019].

[12] Fahrnberger, G., Nayak, D., Martha, V. S. and Ramaswamy, S. (2014). SafeChat: A Tool to Shield Children's Communication from Explicit Messages. *International Conference on Innovations for Community Services (I4CS). 14th. Reims. June 4 -6, 2014*. New York: IEEE, 80-86.

[13] Glossary of Cyberbullying Terms. (2008, January). Retrieved from http://www.adl.org/education/curriculum_connections/cyberbullying/glossary.pdf.

[14] I. H. Witten and E. Frank. (2005). *Data Mining: Practical Machine Learning Tools and Techniques.* Second Edition. San Francisco, CA: Morgan Kauffman,

[15] Kontostathis, A., Reynolds, K., Garron, A. and Edwards, L. (2013). Detecting Cyberbullying: Query Terms and

Techniques. *Annual ACM Web Science Conference. 5th. Indiana. June 23–26, 2013*. New York: ACM, 195-204.

[16] K. Reynolds, A. Kontostathis and L. Edwards. (2011). Using Machine Learning to Detect Cyberbullying. *2011 10th International Conference on Machine Learning and Applications and Workshops, Honolulu, HI*, 241-244.

[17] Nahar, V., Unankard, S., Li, X. and Pang, C. (2012). Sentiment Analysis for Effective Detection of Cyber Bullying. *Web Technologies and Applications*, 767-774.

[18] Nahar, V., Li, X. and Pang, C. (2013). An Effective Approach for Cyberbullying Detection. *Communications in Information Science and Management Engineering*, 3(5), 238.

[19] NaliniPriya. G and Asswini. M. (2015). A Dynamic Cognitive System for Automatic Detection and Prevention of Cyber-bullying Atacks. *ARPN Journal of Engineering and Applied Science,* 10(10), 4618-4626.

[20] Nandhini, B. S. and Sheeba, J. I. (2015a). Online Social Network Bullying Detection Using Intelligence Techniques. *Procedia Computer Science,* 45, 485-492.

[21] Patchin, J., & Hinduja, S. (2006). Bullies Move Beyond the Schoolyard; A Preliminary Look at Cyberbullying. *Youth Violence and Juvenile Justice,* 4(2), 148-169.

[22] Patchin, J. W. and Hinduja, S. (2012*). Preventing and Responding to Cyberbullying: Expert Perspectives*. Thousand Oaks: Routledge.

[23] Rafiq, R.I., Hosseinmardi, H., Han, R., Lv, Q., Mishra, S. and Mattson, S. A. (2015). Careful What You Share in Six Seconds: Detecting Cyberbullying Instances in Vine. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Paris. August 25-28, 2015. ACM*, 617-622

[24] Rajaraman, A., Ullman, J. D. (2011). *Data Mining: Mining of Massive Datasets,* 1-17.

[25] Serra, S. M. and Venter, H. S. (2011). Mobile Cyber-bullying: a Proposal for a Pre-emptive Approach to Risk Mitigation by Employing Digital Forensic Readiness. *Information Security South Africa (ISSA),* 1-5.

[26] Yin, D., Xue, Z., Hong, L., Davison, B. D., Kontostathis, A. and Edwards, L. (2009). Detection of Harassment on Web 2.0. Content Analysis in the WEB. Madrid. April 21, 2009.

[27] Vivek K. Singh, Souvick Ghosh, and Christin Jose. (2017). Toward Multimodal Cyberbullying Detection. *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17*). ACM, New York, NY, USA, 2090-2099.

[28] V. S. Chavan and Shylaja S. S. (2015). Machine Learning Approach for Detection of Cyberaggressive Comments by Peers on Social Media Network. *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Kochi,* 2354-2358.

[29] Willard, N. E. (2007). *Cyberbullying and Cyberthreats: Responding to the Challenge of Online Social Aggression, Threats, and Distress*. Champaign, IL: Research.

[30] Xu, J. M., Zhu, X. and Bellmore, A. (2012). Fast Learning for Sentiment Analysis on Bullying. *International Workshop on Issues of Sentiment Discovery and Opinion Mining*. 1st. Beijing. August 12–16, 2012. New York: ACM.

[31] Y. Chen, Y. Zhou, S. Zhu, and H. Xu. (2012). Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust, ser. SOCIALCOM-PASSAT '12*. Washington, DC, USA, 71-80.

[32] Yin, D., Xue, Z., Hong, L., Davison, B. D., Kontostathis, A. and Edwards, L. (2009). Detection of Harassment on Web 2.0, in Content Analysis in the WEB. Madrid. April 21, 2009.

[33] Z. Hailong, G. Wenyan and J. Bo. (2014). Machine Learning and Lexicon Based Methods for Sentiment Classification: A Survey. *2014 11th Web Information System and Application Conference, Tianjin*, 262-265.

[34] Zhao, R., Zhou, A. and Mao, K. (2016). Automatic Detection of Cyberbullying on Social Networks Based on Bullying Features. *International Conference on Distributed Computing and Net-working. 17th. Singapore*. January 04-07, 2016. ACM, 43.

[35] Zois, D., Kapodistria, A., Yao, M., & Chelmis, C. (2018). Optimal Online Cyberbullying Detection. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP*), 2017-2021.