Synopsis on topic

Detecting Cyberbullying on social media using Machine Learning

SUBMITTED BY

Harsh Agarwal

Jagruti Jadhav

Komal Nagar

Babita Jaybhaye

Group\_id: 1

UNDER THE GUIDANCE OF

Prof. Shrikant Dhamdhere

B.E.(COMPUTER ENGINNERING)

2021(Semester 1)

Department Of Computer Engineering

Genba Sopanrao Moze Trusts Parvatibai Genba Moze College Of Engineering, Wagholi

Technical ACM keywords:

Computing methodologies

* Artificial intelligence
* Natural language processing
* Lexical semantics

Applied computing

* Law, social and behavioral sciences
* Psychology

Problem Statement:

Cyberbullying is a critical global issue that affects both individual victims and societies. Many attempts have been introduced in the literature to intervene in, prevent, or mitigate cyberbullying; however, because these attempts rely on the victims’ interactions, they are not practical. Therefore, detection of cyberbullying without the involvement of the victims is necessary. In this problem we have to classify the statement weather if user is victim of cyberbullying or not.

ABSTRACT

With the exponential increase of social media users, cyber bullying has been emerged as a form of bullying through electronic messages. Social networks provide a rich environment for bullies to uses these networks as vulnerable to attacks against victims. Given the consequences of cyber bullying on victims, it is necessary to find suitable actions to detect and prevent it. Recently, deep neural network-based models have shown significant improvement over traditional models in detecting cyberbullying. Also, new and more complex deep learning architectures are being developed which are proving to be useful in various NLP tasks. The model is trained and evaluated on dataset that is provided by Dataturks. The dataset contained 16000 tweets gathered manually annotated by human experts. Selected Twitter-based features namely text and network-based features were used. Several classifiers are trained for determining cyberbullying

Keywords: Machine Learning, Cyberbullying, detection, embeddings, Social media, Deep learning

Goals and objectives

Implement cyberbullying detection system using given dataset

To study impact of various standard ml algorithms along with different data processing techniques in improving accuracy

Introduction:

Cyber bullying is the use of technology as a medium to bully someone. Social networking sites provide a fertile medium for bullies, and teens and young adults who use these sites are vulnerable to attack. Through machine learning, we can detect language, patterns used by bullies and their victims and develop rules to automatically detect cyber bullying content. Cyberbullying is harassing, threatening, embarrassing someone or making targeting sharing’s about that person through technology. Cyberbullying actions, which are more common among young children and young people, can also be observed in adults. In such cases, severe legal sanctions are imposed on adults, such as prison sentences.

In contrast to the typical “bullying” actions, there is no need for physical force or face-to-face communication for cyberbullying. Anyone using any device with an Internet connection can perform a cyberbullying action. Bullies can be anonymous, as well as from close friends of children and young people. Cyberbullying is the most common in these platforms:

• Popular social networks like Facebook, Instagram, Twitter and Snapchat

• Text messages sent via direct devices (SMS)

• Instant messaging features offered by e-mail providers, applications or social networks

• Chat rooms and e-mails

Although symptoms of cyberbullying vary, children and adolescents who are victims of bullying usually have the following symptoms:

• Emotional anger after using the Internet or mobile devices

• Overprotective behavior about digital life

• Getting away from family members, friends and general routine activities

• Avoiding school and group meetings

• Performance decrease in class and academic success

• To exhibit angry and irritated behavior at home

• Continuous changes in mood, behavior, sleep and appetite

• Extraordinary; stop using devices such as computers and telephones

• Tense and hasty when an instant message or email arrives

• Avoiding discussions about the use of computers and telephone

Despite the various prevention and intervention strategies, cyberbullying perpetration has not decreased in the last one decade. Thus, a complete solution is required for this problem. Cyberbullying needs to stop. The problem can be tackled by detecting and preventing it by using a machine learning approach, this needs to be done using a different perspective. The main purpose of our paper is to develop an ML model so it can detect and prevent social media bullying, so nobody will have to suffer from it. The proposed technique is implemented on the social media bullying dataset which was collected from various sources like Kaggle, GitHub, etc.

Recent studies have looked into automatically detecting cyberbullying incidents, for instance, an affect analysis based on a lexicon and Support Vector Machine (SVM) was found to be effective in detecting cyberbullying, however the accuracy decreased when the size of the data increased, suggesting that SVM may not be ideal in dealing with frequent language ambiguities typical for cyberbullying automatically collected data from an in-game chat (World of Tanks) and found cyberbullying to be a learned behavior (i.e. new players are less likely to engage in cyberbullying).

Methodology

The cyberbullying detection model encompassing the various features. The overall model has three main stages, namely, Preprocessing, feature extractions and cyberbullying detection and classification.

Preprocessing:

The Natural Language Toolkit (NLTK) is used for the preprocessing of data. NLTK is used for tokenization of text patterns, to remove stop words from the text, etc.

• Tokenization: In tokenization, the input text is split as the separated words and words are appended to the list. Firstly, PunktSentence Tokenizer is used to tokenized text into the sentences. Then 4 different tokenizers are used to tokenize the sentences into the words:

o Whitespace Tokenizer

o WordPunct Tokenizer

o TreebankWord Tokenizer

o PunctWord Tokenizer

• Lowering Text: It lowers all the letters of the words from the tokenization list. Example: Before lowering “Hey There” after lowering “hey there”.

• Removing Stop words: This is the most important part of the preprocessing. Stop words are useless words in the data. Stop words can be get rid of very easily using NLTK. In this stage stop words like \t, https, \u, are removed from the text.

• Wordnet lemmatizer: Wordnet lemmatizer finds the synonyms of a word, meaning and many more and links them to the one word.

Feature Extraction:

In this step, the proposed model has transformed the data in a suitable form which is passed to the machine learning algorithms. Features of the data are extracted and put them in a list of features. Also, the polarity (i.e., the text is Bullying or Non-Bullying) of each text is extracted and stored in the list of features.

Algorithm Selection:

To detect social media bullying automatically, supervised Binary classification machine learning algorithms like SVM with linear kernel Logistic regression, Naive Bayes is used. The reason behind this is SVM, Logistic Regression and Naive Bayes calculate the probabilities for each class (i.e., probabilities of Bullying and Non-Bullying tweets). Both SVM, Logistic Regression and NB algorithms are used for the classification of the two-cluster. Machine learning models were evaluated on the same dataset. Classification report is also evaluated. The accuracy, recall, f-score, and precision are also calculated

Software Requirements:

Python3.9, Tableau, anaconda, Vscode

Functional Requirements:

1. For detecting bullying content we need a proper data set which is implemented in real time.

2. A proper algorithm for analyzing the data set.

3. The model should give a respective output and it should not deviate from the expected output

Nonfunctional requirements:

1. Model should have accuracy between 60-90%
2. Data privacy should be maintained
3. Algorithm should not take more than 10 mins for execution

Hardware requirements:

I5 intel processor, 8gb ram, windows OS, basic NVidia graphics card.

Innovativeness and usefulness:

This project helps to prevent attackers to target young teens through social media by detecting real time detection of cyberbullying attackers.

Literature review

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sr. no.: | Authors | Methodology used | Dataset | Limitation |
| 1 | Yin et al. 2009 | Supervised machine learning | Kongregate, Slashdot, Myspace | More features can explore |
| 2 | Dinakar et al. 2011 | Supervised machine learning | YouTube | Victims and attackers not identified |
| 3 | Bayzick et al. 2011 | Rule based extractor | Myspace | Supervised or unsupervised method can be used |
| 4 | Reynolds et al. 2011 | Supervised machine learning | Form spring | Larger dataset should be used |
| 5 | Nahar et al. 2014 | Semi supervised learning | Kongregate, Slashdot, Myspace | Not focused on indirect cyberbullying |
| 6 | Mangaon kar et al. 2015 | Different ml algorithms | Twitter | Less features are used |
| 7 | Zhao et al. 2017 | Deep learning model Semantic Enhanced Marginalized Denoising Auto Encoder | Twitter, myspace | Small dataset size |
| 8 | Rafiq et al. 2018 | Dynamic priority scheduler | vine | Small dataset size |
| 9 | Niharika et al. 2020 | Elaborated on characteristics, forms of bullying, estimates of bullying, risk factors, side effects of bullying in India specific to adolescents going to school | Different social media platform | Lack of proper dataset |
| 10 | Manuel et al. 2021 | Supervised machine learning | Vine | lack of dataset. evaluation based on time, instead of number of posts |

Planner