International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 5, Issue 1
January 2018

Detecting Cyberbullying and Aggression in Social Commentary using NLP and Machine Learning

Kshitiz Sahay, Harsimran Singh Khaira, Prince Kukreja, Nishchay Shukla
University of Petroleum and Energy Studies, Dehradun, India

ABSTRACT

Online bullying and aggression against social media users have grown abruptly, causing serious consequences to victims of all demographics in recent years. It affects more than half of young social media users' worldwide, suffering from prolonged and/or coordinated digital harassment. Tools and technologies geared to understand and mitigate it are scarce and mostly inactive. Also, current implementations of insult detection using machine learning and natural language processing (NLP) have very low recall rates. The researchwork ascertains ways to identify and classify bullying in thetext by analyzing and experimenting with different. The work proposes a robust methodology for extracting text, user, and network-based attributes, studying the properties of bullies and aggressors, and what features distinguish them from regular users. State of the art NLP and machine learning algorithms are studied and evaluated for the task of identifying bullying comments in a dataset.

KEYWORDS

Cyberbullying, Social media aggression, Natural Language Processing (NLP), Machine Learning.

INTRODUCTION

Cyberbullying and cyber aggression are serious and widespread issues affecting increasingly more Internet users. It is defined as an aggressive, intentional act carried out by an individual or group, that takes place in cyberspace.In today's hyper-connected society, bullying, which was once limited to particular places or times of the day (e.g., school hours), can instead occur anytime, anywhere, with just a few clicks on mouse and taps on a keyboard. Cyberbullying and cyber aggression can take many forms and definitions, however, the former typically denotes repeated and hostile behavior performed by a group or an individual and the latter intentional harm delivered via electronic means to a person or a group of people who perceive such acts as offensive, derogatory, harmful, or unwanted. The abundance of public discussion spaces on the Internet has in many ways changed how we communicate with others. These discussions can often be productive, but the anonymity that comes with hiding behind a username has allowed users 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. In 2017, about 50% of young social media users reported being bullied online in various forms. Popular social media platforms like Twitter and Facebook are not immune, as racist and sexist attacks may even have caused potential buyers of Twitter to balk.

The ambition of this research work is to explore the possibilities of classifying hate speech, insults and harassment which are one of the various forms of cyberbullying in social media. The research work extends current research on cyberbullying and online harassment detection.

PROBLEM STATEMENT

With the proliferation of the Internet, cybersecurity 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 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 are most common), incorporates a system to automatically and intelligently identify aggression and instances of online harassment on its platform. Despite the seriousness of the problem, there are very few successful efforts to detect abusive behavior, both from the research community and social media itself, due to several inherent obstacles like grammar, syntactic flaws,and fairly limited context. Aggression and bullying against an individual can be performed in several ways beyond just obviously abusive language – for example, via constant sarcasm, trolling, etc.

OBJECTIVE

The objective of the research work is to combat online harassment and aggression by identifying instances of cyberbullying and abusive behavior on social media and online communities by:

1. Extracting, collecting, and labeling the dataset.
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 thebest model.

The motivation for the research 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.

LITERATURE REVIEW

Since the research field of online aggression 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 aplatform like Instagram [6], YouTube [22], 4Chan [5], Yahoo Finance [14], 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. [1] perform hate speech detection on Yahoo Finance and news data, using supervised learning classification. Kayes et al. [8] find that users tend to flag abusive content posted on Yahoo Answers in theoverwhelmingly correct way.

Dinakar et al. [11] detect cyberbullying by decomposing it into detection of sensitive topics. They collect YouTube comments from controversial videos, use manual annotation to characterize them, and perform a bag-of-words driven text classification. Hee et al. [2] study linguistic characteristics in cyberbullying-related content extracted from Ask.fm, aiming to detect fine-grained types of cyber-bullying, such as threats and insults. Besides the victim and harasser, they also identify bystander-defenders and bystander-assistants, who support, respectively, the victim or the harasser. Hosseinmardi et al. [6] study images posted on Instagram and their associated comments to detect and distinguish between cyber aggression and cyberbullying.

Previous work often used features such as punctuation, URLs, part-of-speech, n-grams, Bag of Words (BoW), as well as lexical features relying on dictionaries of offensive words, and user-based features such as user's membership duration activity, number of friends/followers, etc. Different supervised approaches have been used for detection: [6] uses a regression model, whereas [11, 8] rely on other methods like Naïve Bayes, Support Vector Machines (SVM), and Decision Trees (J48). By contrast, Hosseinmardi et al. [6] use a graph-based approach based on likes and comments to build bipartite graphs and identify negative behavior.

Sentiment analysis of text can also contribute useful features in detecting offensive or abusive content. For instance, Nahar et al. [20] use sentiment scores of data collected from Kongregate (online gaming site), Slashdot, and MySpace. They use a probabilistic sentiment analysis approach to distinguish between bullies

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 5, Issue 1
January 2018

and non-bullies and rank the most influential users based on a predator-victim graph built from exchanged messages. Xu et al. [10] rely on sentiment to identify victims on Twitter who pose ahigh risk to themselves or others. They consider specific emotions such as anger, embarrassment, and sadness.

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. [21] used LSTMs to predict the polarity of tweets and performed comparably to the state-of-the-art algorithms of the time. Huang et al. [7] found that hierarchical LSTMs allow rich context modeling, which enabled them to do much better at sentiment classification. Specifically, they chose to use LSTMs because it solves the vanishing gradient problem. Other researchers have used Convolutional Neural Networks in sentiment analysis.

APPROACH

Data Extraction and Collection

The first step towards cyberbullying detection is to gather raw datasets. Datasets for cyberbullying usually consists of user comments, posts, images, and videos on social media. There are multiple sources to obtain avast number of datasets – UCI Machine Learning Repository which houses thousands of open source datasets for data analysis purpose and Kaggle wherein individuals and businesses contributes data for research and competition. A different approach to gathering data is to extract it from a social networking website in real time.

Twitter constitutes a large ocean of data in the form of tweets and user information available for thepublic using Twitter Streaming API and Twitter REST API. The Streaming APIs give access to (usually a sample of) all tweets as they publish on Twitter. On average, about 6,000 tweets per second are posted on Twitter and developers get a small proportion (less than 1%) of it. The research work involves extracting real-time tweets on various topics and keywords over a course of several days and selecting suitable entries for the training dataset.

Instances of cyberbullying and hate speech are ever increasing on the popular video streaming website YouTube. The research work also considers extracting comment threads using an HTML/CSS parser from popular yet controversial YouTube videos that are suspected to potentially ignite hate speech. These are in delimited JSON format. Suitable entries are selected in a similar way for building ground truth.

Dataset Labelling and Description

A total of 6594 samples of data, 3947 from YouTube and 2647 from Twitter is carefully selected and used to build the final training dataset.

The dataset on cyberbullying detection contributed on Kaggle by Impermium is selected as the test dataset for validation of the model. 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. A total number of samples in the dataset is 2235. There is a small amount of noise (less than 1%) in the dataset as it is not meticulously cleaned.

The gathered dataset is manually labeled. For the purpose of detecting cyberbullying instances through hate speech and insults, each and every sample 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. The dataset is classified in “0” and “1”.

Data Preprocessing and Cleaning

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. It prepares raw data for further processing.

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 5, Issue 1
January 2018

After manually labeling the training dataset, it goes through a series of steps for preprocessing:

1. Data Cleaning: Data is cleansed through processes such filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data.
2. Data Integration: Data with different representations are put together and conflicts within the data are resolved.
3. Data Transformation: Data is normalized, aggregated and generalized.
4. Data Reduction: The step aims to present a reduced representation of the data.

The raw data is first loaded into the memory where it is cleansed of escape sequences like \n, \t and Unicode characters such as \xc2 with a white space. Colloquial words and phrases used mostly in text messages are replaced with its corresponding English word. For example, "u" is replaced with "you"; "em" is replaced with "them"; "da" is replaced with "the" and so on. Contractions such as "won't" and "can't" are replaced with "will not" and "cannot" respectively along with others. The data is further converted to lowercase format. Advanced natural language processing techniques are used to further preprocess the data to ensure a better quality and consistency of data format while building the ground truth and training a classifier.

To build a vocabulary of abusive words and internet slangs, a dictionary of bad words available at (http://urbanoalvarez.es/blog/2008/04/04/bad-words-list/) is used. The dictionary contains a list of bad words in a number of variations used on the internet and its corresponding English dictionary word. Through data preprocessing, different variations of internet slangs are replaced with its dictionary counterpart for which the bad words file is used.

The data is further stemmed down to its root form and occurrences of special characters is removed using regular expressions.

Feature Engineering and Selection

Feature engineering is the process of using domain knowledge of the data to create features that make machine learning algorithm work.

Two kinds of features are created, namely, count vector features and TF – IDF feature. As determined in the previous researches TF – IDF has outperformed other simpler features due to its ability to capture semantic information of the textual data. Therefore TF – IDF is used as a base feature set. Both count vectors and TF – IDF vectors are created by defining n-gram of up to five level. The feature vectors are generated using both word and character as a token. All the extracted features are of the form of asparse matrix.

Feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction.

To select the best features out of five feature sets, a feature selection technique known as SelectKBest is used. SelectKBest scores the features using a chi-squared test and removes all but the K highest scoring features. The result is feature sets with best-selected features to be fed into the machine learning algorithm for classification.

Table 1. Feature vectors of train and test dataset after feature selection.

	Train dataset	Test dataset
UNIGRAM, WHERE N IS 1	(6594, 2000)	(2235, 2000)
BIGRAM, WHERE N IS 2	(6594, 1000)	(2235, 1000)
TRIGRAM, WHERE N IS 3	(6594, 1000)	(6594, 1000)
4-GRAM, WHERE N IS 4	(6594, 500)	(6594, 500)
5-GRAM, WHERE N IS 5	(6594, 100)	(6594, 100)

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 5, Issue 1
January 2018

Table 2. Feature vectors of train and test dataset after feature selection.

	Train dataset	Test dataset
UNIGRAM, WHERE N IS 1	(6594, 2000)	(2235, 2000)
BIGRAM, WHERE N IS 2	(6594, 1000)	(2235, 1000)
TRIGRAM, WHERE N IS 3	(6594, 1000)	(6594, 1000)
4-GRAM, WHERE N IS 4	(6594, 500)	(6594, 500)
5-GRAM WHERE N IS 5	(6594, 100)	(6594, 100)

Building Machine Learning Model and Validation

All the extracted feature vector sets are stacked together into a single feature set consisting of count vectors and TF – IDF vectors of both words and characters as tokens with an n-gram sequencing of up to 5 level. The shape of the final feature set for training dataset is (6594, 4600) where 6594 is the number of samples and 4600 is the number of features or predictors. Similarly, test dataset consists of (2235, 4600) features.

For the purpose of building machine learning model, four kinds of classifiers are used – the most basic Logistic Regression, Support Vector Machine, and the two most popular ensemble methods, Random Forest Classifier and Gradient Boosting Machine. Logistic regression and support vector machine requires an input of sparse matrix feature set whereas random forest and gradient boosting machine requires dense matrix. Thus, the sparse matrix of feature vectors is appropriately converted to adense matrix. To improve the efficiency of the learning, various hyperparameters are studied and tuned. For instance, “C”, a parameter of logistic regression and support vector machine, is the inverse of regularization strength. Smaller values in SVM specify stronger regularization. Other parameters include “number of trees in the forest” in arandom forest, “learning rate”, “number of subsamples” in gradient boosting machine, etc.

Scikit-learn’s implementation of the above classifiers in Python is used to build the modelto classify the examples. The SVM is trained with a linear kernel on the training data. Since the training data is imbalanced, containing mostly negative examples, the cost-factor, a factor representing how much the cost of an error on a positive example should outweigh an error on a negative example. Tuning the parameters was done by hit and trial method where “C” assumed the values [0.002, 0.02, 0.003, 0.03, 300] ad “J”, the cost factor, assumed the values [10, 30, 100]. Training time varied greatly for all the four models. Logistic regression took a total training time of 0.060 seconds whereas SVM took 6.668 seconds. Both the ensemble methods took a much larger time, 196.236 seconds and 324.447 seconds for random forest and gradient boosting machine respectively.

After training the models, they were validatedusing a test dataset provided on Kaggle. Lastly, the predictions generated for test dataset is written down on a file. All the values of prediction lie between 0 and 1, where a score ranging from 0 to 0.5 denotes a “non-insulting” comment whereas a score between 0.5 to 1 denotes “insulting” comment.

RESULTS AND EVALUATION

The results of classifying the gathered training dataset and test dataset provided by Impermium are displayed in Table 2. The table shows various metrics of evaluation of the performance used after training the dataset and validating it with a test dataset. Training accuracy varied from 75% - 90% for all the four classifiers while the test accuracy lies between 70% - 75%.

Table 3. Accuracy score of train and test dataset.

Model	Train Accuracy	Test Accuracy	AUC Score	Cross Validation
Logistic Regression	0.900	0.737	0.777	0.720
SVM	0.966	0.793	0.778	0.759
Random Forest	0.905	0.745	0.739	0.747
Gradient Boost	0.974	0.792	0.779	0.753

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 5, Issue 1
January 2018

CONCLUSION

The state of the art in cyberbullying involves training a machine learning model using supervised learning. The research work mostly focuses on feature engineering, i.e., finding features that can separate bullying comments from non-bullying comments. Finding good features is difficult and problematic. Features that work well for YouTube comments might not work for Twitter comments, due to different social media platforms being likely to have varying vocabulary and expressions in part caused by restrictions on communication, different age groups,and user's interest.

It can be noted that the model performed well on a training dataset, generating a score between 77% - 90% but failed to generalize the test dataset. This is a case of over-fitting or alarge amount of variance in which the model tries its best to fit the training dataset but cannot classify the test dataset correctly. There are a number of possible reasons for such behavior:

1. Size of the dataset. Any machine learning algorithm performs well on a dataset containing a huge number of samples. Whereas the training dataset used for training the algorithm contains a very limited number of samples.
2. Differences in the dataset due to mixing social comments from two different social media platforms.
3. The dataset requires further data cleaning and preprocessing. Upon looking at the normalized dataset after data preprocessing step, it is found that although the preprocessing did a good job in normalizing the dataset, a lot of samples still remain inconsistent. A large number of abusive words and insults is missed out from the vocabulary because of its vastness of usage in many different forms. Apart from the abusive words, there were Unicode characters that remained in the preprocessed data. All these factors contributed greatly towards the poor performance of the model.
4. There were comments present in foreign languages like French and Spanish. The model only learned to classify English comments.
5. The requirement of different kinds of features, for example, Latent Dirichlet allocation (LDA), Latent Semantic Analysis (LSA), Predictive Word Embeddings like Word2Vec features and Doc2Vec features, etc.

The conclusion of the experiments and overall work is that out of the method that was have evaluated, support vector machine and gradient boosting machine trained on the feature stack performed better than logistic regression and random forest classifier in this particular case.

FIGURES

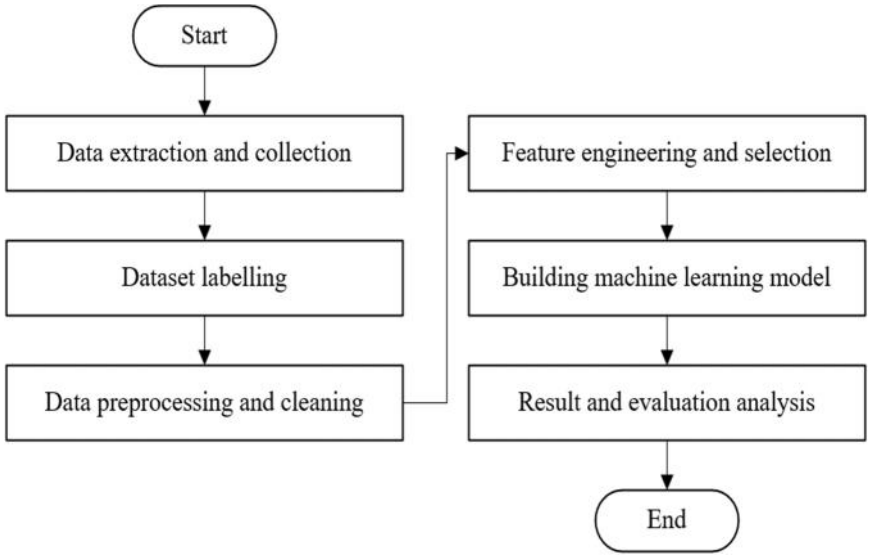

Figure 1. Model pipeline.

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 5, Issue 1
January 2018

	id	Insult	Date	Comment
0	1	0	20120603163526Z	"like this if you are a tribe fan"
1	2	1	20120531215447Z	"you're idiot....."
2	3	1	20120823164228Z	"I am a woman Babs, and the only "war on women...
3	4	1	20120826010752Z	"WOW & YOU BENEFITTED SO MANY WINS THIS YEAR F...
4	5	1	20120602223825Z	"haha green me red you now loser whos winning ...
5	6	0	20120603202442Z	"\nMe and God both hate-faggots.\n\nWhat's the...
6	7	1	20120603163604Z	"Oh go kiss the ass of a goat....and you DUMMY...
7	8	0	20120602223902Z	"Not a chance Kid, you're wrong."
8	9	0	20120528064125Z	"On Some real Shit FUck LIVE JASMIN!!!"
9	10	1	20120603071243Z	"ok but where the hell was it released?you all...

Figure 2. Manually labeled dataset.

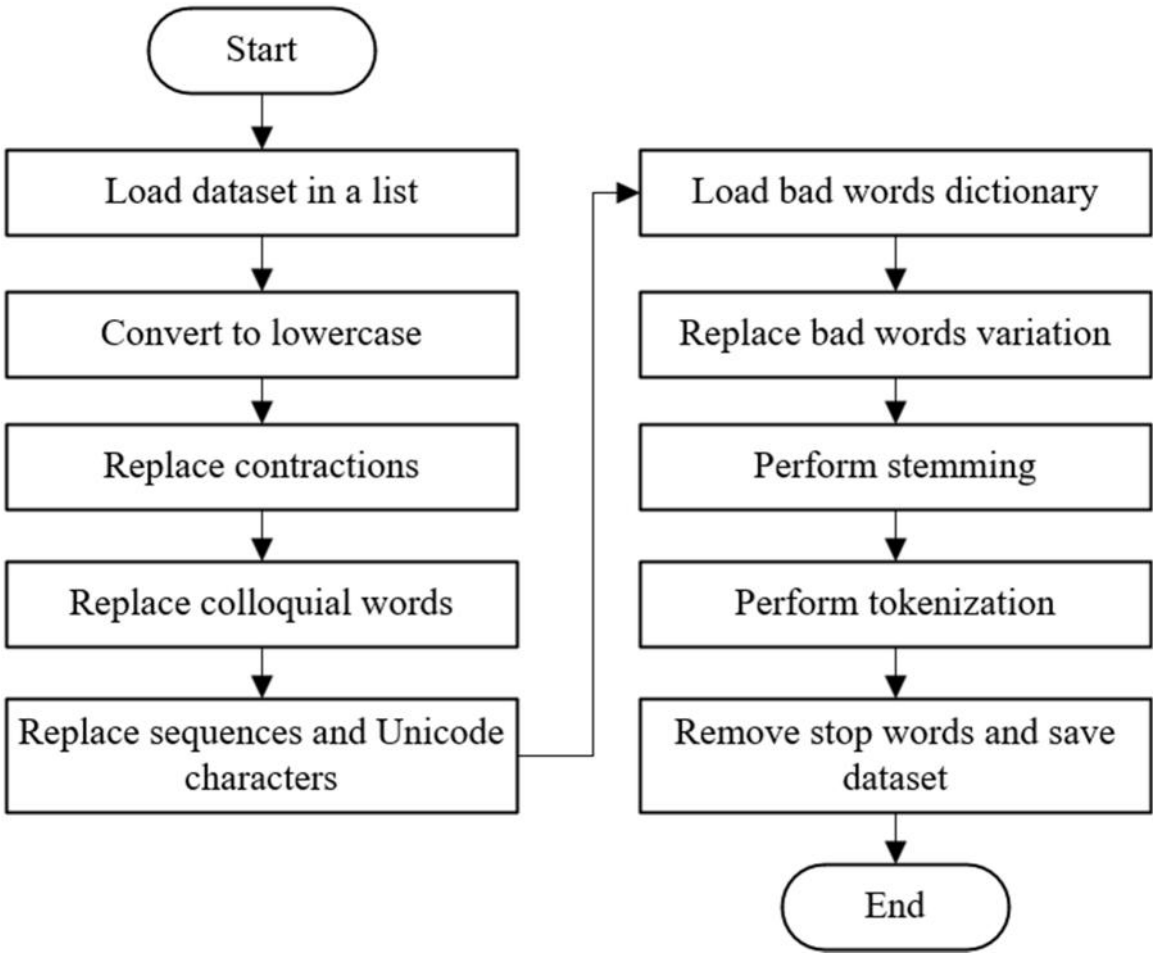

Figure 3. Data preprocessing steps

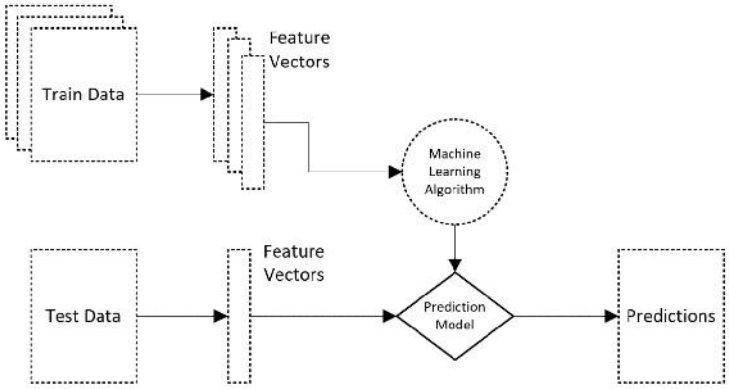

Figure 4. Building machine learning model.

REFERENCES

[1] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. Abusive Language Detection in Online User Content. In *WWW*, 2016.,6
[2] C. Van Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. De Pauw, W. Daelemans, and V. Hoste. Automatic Detection and Prevention of Cyberbullying. In *Human and Social Analytics*, 2015.,8
[3] D. Santos, C. Nogueira, and M. Gatti. Deep Convolutional Neural Network for Sentiment Analysis of Short Texts. *COLING*, 2014.,13
[4] Divyashree, H. Vinutha, N. S. Deepashree. An Effective Approach for Cyberbullying Detection and Avoidance. *International Journal of Innovative Research in Computer and Communication Engineering*, 2016.,14
[5] G. E. Hine, J. Onaolapo, E. De Cristofaro, N. Kourtellis, I. Leontiadis, R. Samaras, G. Stringhini, and J. Blackburn. A Measurement Study of 4Chan's Politically Incorrect Forum and its effort on the web. In *ICWSM*, 2017.,3
[6] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra. Analyzing Labelled Cyberbullying Incidents on the Instagram Social Network. In*SocInfo*, 2015,1
[7] Huang, Minlie, Y. Cao, and C. Dong. Modelling Rich Contexts for Sentiment Classification with LSTM. *arXiv preprint arXiv: 1605.01478*, 2016.,12
[8] I. Kayes, N. Kourtellis, D. Quercia, A. Iamnitchi, and F. Bonchi. The Social World of Content Abuser in Community Question Answering. In *WWW*, 2015.,5
[9] J. M. Xu, X. Zhu, A. Bellmore. Learning from Bullying Traces in Social Media. *University of Wisconsin-Madison*, 2016.,22
[10] J. M. Xu, X. Zhu, and A. Bellmore. Fast Learning for Sentiment Analysis on Bullying. In *WISDOM*, 2012.,10
[11] K. Dinakar, R. Reichart, and H. Lieberman. Modelling the Detection of Textual Cyberbullying. *The Social Mobile Web*, 11, 2011.,7
[12] K. Heh. Detection of Insults in Social Commentary. *Stanford University*, 2013.,19
[13] L. Engman. Automatic Detection of Cyberbullying on Social Media. *UMEA UNIVERSITY*, 2016.,15
[14] N. Djuric, J. Zhou, R. Morris, M. Gravois, V. Radosavljevic, and N. Bhamidipati. Hate Speech Detection with Comment Embeddings. In *WWW*, 2015.,4
[15] P. Ravi. Detecting Insults in Social Commentary. The *University of Illinois at Urbana Champaign*, 2016.,18
[16] R. K. Amplayo, J. Occidental. Multi-level Classifier for the Detection of Insults in Social Media. In *ResearchGate*, 2015.,20
[17] R. Sugandhi, A. Pande, A. Agrawal, H. Bhagat. Automatic Monitoring and Prevention of Cyberbullying. In *International Journal of Computer Applications*, 2016.,16
[18] T. Chu, K. Jue. Comment Abuse Classification with Deep Learning. *Stanford University*, 2017.,17
[19] U. Bretschneider, T. Wohner, R. Peters. Detecting Online Harassment in Social Networks. *Martin Luther University*, 2016.,21
[20] V. Nahar, S. Tankard, X. Li, and C. Pang. Sentiment Analysis for Effective Detection of Cyberbullying. In *APWeb, 2012.,9*
[21] Wang, Xin. Predicting Polarities of Tweets by Composing Word Embeddings with Long Short-Term Memory. *ACL*, 2015.,11
[22] Y. Chen, Y. Zhou, S. Zhu, and H. Xu. Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. In PASSAT and SocialCom, 2012.,2