Detecting Hate Speech and Insults on Social Commentary using NLP and Machine Learning

Hitesh Kumar Sharma, T.P. Singh, K Kshitiz, Harsimran Singh, Prince Kukreja
University of Petroleum & Energy Studies, Dehardun
Uttarakhand, India

Abstract:
In the era of social media and networking, the usage of bad words and aggressive words has increased significantly. The young population is playing a major role in it. Cyberbullying affects more than half of the young population using social media. Insults in social media websites create negative interactions within the network. These remarks build up a culture of disrespect in cyberspace. Tools and technologies geared to understand and mitigate it are scarce and mostly inactive. Also, current implementations on insult detection using machine learning and natural language processing have very low recall rates. In short, the paper involves determining ways to identify bullying in text by analyzing and experimenting with different methods to find the most suitable way of classifying bullying comments. We proposed a efficient algorithm to identify the bullying text and aggressive comments and analyses these comments to check the validity. NLP and Machine learning is used for analyzing the social comment and identified the aggressive effect of an individual or a group. The best performing classifier acts as the core component in a final prototype system that can detect cyberbullying on social media.

Keywords: Cyberbullying, Natural Language Processing, Machine Learning.

INTRODUCTION
Usage of unwanted words or cyberbullying has been increased on cyber platforms by the young community. It is increasing day by day due to actively participation on social media. Users are increasing in an exponential manner and their time devotion is also increasing in the same manner. They are free to say anything on these platform without any rules and regulation. It affects the accessibility of other people for these cyber or social platform. Analyzing these comments which contain cyberbullying words may help to limiting these content on these platform so that a healthy discussion can take place for which these community platform is developed. The abundance of public discussion spaces on the Internet has in many ways changed how we communicate with others. Whether it is a comments section for a controversial news article or a forum for discussing a particular video game, these online spaces allow us to easily share our own opinions and findings, as well as hear about the thoughts of others. Though these discussions can often be productive, the relative anonymity that comes with hiding behind a username has allowed people to post insulting or inappropriate comments. These posts can often create a hostile or uncomfortable environment for other users, one that may even discourage them from visiting the site. This problem is a serious one that website owners commonly face. In earlier year cyberbullying was not taken seriously and it was suggest to screen off or disconnect if you get some kind og bullying comments. But now the scenario is totally changed. In 2017, Half population on social platform is using these comment and these can not be ignored. Major Social platform like, Twitter and Facebook also facing this major problem. Many legal cases has been filed due to these abused word used against and individual and a community.

The ambition of this research work is to explore the possibilities of developing a prototype that can automatically detect cyberbullying on social media. This work extends current research on cyberbullying and online harassment detection.

PROBLEM STATEMENT
With the proliferation of the Internet, cyber security is becoming an important concern. While web provides easy, interactive, anytime and anywhere
access to the online communities, it also provides an avenue for cybercrimes like cyberbullying and online harassment. Cyberbullying is defined as an aggressive, intentional act carried out by an individual or group, that takes place in cyber space through various mediums including social commentary, online chats, text messages and e-mails. It is a huge problem on social media websites like Facebook and Twitter. A number of life threatening cyberbullying experiences among young people have been reported internationally thus drawing attention to its negative impact. In the USA, the problem of cyberbullying has become increasingly evident and has officially been identified as a social threat.

The challenges in fighting cyberbullying include: detecting online bullying when it occurs; reporting it to law enforcement agencies; and identifying predators and their victims. No present online community or social media websites (for example, Facebook and Twitter; where cyberbullying is most common), incorporates a system to automatically and intelligently identify aggression and instances of online harassment on its platform. Due to non-seriousness of this major issue earlier it is not considered the issue of research but now it is in dangerous phase. No-one can ignore this effect on cyber platform. It require a serious attention by researchers and cyber crime agencies to control this activity.

LITERATURE REVIEW
Since the research field of online harassment and cyberbullying is still emerging, there is only a limited amount of work available. Over the past few years, several techniques have been proposed to measure and detect offensive or abusive content/behavior on platform like Instagram [1], YouTube [2], 4Chan [3], Yahoo Finance [4], and Yahoo Answers [5]. Chen et al. [11] use both textual and structural features (for example, ratio of imperative sentences, adjective and adverbs as offensive words) to predict a user’s aptitude in producing offensive content in YouTube comments, while Djuric et al. [4] rely on word embeddings to distinguish abusive comments on Yahoo Finance. Nobara et al. [6] perform hate speech detection on Yahoo Finance and news data, using supervised learning classification. Kayes et al. [5] find that users tend to flag abusive content posted on Yahoo Answers in overwhelmingly correct way. Dinakar et al. [7] identified the aggressive and cyberbullying words from YouTube videos and decompose them for better classification. They collect these words mainly from controversial videos. Hee et al. [8] did their study on ask.fm web application. They found cyberbullying from audio content and do its study for aggressiveness. Hosseinmardi et al. [1] did their research work on images. They did it on Instagram and find the abusing content in Images and aggressiveness in representation in visual content.

The current technologies like part of speech, URLs BoW (Bag of Words), lexical features are useful for our study on this context.

Sentiment analysis plays an important role to find the abusement and aggressiveness in list of comments. It help to categories the comments in good or bad categories. In this study we made two main categories bullies and non-bullies and the use of probabilistic sentiment analysis approach is used for filtering in these two categories.

More recently, research into applying deep learning to related fields such as sentiment analysis has proven quite fruitful. Recurrent Neural Networks have been known to perform well in sentiment analysis tasks since RNNs use a sequencing model. Wang et al. [11] used LSTMs to predict the polarity of tweets and performed comparably to the state-of-the-art algorithms of the time. Huang et al. [12] found that hierarchical LSTMs allow rich context modelling, which enabled them to do much better at sentiment classification. Specifically, they chose to use LSTMs because it solves the vanishing gradient problem. Other researches have used Convolutional Neural Networks in sentiment analysis.

OBJECTIVES:
The objective of the research work is to combat online harassment and aggression by developing a prototype that can automatically detect cyberbullying and abusive behavior on social media and online communities by:
1. Extracting, collecting, and labelling the data set.
2. Preprocessing, cleaning, and experiment with various features to improve accuracy.
3. Classification of text, comment, or posts into one of the many classes.
4. Evaluation and analysis of best model.

The motivation for the work is to learn the application and implementation of Natural Language Processing and Machine Learning in a real-world problem, i.e., cyberbullying and online harassment.

**METHODOLOGY:**

We propose a novel methodology employing Natural Language Processing and Machine Learning to analyze texts and predicting abusive behavior. The pipeline involves extraction of a suitable data set from various online sources, preprocessing, ground truth building, feature engineering and selection, classification. Being a supervised learning problem, the goal is to classify a text from an online user which could be in the form of comments, and status/post updates into two categories – “Bully” & “Non-Bully”.

The first step towards detection of hate speech and cyber bullying is to get raw data sets from various online sources. Data sets for cyber bullying usually consists of user comments, posts, images, and videos on social networking websites. Twitter provides free access to 1% of its tweets to open source developers through its Streaming API which could be used in case of absence of a relevant data set. After data collection and extraction, the second step involves preprocessing or cleaning of the data set – noise reduction, lowercasing, tokenization, stemming, lemmatization, stop words removal, discarding URLs and punctuations, normalization, removal of spam content and handling missing values. If the gathered data set is not labelled/classified, each sample will be labelled or categorized into “Bully” & “Non-Bully”. “Bully” represents an aggressive/harassing text/comment/post update with possible signs of cyber bullying and “Non-Bully” otherwise. Depending upon the data set, there could be multiple labels, for example, “Bully”, “Aggressor”, “Spammer”, and “None”, making it a multi-classification problem. A learning algorithm learns better when a large amount of data is provided to it and it learn even better when additional information about the data is fed. In order to increase the accuracy of model, the next step would be feature engineering – extracting user, textual, and network features. Some possible features could be – Lexical Syntactic Features, TF-IDF (Term Frequency – Inverse Document Frequency), count of offensive words in a sentence, count of positive words in a sentence, count of second person pronoun in a sentence, Character 4-gram and 5-gram count, Word/Document Vectors. The final step is to perform classification using the (extracted) features and the ground truth. Naturally, different machine learning techniques can be used for this task, including probabilistic classifiers (for example, Naïve Bayes), Decision Trees, Ensembles (for example, Random Forest) or Artificial Neural Network. For evaluation and analysis of the result, various accuracy metrics – model accuracy score, test accuracy score, cross-validation score, precision, recall, sensitivity, specificity, AUC score will be noted. Based on the evaluation, a prototype of automatic detection and flagging of comments on a dummy social community would be created to demonstrate the final data product.

![Data Product Pipeline](image)

**Data Extraction and Collection:**

As discussed earlier, the first step towards detection of cyber bullying in the form of hate speech and insults is to get raw data sets. Data sets for cyber bullying usually consists of user comments, posts, images, and videos on social networking and social media. There are multiple sources to obtain vast amount of data sets – UCI Machine Learning Repository which houses thousands of open source data sets for data analysis purpose, Kaggle wherein individuals and businesses contributes data for research and competition, etc. Of the multiple data sets available, MySpace group data crawled in 2011, Formspring.me data crawled in 2010 and...
Instances of cyberbullying and hate speech are ever increasing on the popular video streaming website YouTube.com. The research work considers extracting comment threads from popular YouTube videos that are suspected to potentially ignite hate speech. The method uses a non-API approach due to recent instability of YouTube Data API v3. YouTube comments are parsed and extracted in delimited JSON format.

**Data set Description and Labelling:**
Among all the collected data sets, data set on cyberbullying detection contributed by Impermium on Kaggle is selected to build the ground truth. The data consists of two attribute fields and an identifier. The first attribute is the time at which the comment was made. There are multiple null instances which means an accurate timestamp is not possible. It is in the form “YYMMDDHHMMSS” followed by a Z character. It is on a 24-hour clock and corresponds to the local time at which the comment was originally made. The second attribute is the unicode-escaped text of the content, surrounded by double quotes. The content is mostly English language comments, with some occasional formatting. Total number of samples in the data set is 2235. There is a small amount of noise (less than 1%) in the data set as it is not meticulously cleaned. The gathered data set was manually labelled because it was not readily labelled. Data set labelling was the most time consuming and labor intensive. It For the purpose of detecting cyberbullying instances through hate speech and insults, each and every unit of textual data is carefully read, understood and classified. List of the...
possible classes – “Bully” and “Non-bully” which constitutes a binary classification problem; “Bully”, “Aggressor”, “Spammer” and “None” which constitutes a multi-class classification problem; “0” and “1” which refers to bully and non-bully comments respectively.

For the sake of simplicity, the selected data set is labelled into two classes – “0” meaning neutral comment and “1” meaning insulting comment. Neural comment can be considered as the one not belonging to the insulting class. Prediction must be a real number in the range [0,1] where 1 indicates 100% confident prediction that comment is an insult. Labelling the data set into two classes makes it a binary classification problem.

The data set is labelled on the basis of following guidelines:

1. Comments that are intended to be insulting to a person who is a part of the larger blog/forum conversation are labelled as insulting comments.
2. Insults directed to non-participants such as celebrities, public figures, etc. are labelled as neutral comments.
3. Comments which contain profanity, racial slurs, or other offensive language are labelled as insulting comments.
4. Comments which contain profanity or racial slurs, but are not necessarily insulting to another person are considered not insulting.
5. The insulting nature of the comment are obvious, and not subtle.

<table>
<thead>
<tr>
<th>id</th>
<th>Insult</th>
<th>Date</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>20120603103520</td>
<td>&quot;like this if you are a tribe fan&quot;</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>20120531215447</td>
<td>&quot;you're idiot....................&quot;</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>20120823164228</td>
<td>&quot;I am a woman Babs, and the only war on women...&quot;</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>20120828010752</td>
<td>&quot;WOW &amp; YOU BENEFITTED SO MANY WINS THIS YEAR F...&quot;</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>20120602223825</td>
<td>&quot;ahaha green me red you now loser whos winning...&quot;</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>20120603202442</td>
<td>&quot;InMe and God both hate-faggots. InWha's the...&quot;</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>20120603103042</td>
<td>&quot;Oh go kiss the ass of a goat...and you DUMMY...&quot;</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>20120602239022</td>
<td>&quot;Not a chance Kid, you're wrong.&quot;</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>20120528064125</td>
<td>&quot;On Some real Shit FYou LIVE JASMII!!&quot;</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>20120603071243</td>
<td>&quot;ok but where the hell was it released? you all...&quot;</td>
</tr>
</tbody>
</table>

Figure 3. Manually Labelled Data set

Data Preprocessing and Cleaning:

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing.

Figure 4. Data Cleaning Process
RESULTS & CONCLUSION

Real-time extraction of tweets using Twitter Streaming API. The tweets are filtered on the basis of keywords, in this case “Modi”, “NTPC”, and “Yogi”. The extracted tweets are formatted into a JSON file. Although the stream also give access to a number of meta-data, a query to extract only the name of username and the tweet will exclude every other data.

YouTube comments are extracted without using an API key through parsing the HTML and CSS data of a queried YouTube video using its video ID. All the comments including replies in the thread are extracted and saved in a text file in delimited JSON format.

REFERENCES


