Detecting Cyberbullying activities Over Social Media

John Hani, Mohamed Nashaat, Mostafa Ahmed, Zeyad Emad
Supervised by Dr. Eslam Amer and Eng. Menna Gamil

September 25, 2018

Abstract

Cyberbullying, defined by Cambridge dictionary, as the use of technology to harm or offend other people by sending them obnoxious messages. Without a doubt, cyberbullying is one of the most critical topics nowadays, because it has a huge impact on our society. Cyberbullying causes big problems to the society and must be prevented as fast as possible. Lately, there has been some serious work done to help in the detection and prevention of cyberbullying. Although, it39;s not easy to detect it on social media. There are many challenges to detect cyberbullying automatically like: how to make the AI understand the meaning of a post or How to deal with anonymous users. In order to detect offensive comments, you have to use some preprocessing to remove any noise from the text. Like removing special characters and stop words and to return obfuscate words to its original form. Our aim in this project is to build an application that Detects cyberbullying on social media by using machine learning classifiers and take an appropriate action against the bullies.

1 Introduction

1.1 Background

Social media has been increased largely in the Middle East in the last decade. According to social media in the Middle East in 2017 the number of the active social media users are 93 million people a day. As we know the social media is a good place for communication, sharing information and maintaining the old relationships. On the other hand, it has many bad impacts on the society especially the teenagers. One of the biggest bad impacts is bullying. According to very well family .com cyberbullying can lead to many psychological, physical and mental effects like feeling lonely, depression, anxiety, and the dangerous thing is that bullying can lead to suicide. So due to the prevalence of cyberbullying according to bullyingstatistics.org over half of the youth have been cyberbullied

and equal to this number have been involved in cyberbullying .our aim is to detect cyberbullying in the existence of sarcasm using machine learning classifiers and hybrid classifiers.

1.2 Market Motivation

In the last couple of years, the usage of the internet and social media has increased drastically and the usage will continue to grow over time, with this increase, the amount of cyberbullying will be huge in the near future. There are many non-profit organizations that call to stop cyberbullying like UNICEF is now making a great campaign to eliminate bullying in Egypt. As we mentioned before, the community is suffering from cyberbullying and some action has to be done to prevent it.

1.3 Academic motivation

Our work is motivated by the previous work by the field. Rui Zheo, et al developed enhanced bag of words [10]:. Michele Di Capua, etal in their paper Unsupervised Cyber Bullying Detection in Social Networks [7]: that they want to detect sarcasm.also the large number of false positive in "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network" paper.

1.4 Problem Definitions

Data in the internet nowadays are too huge to be monitored manually by humans to detect cyberbullying. In previous cyberbullying detection frameworks there has been a problem in detecting false positive cyberbullyings cases. So we aim to Enhance the accuracy of cyberbullying detection using machine learning classifier and contextual analysis. Also In cyberbullying detection frameworks they cant detect sarcasm so we aim to detect sarcasm along with cyberbullying.

2 Project Description

2.1 Objective

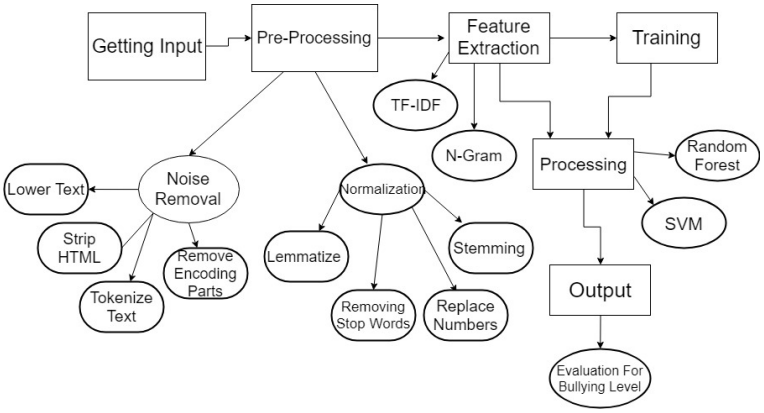
Our goal is to build an application that39;s able to detect and prevent cyberbullying using sentiment and contextual analysis and machine learning classification. First, we want to make the system able to understand the posts in social media to detect emotions in the posts and decide whether there is cyberbullying or not. Moreover, the system should be able to detect sarcastic posts and cyberbully posts by understanding the context. Finally, when a cyberbullying is detected a warning appears to the parents of the bullied child or to the administrator of the social media platforms to take the appropriate action. The system will only cost the cloud server hosting and a laptop to develop on.

2.2 Scope

The system will cover in its scope:

1. Sentiment analysis of text combined with contextual analysis.
2. The system will work on word level analysis and also phrase level analysis.
3. The system will use machine learning classifiers and try hybrid approach to increase the accuracy.

2.3 Project Overview



3 Similar System Information

1. Sentiment Informed Cyberbullying Detection in Social Media [6]:
   (a) In this paper the researchers were motivated by psychological and sociological findings, wanted to investigate the relationship between sentiment information and Cyberbullying behaviors.
   (b) The main problem is to used sentiment analysis to detect cyberbullying and deal with 2 problems: short, noisy and unstructured content information and the obfuscation of the obnoxious words by the users.
   (c) Researchers proposed a principle learning framework called (SICD) and they study whether sentiment information is particularly correlated with cyberbullying behaviors and how to deal with short and unstructured content.
   (d) Researchers conducted extensive experiments on two real-world datasets. The experimental results show the effectiveness of the proposed model as well as the impact of sentiment information.
   (e) This Paper is going to help us in the sentiment analysis section as they done many experiments that investigate the effectiveness of the sentiment analysis on cyberbullying.

2. Automatic Detection of Cyberbullying on Social Networksbased on Bullying Features[10]:

   (a) They made this program because the increasing of social media which increase the cyber bullying that give bad impacts on children and teenagers such as depression and suicidal thoughts.

   (b) The main problem that BOW is every word is independent from the other that fail to see the sentence as all.

   (c) They made a framework that detect the cyber bullying, based on word embeddings they made a list of insulting words then they assign to them weights. After this they concatenate latent semantic feature with bag of words then they classified them with SVM.

   (d)

Measures	BoW	sBoW	LSA	LDA	EBoW
Precision	75.6	75.7	75.9	74.0	**76.8**
Recall	77.8	78.3	78.2	76.5	**79.4**
F1 Score	76.6	76.9	77.0	74.9	**78.0**

   (e) It is important to us because they concatenate bag of words with latent semantic feature.

3. Cybercrime detection in online communications: The experimental Case of cyberbullying detection in the Twitter network [1]:

   (a) The bad effects of social media like cyberbullying that make the cyberbullied person suffering from many things such as suicidal thoughts and depression.

   (b) They dont have word embeddings or sentiment analysis they rely their work on classification.

   (c) Their model takes network, tweet content, activity and user features from tweets then they train random forest with SMOTE classifiers to classify cyberbullying and non-cyberbullying.

   (d) Results: under the receiver operating characteristic (ROC) curve (AUC) of 0.943 fmeasure of 0.936 using random forest with SMOTE.

   (e) It is important to us because they use hybrid classifiers which one of them is random forest and we plan to use these methods.

4. Unsupervised Cyber Bullying Detection in Social Networks [7]:

   (a) While cyber bullying is a well-studied problem from a social point of view, only recently it has attracted the attention of computer scientists, especially towards automatic detection tasks. For this reason, only relatively few articles on the subject and very few datasets are available.

   (b) We proposed to adopt an unsupervised approach to detect cyber bully traces over social networks.

(c)

TABLE I
RESULTS OBTAINED ON FORMSPRING.ME DATASET

Precision	Accuracy	Recall	F1	Method
0.72	0.73	0.69	0.71	GHSOM
0.60	-	0.40	-	C4.5
-	-	0.67	-	SVM

TABLE II
AVERAGE RESULTS OBTAINED ON YOUTUBE DATASET.

Precision	Accuracy	Recall	F1	Method
0.60	0.69	0.94	0.74	GHSOM

TABLE III
AVERAGE RESULTS OBTAINED ON TWITTER DATASET.

Precision	Accuracy	Recall	F1	Method
0.81	0.72	0.26	0.4	GHSOM
-	0.67	-	-	Naive Bayes

(d) We now know multiple sources that we can setup as our data sets (YouTube, twitter, FormSpring)

5. Experts and Machines against Bullies: A Hybrid Approach to Detect Cyberbullies [5]:

   (a) Most of the technical studies have focused on the detection of cyberbullying through identifying harassing comments rather than preventing the incidents by detecting the bullies.

   (b) Proposed methods: we introduce the three types of models used for calculating and assigning the bulliness score to the social network users: a multi-criteria evaluation system, a set of machine learning models and two hybrid models that combine the two.

   (c) Machine Learning Approaches: We used three well-known machine learning methods, which use pre-labelled training data for automatic learning: a Naive Bayes classifier, a classifier based on decision trees and Support Vector Machines (SVM) with a linear kernel

   (d) Results: The discrimination capacity of the MCES was 0.72.

6. Cyberbullying System Detection and Analysis [9]:

   (a) Cyber-bullying has recently been reported as one that causes tremendous damage to society and economy.

   (b) The system relies on the detection of three basic natural language components corresponding to Insults, Swears and Second Person.

(c) Proposed Methods: the whole is greater than the sum of its parts. A combination of modestly accurate features coming from heterogeneous data modalities can outperform methods that employ a single modality.

(d)

Feature	Acc.	Prec.	Reca.	F1-mea	F2-mea
tf-Idf	97,3%	31,2%	68,4%	42,85%	55.23%
LIWC	76,4%	28,4%	57,1%	32,56%	41,97%
Depen	67,5%	27,3%	60,6%	37,64%	48,72%
tf-Idf+LIWC	97,8%	42,4%	75,1%	54,20%	65,01%
LIWC + Depen	82,1%	38,4%	69,5%	49,47%	59,81%
tf-Idf+Depen	97,9%	58,9%	78,4%	67,26%	73,53%
All features	**99,4%**	**69,0%**	**84,9%**	**76,13%**	**81,15%**

(e) This work opens up new direction for future research through using advanced parser, dimension reduction and taking into account users profile in order to strengthen the detection capabilities.

7. Common Sense Reasoning for Detection, Prevention, and Mitigation of Cyberbullying [8]:

(a) Cyberbullying or harassment on social networks is as much a threat to the viability of online social networks for youth today as spam once was to email in the early days of the internet.

(b) Proposed models: To detect explicit bullying language pertaining to (1) sexuality, (2) race and culture and (3) intelligence. Binary classifiers outperform their multiclass counterparts: JRip and Support Vector Machines were the best performing in terms of accuracy and kappa values.

(c)

	Naive Bayes			Rule-based JRip			Tree-based J48			SVM (poly-2 kernel)		
	Acc.	F1	kappa	Acc.	F1	kappa	Acc.	F1	Kappa	Acc.	F1	kappa
Sexuality	66%	0.67	0.657	**80%**	0.76	0.598	63%	0.57	0.573	66%	0.77	**0.79**
Race and Culture	66%	0.52	0.789	**68%**	0.55	0.789	63%	0.48	0.657	66%	0.63	**0.71**
Intelligence	72%	0.46	0.467	**70%**	0.51	0.512	70%	0.51	0.568	72%	0.58	**0.72**
Mixture	63%	0.57	0.445	63%	0.60	0.507	61%	0.58	0.456	66%	0.63	0.653

(d) Future work: They are currently embarking on the use of a family of latent variable models to model, understand and predict self-harm in adolescents, a phenomenon that is not very well understood in the field of abnormal psychology.

8. Improved Cyberbullying Detection Using Gender Information [4]:

(a) We used a supervised learning approach to detect cyberbullying. We constructed a Support Vector Machine classifier using WEKA.

(b) Four types of features: Profane words, second person pronouns, other personal pronouns, and the weight of the words in each sentence.

Table 1. The accuracy measures for basic and gender-based approaches for cyberbullying detection in a MySpace corpus

Feature used in classifier	Precision	Recall	F-measure
Baseline	0.31	0.15	0.20
Gender-specific	0.43*	0.16*	0.23*
Female-specific (34% corpus)	0.40	0.05	0.08
Male-specific (66% corpus)	0.44	0.21	0.28

(c)

(d) Future work: Considering contextual features of the text as well as the word level features. The ground truth annotation can be done through crowdsourcing, investigate other features which may differentiate the writing styles of the users such as age, profession, and educational level.

9. Machine Learning Approach for Detection of Cyber-Aggressive Comments by Peers on Social Media Network [3]:

   (a) There is an enormous amount of information to manually flag offensive comments or posts. So an automatic classifier that is fast and effective is needed to solve this problem.

   (b) There are challenges such as the comments containing special character and they used machine learning techniques to detect insults and offensiveness of the comments.

   (c) They proposed 2 new hypotheses for detecting cyberbullying and it has increased the precision by 4 %.

   (d) The achieved 70 % precision using SVM classifier and 64 % precision using logistic regression.

10. A Pattern-Based Approach for Sarcasm Detection on Twitter [2]:

    (a) Sarcasm is a sophisticated form of irony widely used in social networks and micro-blogging websites. It is usually used to convey implicit information within the message a person transmits.

    (b) Recognizing sarcastic statements can be very useful to improve automatic sentiment analysis of data.

    (c) They used NPL and SVM and for features extraction:sentiment-related features, punctuation-related features, syntactic AND semantic features and pattern-related features.

    (d) Their proposed approach reaches an accuracy of 83.1% with a precision equal to 91.1%.

3.1 Similar System Description

Cyberbullying is gender specific which means that each gender has their own set of preferences when it comes to terminologies, and each gender-specific language features into account improve the discrimination capacity of classifiers. Females tend to use a relational style of aggression such as cutting a person out of a group, while males use more threatening expressions and profane words. These words, including their abbreviations and acronyms, are treated as a single text and compared in relation to the whole post. Their methodology however also includes the usage of personal pronouns and second person pronoun and last but not least the TF-IDF.

3.2 Comparison with Proposed Project

	Previous System	**Our System**
Accuracy	70.29%	Definitely higher
Methodology	1. N-gram 2. Counting 3. TF-IDF 4. Pronoun occurrence 5. Skip-grams 6. Classification a. SVM b. Logistic regression	1. Sentiment and contextual features analysis 2. Syntactic features 3. Semantic features 4. Sentiment features 5. Linguistic Inquiry and Word Count 6. TF-IDF 7. N-gram 8. Noise removal 9. Normalization 10.Classification a. SVM b. Random Forest 11.Hybrid classifiers 12.Deep learning*
Application	No Application	A graphical user interface will be used for furthermore illustration
Dataset	Small scale of dataset	Large scale of dataset

* If the results are not satisfying we will use deep learning.

4 Project Management and Deliverable

4.1 Tasks and time plan

Task Name	Start Time	Finish
Idea Discussion	1/8/2018	1/8/2018
Idea Research	1/8/2018	13/9/2018
Proposal Writing	13/9/2018	16/9/2018
Implementing Prototype	16/9/2018	17/9/2018
Delivering Rehearsal	18/9/2018	18/9/2018
Delivering Proposal	18/9/2018	26/9/2018
Doing Survey	10/10/2018	20/10/2018
Implementing Second Prototype	20/10/2018	25/10/2018
Writing SRS	25/10/2018	30/10/2018
Implementing	30/10/2018	25/11/2018
Preparing For External Examiner	25/11/2018	3/12/2018
Implementing	3/12/2018	18/1/2019
Writing SDD	18/1/2019	1/2/2019
Implementing	1/2/2019	1/4/2019
Preparing For Implementation Evaluation	1/4/2019	25/4/2019
Writing 8 Pages Paper	25/4/2019	28/4/2019
Finalizing Implementation	28/4/2019	7/5/2019
Writing Final Thesis	10/5/2019	25/5/2019
Presenting Final Thesis	25/6/2019	25/6/2019

4.2 Budget and Resources Costs

- Cloud server 28$/month.
- Laptop 8000 EGP.

5 References

References

[1] M. A. Al-garadi, K. D. Varathan, and S. D. Ravana, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the twitter network," *Computers in Human Behavior*, vol. 63, pp. 433–443, 2016.

[2] M. Bouazizi and T. O. Ohtsuki, "A pattern-based approach for sarcasm detection on twitter," *IEEE Access*, vol. 4, pp. 5477–5488, 2016.

[3] V. S. Chavan and S. Shylaja, "Machine learning approach for detection of cyber-aggressive comments by peers on social media network," in *Advances*

in computing, communications and informatics (ICACCI), 2015 International Conference on. IEEE, 2015, pp. 2354–2358.

[4] M. Dadvar, d. F. Jong, R. Ordelman, and D. Trieschnigg, "Improved cyberbullying detection using gender information," in *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012).* University of Ghent, 2012.

[5] M. Dadvar, D. Trieschnigg, and F. de Jong, "Experts and machines against bullies: A hybrid approach to detect cyberbullies," in *Canadian Conference on Artificial Intelligence.* Springer, 2014, pp. 275–281.

[6] H. Dani, J. Li, and H. Liu, "Sentiment informed cyberbullying detection in social media," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases.* Springer, 2017, pp. 52–67.

[7] M. Di Capua, E. Di Nardo, and A. Petrosino, "Unsupervised cyber bullying detection in social networks," in *Pattern Recognition (ICPR), 2016 23rd International Conference on.* IEEE, 2016, pp. 432–437.

[8] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, "Common sense reasoning for detection, prevention, and mitigation of cyberbullying," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 2, no. 3, p. 18, 2012.

[9] Y. J. Foong and M. Oussalah, "Cyberbullying system detection and analysis," in *2017 European Intelligence and Security Informatics Conference (EISIC).* IEEE, 2017, pp. 40–46.

[10] R. Zhao, A. Zhou, and K. Mao, "Automatic detection of cyberbullying on social networks based on bullying features," in *Proceedings of the 17th international conference on distributed computing and networking.* ACM, 2016, p. 43.