Cyberbullying Intervention Interface Based on Convolutional Neural Networks

Qianjia Huang
Department of Computer Science
University of Ottawa
qhuan035@uottawa.ca

Jianhong Zhang
Department of Computer Science
University of Ottawa
jzhan410@uottawa.ca

Diana Inkpen
Department of Computer Science
University of Ottawa
Diana.Inkpen@uottawa.ca

David Van Bruwaene
SafeToNet Canada
dvanbruwaene@safetonet.com

Abstract

This paper describes the process of building a cyberbullying intervention interface driven by a machine-learning based text-classification service. We make two main contributions. First, we show that cyberbullying can be identified in real-time before it takes place, with available machine learning and natural language processing tools, in particular convolutional neural networks. Second, we present a mechanism that provides individuals with early feedback about how other people would feel about wording choices in their messages before they are sent out. This interface not only gives a chance for the user to revise the text, but also provides a system-level flagging/intervention in a situation related to cyberbullying.

1 Introduction

Cyberbullying, which can be defined as *'when the Internet, cell phones or other devices are used to send or post text or images intended to hurt or embarrass another person'* (Dinakar et al., 2012), has become a pernicious social problem in recent years. This is also worrying, as multiple studies found that cyberbullying victims often have psychiatric and psychosomatic disorders (Beckman et al., 2012), and a British study found that nearly half of suicides among young people were related to bullying (BBC News[1]). These factors underscore an urgent need to understand, detect, and ultimately reduce the prevalence of cyberbullying.

In contrast to traditional bullying (e.g., school bullying), cyberbullying is not limited to a time and place, which makes cyberbullying potentially more prevalent than traditional bullying. Cyberbullying victims may not recognize their experiences as bullying and they may not report them or seek help for associated emotional difficulties. Kowalski and Limber (2007) reported that almost 90% of young cyberbullying victims did not tell their parents or other trusted adults about their online negative experiences. These factors are especially worrying as multiple studies have reported that the victims of cyberbullying often deal with psychiatric and psychosomatic disorders (Beckman et al., 2012; Sourander et al., 2010), and the worst cases are suicides (Tokunaga, 2010).

Given the importance of the problem, content-based cyberbullying detection is becoming a key area of cyberbullying research. Current state-of-the-art methods for cyberbullying detection

[1]http://www.bbc.co.uk/news/10302550

Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, pages 42–51
Santa Fe, USA, August 25, 2018.

combine contextual and sentiment features (e.g., curse word dictionaries, histories of users activities, grammatical properties, and sentiment features derived from online users content) with text-mining approaches. While performance can be improved by training on text-external features, the scarcity of platform-ubiquitous external features requires a cross-platform new-media text classification algorithm to be trained strictly on text. This reduces the presence of features that are significant in training, but absent from data used in out-of-domain contexts. Hence, we introduce a text-driven model which covers six social media platforms (Facebook, Instagram, Twitter, Pinterest, Tumblr, Youtube), and it could be an ideal solution for this problem.

While presenting their methods for cyberbullying detection7, scholars have also suggested different interfaces for intervention. Dinakar et al. (2012) describe cyberbullying intervention mock-ups for both sender and receiver. With the aid of the text-driven model, this project also implements an Android-based interface which combines and optimizes the mock-ups from Dinakar et al. (2012). For instance, an interface giving the sender a chance to retype/cancel the message (as shown in Figure 1) is considered in our project. This project could be developed into a thrid-party application between users and social media providers for creating a healthy online environment.

Figure 1: Mock-up for delaying and undoing the issuance of messages from Dinakar et al. (Dinakar et al., 2012)

2 Online Cyberbullying Model

2.1 Data Set

This paper uses a dataset that was built for cyberbullying detection. Visr, a predictive wellness company, produced this dataset for use with an application (app) that analyzes online activities and interactions, and then alerts parents to potentially harmful issues their children may be experiencing.[2] Issues that parents are alerted about include bullying, anxiety, and depression. By making parents immediately aware of emerging issues on Instagram, Gmail, Tumblr, YouTube, Facebook, Twitter, and Pinterest, the Visr app aims to help parents address such issues before they grow into thornier problems. The app raises a red flag to warn parents when signs of these issues are detected in a child's online activities, including signs of possible mental health consequences like nascent depression, eating disorders, and self-harm. Visr accesses children's social media content through the API's of these social media channels with the consent of the children who are the account holders. This provides a unique cross-platform dataset with rich information.

[2]The app is available at https://app.visr.co, through the Apple App store, or through the Google Play store.

The data was collected by the Visr child safety app from September 2014 to March 2016. Over a half-million online posts were selected from among the social media platforms (Facebook, Instagram, Twitter, Pinterest, Tumblr, Youtube) and Gmail. These posts were randomly chosen among posts that had been viewed, received, or sent by the adolescents (between age 13 to 17). Personally identifying information was removed to ensure the privacy of Visr users. Demographic information such as gender, age, location's time-zone, post time, and the number of likes were recorded as well.

The specific cyberbullying detection dataset is one of Visr's labeled datasets. After combining an enriched selection of those posts with 3,072 'real issues' (posts with issues confirmed by parents), an annotation process was performed by three annotators for the *cyberbullying* label. 1,753 posts were determined to be positive for *cyberbullying*), 304 were labeled as 'unsure', and 12,441 were labeled as negative for *cyberbullying* (Agreement percentage: 95.07%; Cohen's Kappa: 0.805). The posts either labeled as 'unsure' or about which annotators disagreed were removed, leaving a corpus of 14,194.

2.2 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are known to have good performance on data with high locality, when words get more care weight about the features surrounding them. For our classification problem, we are trying to get high locality in text given their short length and their tendency to focus on cyberbullying.

We used CNNs that received input text in the form of sequences of integer representations of stemmed unigrams. Our character processing included the conversion of emoticons into word representations, and the removal of non-Latin characters. We also removed frequently occurring url components (e.g., names of popular websites), metadata encoded in the main body-text (e.g., 'RT: '), and a variety of social media platform-specific features. Hashtags and @-mentions were reduced to binary features. The text was then lower-cased and tokenized using NLTK's Tweet-Tokenizer[3]. The tokenized text was next encoded using a dictionary of integers, with the original ordering of the tokens preserved. The encoded text was converted into dense vectors of fixed size. This one-dimensional embedding was fed into a single-layer CNN with 200 embedding dimensions, 150 output dimensions, and 200 convolution kernels. The kernels were optimized using Tensorflow's 'adagrad' optimizer (lr=0.001) using categorical cross-entropy as the loss function. The 150 output dimensions were flattened using a sigmoid function into two output nodes whose values are floats between 0 and 1, with 1 representing *bullying* and 0 representing non-bullying.

To test the performance of our model, we took 70% of the dataset as training set, and 30% of it for testing. As suggested by previous research, we also added textual features (total used: 93) from LIWC 2015 [4] to build another model for comparison. We set the threshold which got the best result (here we used highest F-measure to represent the performance). Comparison can be seen in Table 1. We put ZeroR and SVM (Support Vector Machine) models as baselines for comparison. Because the ZeroR model puts everythoing in the majority class, labeling all of the positive instances as negative ones, both the F-measure for the positive class and True

[3]http://www.nltk.org/api/nltk.tokenize.html

[4]https://liwc.wpengine.com/

Positive score are 0. Meanwhile, the AUC valus of the ZeroR model is 0.5 and the accuracy measure depends on the distribution of positives and negatives in the dataset. It is obvious that the performance of CNN model is better than that of the SVM model in terms of F-measure, AUC, and True Positive rate. It is also expected that adding LIWC features could help to improve the F-measure and accuracy. However, for other important parameters (i.e., True Positives and AUC value), our original model with NLTK-tokenized features got a better index. Note that all the thresholds are taken as 'optimal' because they lead to the highest F-measure, which is not only influenced by True Positive rate but also by the True Negative rate. However, in real-life, we care more about the true positives than the true negatives; in other words, detecting the normal cases (which is much more frequent than cyberbullying) is not the goal for this project. Hence, we are setting up the thresholds with another method which will be described in section 2.3.

Model name	**F-measure**	AUC	**Accuracy**	**True Positive**
ZeroR	0	0.5	87.6%	0%
SVM	0.517	0.851	87.1%	58.9%
SVM + LIWC	0.585	0.892	90.2%	55.9%
CNN	0.523	0.860	87.1%	60.4%
CNN + LIWC	0.597	0.898	90.0%	60.2%

Table 1: Comparison of the results of the CNN models

2.3 Thresholds Setting

To set up the thresholds of our application, we built an electronic survey which contains 45 cyberbullying posts (being labeled as cyberbullying) from the VISR dataset and there are five posts which did not belong to cyberbullying at all (e.g., non-bullying–'he was a complete ass hole, .. He used 2 tell me that my mom tried to abort me because she didn't want to have another kid', cyberbullying–'Go fuck yourself!'). Then we invited four colleagues (two males and two females, all PhD students) to give feedbacks of their feelings about those cyberbullying-related online posts. The survey goes as follow:

"Assume someone (online, you might know the person or not) is sending you a message via phone/posting a message which @yourid/leaving a comment under your profile, etc. Please give the feedback score about how you feel:
1. It is totally fine;
2. Well, not that comfortable, but there is no need to hide it;
3. Not acceptable, I don't want to see this ever, it should be blocked."

After reviewing the feedback scores from our colleagues to ensure they understand the assumption properly, one survey was deleted as the participant didn't get the whole image and returned the feedback with 47/50 'score 3' ("I suggest to hide every F-word, that's really annoying" the participant wrote in the feedback survey). Thus, three surveys were kept for the threshold setting.

The results from the three participants' feelings and the related bullying index of the chosen text are shown in Figure 2. We averaged the 'feeling score' of the three participants; for instance, the average 'feeling score' 1.67 represents the total score of 5, which means that probably two chose 2 and one chose 1. The cyberbullying index of the selected texts is from 0.0431 to 0.8614. We separated the text with 'feeling score' 1 - 1.33 as 'totally fine' group, 1.67 - 2.33 as 'uncomfortable but not that bad' group, 2.34 - 3 as 'not acceptable' group. With an ANOVA test, we found the 'totally fine' group (mean = 0.17) is significantly different (p = 0.0101) from the other two groups ('uncomfortable but acceptable': 0.40, 'unacceptable': 0.44).

From the results, six messages are mainly considered as 'totally fine' ('feeling score' from 1 to 1.33, similar to the number of not bullying at all) and the average index is 0.17; we set this index as the threshold_1. From our testing document, most of the 'uncomfortable' and 'not acceptable' scores are higher than threshold_1; only three of them would be identified as 'totally fine'. For the threshold_2, we set it with the average index of 'uncomfortable but not that bad' group (0.40); seven of the 'not acceptable' messages get lower index (which would not be sent with warning, as discuss in the following section), but because of the threshold_1, only one of these 'not acceptable' would not be filtered as 'uncomfortable'.

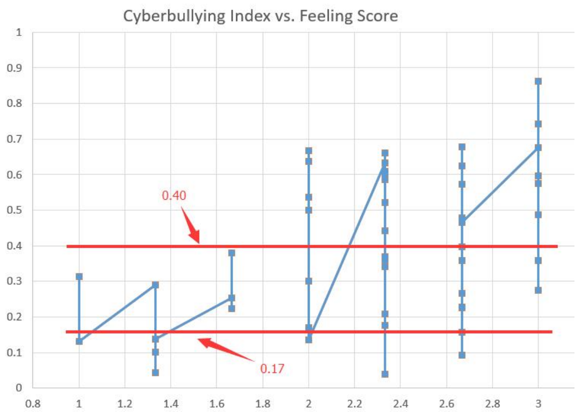

Figure 2: Threshold Setting

3 Application

3.1 Platform and Development Environment

We chose Android OS as our application development platform as it is the most popular mobile OS, and it is an open source, real time operating system which meets our development requirements. We used Android Studio for the development, since it is optimized for all devices, and it provides various APIs and layouts.

3.2 Application Design

3.2.1 Application Subsystem Design

The user interface has a text input field which allows the user to type the message and send it by using the HTTPS to Visr API over the network. This app will extract the cyberbullying index calculated by the Visr API and compare it with the set thresholds. Based on the results of the comparison, a corresponding prompts will be given to the user.

We used HTTPS to transmit data, and we used Volley module to implement the HTTP functions. Volley [5] is a HTTP library developed by Google, it provided us the: 1. Scheduling network request; 2. JSON and images asynchronous downloading; 3. Network request priority handling; 4. Caching; and 5. Powerful APIs.

3.2.2 System Architecture

The system dialog is shown in Figure 3. Regarding users' communication behavior and freedom of speech, the challenges to this interface are presented as follow. On one hand, to build a better online environment, we do not want to miss any of the 'uncomfortable'/'unacceptable' texts and allow them to be sent without filtering. On the other hand, we do not want the users to feel annoying about the intervention; if they really want to send the message, it is not possible to stop them. Hence, we built a win-win solution with the two thresholds and the related interventions.

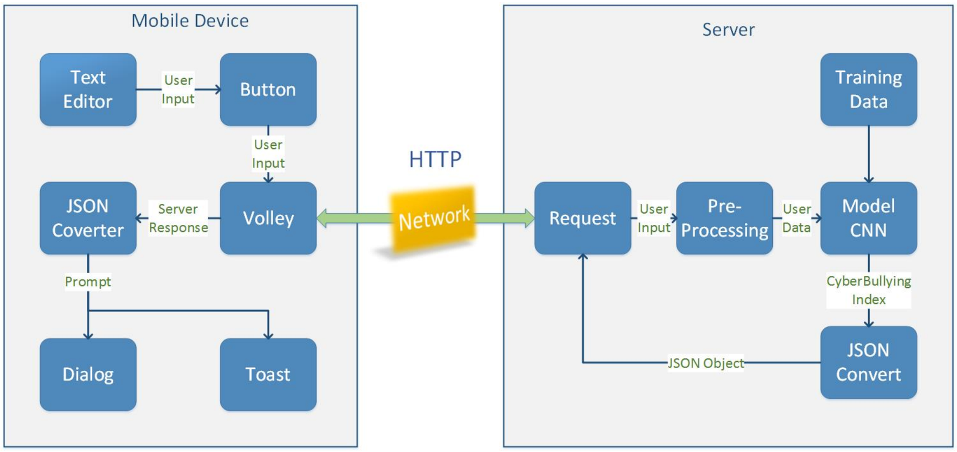

Figure 3: System Diagram

In our interface application, for a text which has index between two thresholds (0.17, 0.40), we send a prompt to the user ('Your message may be aggressive to others, do you really want to send?'). As in this level, the text is not identified as unacceptable; the user could either send it or change it. Similar to figure 1, the prompt is displayed as a delay/chance for user to think about the feelings of the receiver. For the text which has a extremely high index (>0.40) of cyberbullying, the app will first send a prompt as 'Your message may make others feel uncomfortable, please change the tone.' and stay on the same page for the sender to modify the message. If the user

[5] https://developer.android.com/training/volley/index.html

changes it (or not) and the index is still high (above 0.40), the app will send another prompt as 'Your message may make others feel uncomfortable, do you really want to send with warnings?'. This second step comes as we do not want to infringe on users' freedom of speech and we respect the users' communication behavior as well. The only idea is to give advice to the writer about how other people would feel when reading the message. The system flow chart is shown in Figure 4.

4 Examples of the interface

To understand the system better, here are examples of situations with different outputs from the Visr API.

Imagine a user typed a message which receives the cyberbullying index as 0.20 which belongs to the 'uncomfortable but acceptable' level; the prompt would pop us as 'Alert: Your message may be aggressive to others, do you really want to send?' If the user clicks 'yes', it will send the message immediately; if user clicks 'no', there would be a chance to change the message.

For the 'extremely high index' message, such as 'what's your opinion for this fucking shit? You are retard!', the system will display the prompt as shown in figure 5. Please note there is no way to submit this message at the first time. No matter whether the user changes the content or not, if the second time the submitted text's index is still higher than threshold_2, a prompt which is similar to 'uncomfortable but acceptable' would pop up. This prompt is shown in figure 6. If the user wants to send it anyways, there will be warnings with this message, interventions such as hiding the message or sending the message with a warning could be applied at the receiver's end.

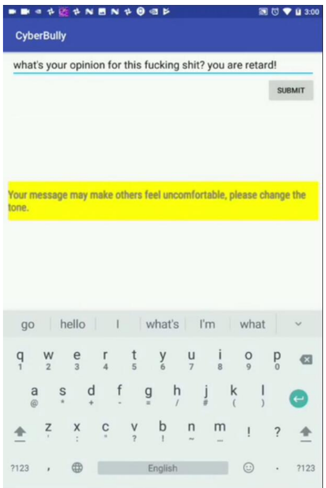

Figure 5: 'Extremely high index' message

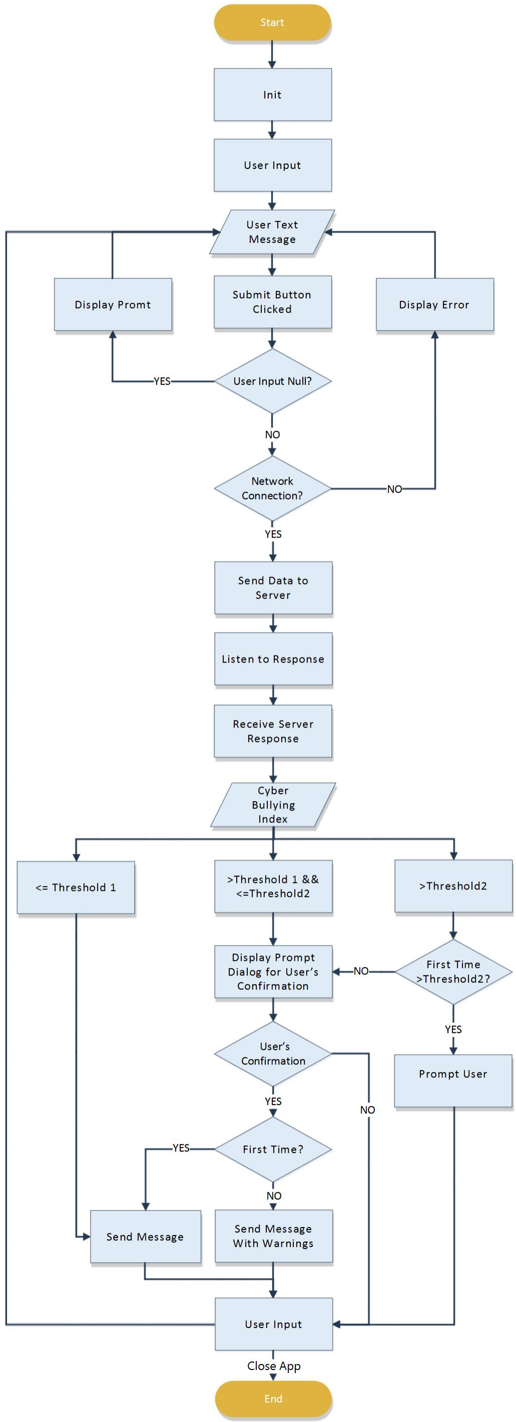

Figure 4: Flow Chart

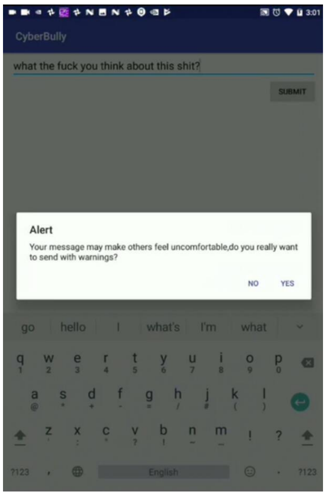

Figure 6: The second time 'extremely high' index of the message

5 Conclusion and Future Work

This paper described the process of creating a cyberbullying intervention application based on a convolutional neural network learning model trained on a multi-platform dataset. The model is then used to compute a cyberbullying index for any new message.

For setting the two thresholds based on the feedbacks of the participants, different interventions could be taken for different levels of the cyberbullying index. Finally, as discussed, this project could be seen as the first step towards a framework of building an effective cyberbullying intervention application for online applications. Social media platforms could use the textual cyberbullying index and our proposed thresholds in order to make interventions to safeguard users from cyberbullying.

Several possible optimizations for future work are as follow:

- Word embeddings, such as GloVe[6] or Word2Vec[7] could be utilized to initialize our CNN models, which might lead to better results.
- As sending images and videos is becoming popular among adolescents (Singh et al., 2017), image/video processing would be another important area for cyberbullying detection.
- The user interface designed in this project takes the user's input directly. In the future, it can be designed as a background running application, which can collect user's input from different applications (while respecting privacy and security issues).
- A complete social network relationship graph (Huang et al., 2014) (for example, whether this conversation is between two good friends or not) could be taken into consideration for improving the cyberbullying identification.

[6] https://nlp.stanford.edu/projects/glove/
[7] https://code.google.com/archive/p/word2vec/

References

Linda Beckman, Curt Hagquist, and Lisa Hellström. 2012. Does the association with psychosomatic health problems differ between cyberbullying and traditional bullying? *Emotional and behavioural difficulties*, 17(3-4):421–434.

Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):18.

Qianjia Huang, Vivek Kumar Singh, and Pradeep Kumar Atrey. 2014. Cyber bullying detection using social and textual analysis. In *Proceedings of the 3rd International Workshop on Socially-Aware Multimedia*, pages 3–6. ACM.

Robin M Kowalski and Susan P Limber. 2007. Electronic bullying among middle school students. *Journal of adolescent health*, 41(6):S22–S30.

Vivek K Singh, Marie L Radford, Qianjia Huang, and Susan Furrer. 2017. " they basically like destroyed the school one day": On newer app features and cyberbullying in schools. In *CSCW*, pages 1210–1216.

Andre Sourander, Anat Brunstein Klomek, Maria Ikonen, Jarna Lindroos, Terhi Luntamo, Merja Koskelainen, Terja Ristkari, and Hans Helenius. 2010. Psychosocial risk factors associated with cyberbullying among adolescents: A population-based study. *Archives of general psychiatry*, 67(7):720–728.

Robert S Tokunaga. 2010. Following you home from school: A critical review and synthesis of research on cyberbullying victimization. *Computers in human behavior*, 26(3):277–287.

DOI: http://dx.doi.org/10.26483/ijarcs.v9i1.5396
Volume 9, No. 1, January-February 2018
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
RESEARCH PAPER
Available Online at www.ijarcs.info

CYBERBULLYING REVELATION IN TWITTER DATA USING NAÏVE BAYES CLASSIFIER ALGORITHM

Vandana Nandakumar
MTech Student
Department of Information Technology, CUSAT
Kerala, India

Binsu C. Kovoor
Assistant Professor
Department of Information Technology, CUSAT
Kerala, India

Sreeja M. U.
Research Scholar
Department of Information Technology, CUSAT
Kerala, India

Abstract: Cyberbullying can be visualized as a potential issue affecting children and all categories of people. One demanding concern is effective representation for learning of content messages. The proposed system deals with cyberbullying revelation in email application using Naive Bayes Classifier Algorithm. The Classification Algorithm is a baseline method for content classification; the method of analyzing documents as relating to one classification or the other with word prevalence as features. The technique deals with the identification and filtering of spam words. The denoised messages are classified with the help of Naive Bayes Classifier Algorithm. The messages are processed under feature set extraction method. The feature probabilities are found out using Naive Bayes Classifier Algorithm .The efficiency factor is compared among the two algorithms, Naive Bayes Classifier Algorithm and Support Vector Machine and a graph is plotted. Comparison on the basis of precision factor is also done with the fact that the probabilities for each feature set are calculated independently from the twitter dataset and can evaluate the performance by predicting the output variable.

Keywords: e-mail; twitter; spam; Naïve Bayes classifier; support vector machine

I. INTRODUCTION

Cyberbullying can be explained as the technique of bullying a person or a character with the advent of internet technologies [1, 2]. Cyberbullying Revelation is a technique which is implemented in email application with the help of Naive Bayes Classifier Algorithm. The algorithm uses the classification method in order to categorize the messages which are having spam words. The proposed system categorizes the emails which are having cyberbullying content versus the emails which are not having cyberbullying content. The denoised value for each word is calculated by grouping the messages and is done by using the classification method. The feature set extraction technique is done for each twitter message which selects data attributes that best characterize a predicted variable [3]. The feature probabilities are calculated using the naïve Bayes classifier algorithm. The proposed system alerts the sender if they are using any vulgar languages and the messages are redirected as such. The technology provides hope to concerned parents and is a sign of relief to all categories of people who are affected.

Cyberbullying problem is also occurring in school premises. Teachers try to make their students aware about cyberbullying practices and its negative effects [4]. The proposed system uses Word Embedding Technique as its framework which obtains the bullying characters automatically. Finally these specific alterations make the new feature space more selective and thus facilitate Cyberbullying Revelation.

Machine Learning Techniques can make automatic revelation of bullying messages in social media networks possible and will create a clear social environment [5]. Data mining can be explained as an absolute subfield of computer science. It is the technique of analyzing patterns in large data sets including techniques at the intersection of machine learning methodologies and database system techniques. The main aim of data mining technology is to extract information from a large data set and convert it into useful methods so that it can be used for further extensions [6]. Cyberbullying can be compared with traditional bullying but the latter encompasses a range of public areas like college, school with the victim often experiencing it. The predator is the first person who is capable of molesting the victim in both cases. But, cyberbullying is done with the help of online methods where physical presence of victim is not a relevant factor.

In [7], a real time system has been implemented which minimizes the amount of bullying rather than detecting and preventing them. An analysis of common users in social networks is done in [8]. The posting activity of common users and relation with negativity is examined here. Negativity in anonymous messages is also analyzed. The accuracy of predicting the level of cyberbullying attack using classification methods is studied in [9]. A Facebook watchdog application is developed in [10] that make use of image analysis, social media analytics, and text mining techniques to detect cyberbullying activity.

The rest of the paper is organized as follows. Section 2 describes the methodology adopted for the proposed system. Experimental results based on different parameters are described in section 3. Conclusions along with future enhancements possible are detailed in section 4.

II. METHODOLOGY

The various modules for the proposed system are GUI designing, Training dataset, classification and analyzing the twitter messages for the presence of spam content. The classification technique is implemented using Naïve Bayes Classifier Algorithm. The Revelation consists of the following steps. The primary step is to accept data sets from numerous online network sites. The dataset deals with the twitter messages and the wordcount for each message are calculated by grouping the messages. In the layers, the values of words are found out using the wordcount. The datasets include the comments that are posted by users, images, video clips on networking sites, social networks network, etc. Using Twitter API, tweets can be easily analyzed and verified. The next stage is Preprocessing of Data where the dataset is processed so that data contains only required information. Subsequently, denoised words are analyzed. The removal of whitespaces and stop words can be considered as a way of data preprocessing after which tokenization and lemmatization occurs. Various other methods are also there to clear the datasets. The final stage deals with the classification of data. The classification is done using classifiers and a classification algorithm is considered as a part of it. Messages are classified into a set of classes. The probabilities are found out using the feature set extraction method. A message with a value less than the threshold value which is 1.5 is considered as a cyberbullied message. Data is classified into positive and negative instances by comparing text content having cyberbullying content with data which has no admissible cyberbullying content. After processing the messages the denoised values are found out. The message classification is done using Naïve Bayes Classifier Algorithm. The denoised values are trained using the Naïve Bayes Classifier. Before a new data is classified, the classification algorithm is in need of training sets for training a classifier and thus in turn facilitates potent and discriminative representation of learning of text messages. A classifier is in need of labeled examples. So that it could analyze the label of an input and this learned classifier is then used to validate a bullying message. Numerous algorithms and techniques are used for the classification of data like Bag-of-Words(BoW),Support Vector Machine(SVM),Naive Bayes Classifier Algorithm etc. Each technique deals with both the positive and negative versions of cyberbullying aspects.

Data Preprocessing improves the data set so that the dataset includes only required information. The various stages in data preprocessing are:

A. *Tokenization*

Tokenization is the process of distributing large set of unstructured messages into a small subset of tokens. These are classified with the help of various aspects such as white spaces, punctuation marks and is categorized as phrases, sentences etc.

B. *Stop words Removal*

The most common words that are used in a text are words such as 'a' 'and' 'are' and so on. The main drawback of these words are that such words only contribute very little meaning to text and aids only a very small value in classifying text. Stop words removal from messages results in more convenient recognition of text in further steps.

C. *Replacement of Special Characters*

The method deals with the replacement of special characters like '@' with its exact word 'at'. In tweets, this step has larger importance because of the extensive occurrences of special symbols.

D. *Stemming and Lemmatization*

This method finds the root of a single word and is considered as a heuristic technique it simply abridges prefixes as well as suffixes. It uses word-based approach in order and is so called dictionary based approach. Lemmatization method can be considered as the further extension of stemming technology. For the grammatical categorization of characters to get the base method of a single word called lemma, this technique is widely used. One of the algorithms which are broadly used for this purpose is Porters algorithm and can be more adequately used and is more specific.

E. *Coreference Resolution*

Coreference Resolution is the method of analyzing all expression that focuses to the same entity in text content. One of the relevant co-referential equipment in a written document which can be considered is repetition and it could make string-comparison characteristics more relevant to all co-reference resolution methods. This is one of the most promising steps in advanced Natural Language Processing methods that comprise minor language analysis such as a text content summarization, answering to a particular question and retrieval of the correct information.

The flowchart of the proposed system is shown in fig.1.

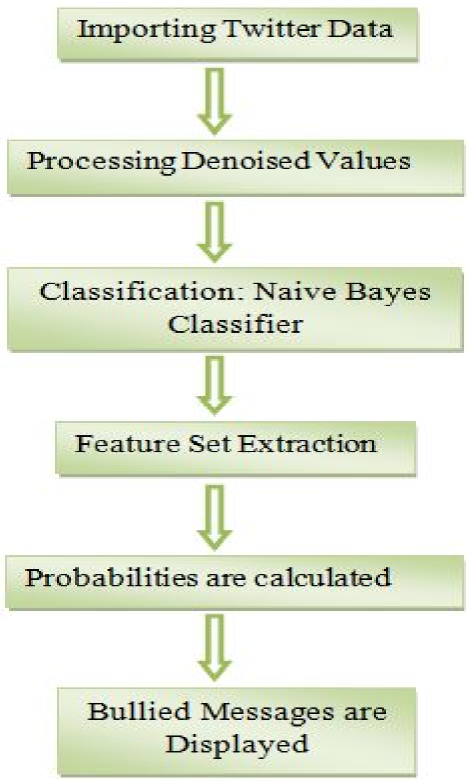

Figure 1. Flowchart of proposed system

The main limitation in the research related to cyberbullying is that the bullying begins when there is rivalry between the bullying person and victims in real life. Since real life incidents are difficult to be deducted from social networks, the main reason for bullying is difficult to be analyzed. But there are some researches depicting the common reasons to bully a person like love failure, envy, etc. The first step here is to identify the popular newsmakers for a week and term frequency methods can be used as an output. The next stage is to identify the news that made the newsmakers popular and fetch the corresponding news and cluster them. Later, extract the comments and posts related to that news using the similarity index. The last method is to identify the cyberbullying terms and negative words so it can be more extensively recognized. This can be considered as the challenging parts in cyberbullying revelation .A graph can be introduced after the classification of data by considering the data which definitely has cyberbullying content with the data which has no significant cyberbullying content.

III. Results and Discussion

The results and discussion deals with the analysis of comparison graph on the basis of precision and run time complexity. When the algorithm runs it takes each twitter messages and breaks it down into individual words. Each word is compared to the words in the bully dictionary. If it matches any of the words then it is added to the precision value. The precision values are compared with the threshold value and if it is less than the threshold value it is considered as cyberbullied message. Finally, the algorithm adds all the twitter messages having precision values less than the threshold. Fig. 2 compares precision values obtained using Naïve Bayes classifier and support vector machine.

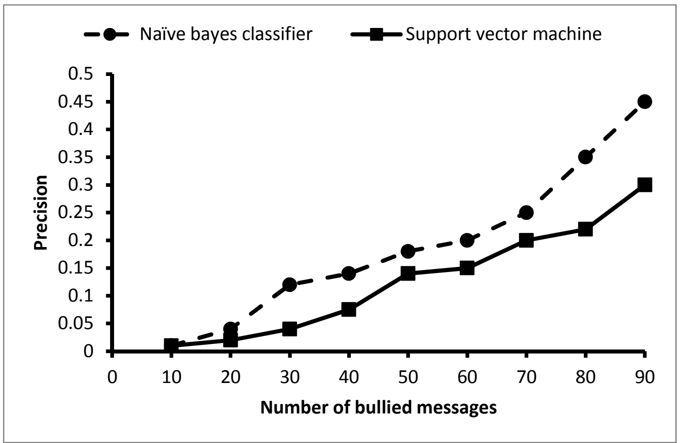

Figure 2. Comparison based on Precision

Using the precision factor, the probabilities for each feature set are calculated independently from the twitter dataset and performance is evaluated by predicting the output variable. The Feature extraction selects the data attributes that best characterize a predicted variable. It can be done more conveniently using the precision factor analysis of Naïve Bayes Classifier Algorithm. The identification and separation of segments can also be done more effectively using this technique.

The time complexity graph in fig.3 shows that Naïve Bayes Classifier is having lower run time complexity. Run time complexity is calculated as the absolute difference between the time before the algorithm starts and time at which the algorithm finishes running. It is calculated in milliseconds.

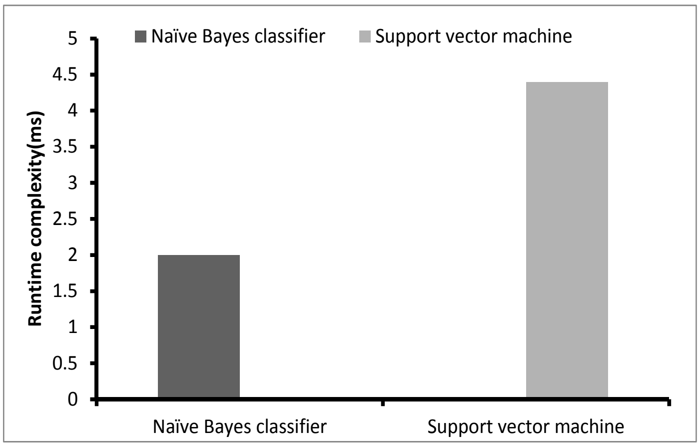

Figure 3. Comparison based on Run time complexity

IV. Conclusion and Future Scope

The rapid increase of social networks has shown a consistent growth in cyberbullying activities. Cyberbullying has become a major social problem. Cyberbullying has become an important area of research due to its impact on society. Various researches try to recognize the reason of cyberbullying and its aftereffect. But only a few try to enhance software to prohibit cyberbullying. Robust and selective representation of learning of text messages is crucial for consistent detection system [11]. Machine Learning representation and authentication makes automatic revelation of bullied messages in online media possible and ensures building a relevant and clear social media environment. The Email based cyber stalking is also a huge problem. Email based cyber stalking detection involves two phases; the first is to analyze and detect cyber stalking emails and the second phase is to verify the proof for finding out the cyber stalkers as a prohibition and detection mechanism [12].

Cyberbullying is a major problem that is happening on the Internet. Internet is a convenient environment for bullying the most vulnerable community. The main procedure for cyberbullying revelation is web based mining technologies. An acceptable level of precision can be acquired with the proposed system and the results are promising. For social network medias when the level of precision is met it can be copied and verified in software and can be done for miscellaneous implementation phases. The proposed system can be modified for cyberbullying revelation in Non-English applications. The productive (effective) visualization can be promisingly met with the help of simpler notification about the occurrences of bullies. After identifying cyberbullying problem, relevant measures should be taken to prohibit further molestation of victim, preventing the spread of vulgar and immense messages. Analyzing, verifying and providing additional information helps the victims to take measures for getting rid of the problem.

V. References

[1] Weider D. Yu, Maithili Gole, Nishanth Prabhuswamy, Sowmya Prakash, Vidya Gowdru Shankaramurthy, “An Approach To Design and Analyze the Framework for Preventing Cyberbullying”, IEEE International Conference on Services Computing, 2016, pp.864-867, doi: 10.1109/SCC.2016.125 .

[2] Samaneh Nadali, Masrah Azrifah Azmi Murad, “A Review of Cyberbullying Detection :An Overview”, International Cmiference on Intelligent Systems Design and Applications (ISDA), 2013, pp.325-330, doi:10.1109/ISDA.2013.6920758 .

[3] Amrita Mangoankar, Allenoush Hayrapetian, ” Collaborative Detection of Cyberbullying Behaviour In Twitter Data”, 2015, IEEE Conference, pp.612-616, doi:10.1109/EIT.2015.7293405 .

[4] Silvana Tambosi, Gustavo Borges, Vanessa Mondin, Maria José Domingues, ” Cyberbullying: Concerns of Teachers and School Involvement”, International Conference on Complex, Intelligent, and Software Intensive Systems Cyberbullying, 2015, pp.416-420, doi: 10.1109/CISIS.2015.92

[5] Kelly Reynolds, April Kontostathis, Lynne Edwards, ” Using Machine Learning to Detect Cyberbullying”, International Conference on Machine Learning and Applications, 2011, pp.242-244, doi:10.1109/ICMLA.2011.152 .

[6] Ana Kovacevic, ”Cyberbullying Detection Using Web content Mining”, Telecommunications forum TELFOR, 2014, pp. 25-27, doi:10.1109/TELFOR.2014.7034560 .

[7] Tibor Bosse, Sven Stam ,”A Normative Agent System To Prevent Cyberbullying”, IEEE/WIC/ACM International

Conferences on Web Intelligence and Intelligent Agent Technology, 2011, pp. 425-430, doi: 10.1109/WI-IAT.2011.24.

[8] Homa Hosseinmardi Shaosong Li, Zhili Yang and Qin Lv, Rahat Ibn Rafiq Richard Han and Shivakant Mishra," A Comparison of Common Users across Instagram and Ask.fm to Better Understand Cyberbullying", IEEE International Conference on Big Data and Cloud Computing, 2014, pp. 355-362, doi: 10.1109/BDCloud.2014.87.

[9] Nektaria Potha, Manolis Maragoudakis," Cyberbullying Detection using Time Series Modeling", IEEE International Conference on Data Mining Workshop, 2014, pp.373-382, doi:10.1109/ICDMW.2014.170 .

[10] Marlies Rybnicek, Rainer Poisel and Simon Tjoa, "Facebook Watchdog:A Research Agenda For Detecting Online Grooming And Bullying Activities", IEEE International Conference on Systems, Man, and Cybernetics, 2013, pp.2854-2859, doi: 10.1109/SMC.2013.487 .

[11] Rekha Sugandhi, Anurag Pande, "Method for Detection of Cyberbullying:A Survey", International Conference on Intelligent Systems Design and Applications (ISDA), 2015, pp.173-177, doi: 10.1109/ISDA.2015.7489220 .

[12] Zinnar Ghasem, Ingo Frommhoz, "A Hybrid Aprroach to Combat Email Based Cyberstalking", FGCT International Conference on Future Generation Communication Technologies, 2015, pp.106-111, doi:10.1109/FGCT.2015.7300257

Cyberbullying Comment Classification on Indonesian Selebgram Using Support Vector Machine Method

Miftah Andriansyah [1]
Departement of Informatic Engineering
Sekolah Tinggi Teknik Multimedia Cendika Abditama
Indonesia
miftah@cendekia.ac.id

Ali Akbar [2]
Departement of Industrial Engineering
Universitas Gunadarma
Indonesia
Ali_akbar@staff.gunadarma.ac.id

Afina Ahwan [3]
Departement of Informatic Engineering
Universitas Gunadarma
Indonesia
afinaahwaan@gmail.com

Ardiono Roma Nugraha [4]
Departement of Informatic Engineering
Universitas Gunadarma
Indonesia
roma058a@gmail.com

Nico Ariesto Gilani[5]
Departement of Informatic Engineering
Universitas Gunadarma
Indonesia
nico.gilani@gmail.com

Rizki Nofita Sari[6]
Departement of Informatic Engineering
Universitas Gunadarma
Indonesia
rizkinofita@gmail.com

Remi Senjaya[7]
Departement of Informatic Engineering
Universitas Gunadarma
Indonesia
remi@staff.gunadarma.ac.id

Abstract- This paper aims to classify comments containing cyberbullying on instagram social media. Data were taken from the comments on instagram accounts of some well-known Indonesian Instagram celebrities/*Selebgram*, 1053 comments were taken as a training document and 34 comments were taken as a test document. Text classification method used is Support Vector Machine (SVM). The Support Vector Machine (SVM) method is used for classification types that only have two values, namely -1 and 1. In this classification process, the Support Vector Machine (SVM) method is used to see how far this method can classify comments on the Indonesian account Contains cyberbullying or not. The language used to create a text classification program with Support Vector Machine (SVM) is the R language using RStudio's Integrated Development Environment (IDE). The result shows that the classification of cyberbullying comments on instagram account of Indonesian program using Support Vector Machine (SVM) method shows the result of accuracy percentage of 79 , 412% .

Keywords ;cyberbullying; text classification; support vector machine; selebgram; R)

INTRODUCTION

Currently the use of social media instagram is very popular with Indonesian society. Quoted from www.wartaekonomi.co.id that Indonesia is the fourth most active market of instagram users in the Asia Pacific region with 54% of total internet users. Various age users that create social media instagram. Teen age is the dominant age that often accesses the internet, especially in instagram a social media platform.

The development of communication technology into a new container for teens at risk for violence. Negative effects of the internet ultimately lead to violent behavior on the virtual world called cyberbullying. The definition of cyberbullying is any harassment that occurs via internet, cell phones or other electronic devices. This type of bullying uses communication technologies to intentionally harm other people through hostile behaviour such as sending intimidating text messages and posting ugly comments on the Internet [1]. Supervised learning method most widely used are NBC and SVM with reason that NBC calculations is easier to do with a training on the data. SVM is widely used because it produces quite significantly the accuracy [9].

Cyberbullying consists of two individuals involved, namely the bully as the bullying actor and the victim as a bullying target [2][10]. In this study, the victims of cyberbullying observed were selebgram (celebrity instagram) which was popular with Indonesia. Instagram is chosen as the social media object that being observed because Instagram is a social network with the highest number of cyberbullying. According to the data from The Annual Bullying Survey 2017, 42% of respondents claim to have bullied in Instagram, this is the highest between otherdikara social network, like Facebook (37%), Snapchat (31%), Whatsapp (12%), YouTube (10%), Twitter (9%), and Tumblr (5%) [12]. Cyberbullying itself can have a more dangerous effect than physical bullying. Cyberbullying can potentially make the victim do suicide [15].

To avoid cyberbullying, Instagram had developed a system using machine learning technology to filter negative words. But, this system's word reference is still manually inputted by the user, so if the user got so many negative words in the comments, it is not so effective [16].

Selebgram often received comments from netizens that contain cyberbullying in their Instagram account. With the SVM (Support Vector Machine) method, these comment can be categorized as comments containing cyberbullying or not. SVM method is a learning system that uses space hypothesis in the form of linear functions of a feature space (feature space) high-dimensional, trained with learning algorithms based on optimization theory by implementing learning bias derived from the theory of statistical learning. The theory underlying SVM itself has evolved since the 1960s, but it was only introduced by Vapnik, Boser and Guyon in 1992 and since then SVM has grown rapidly. SVM is one of the relatively new techniques compared to other techniques, but it has better performance in various application fields such as text classification, bioinformatics, handwriting recognition, and so forth. The goal in this article is trying to classify a comment in an Instagram post, and categorized which words or comments that are containing cyberbullying, so it can prevent Instagram users from cyberbullying.

LITERATURE REVIEW

In previous studies cyberbullying detection in social media was detected by a binary classification task that is cyberbullying and non-cyberbullying [3]. Divyashree et al. (2016) conducted a research through a SVM classifier algorithm to detect cyber bullying messages from social media. The study applies a SVM classifier algorithm to trained the extracted features from training phase input sentences. Training phase input sentences are used to identify the cyberbullying from the testing phase by using the user comments. User comments are used to classify whether the comment belongs to cyberbully or not from each comment lexical and syntactic features are extracted these features [4]. Ducharme et al. (2017) conducted research to classify cyberbullying comments using a K-Nearest Neighbor/Support Vector Machine hybrid model. The training data in the study consisted of 350 comments sampled from the original data, ensuring that the two classes were balanced, that included 175 bullying comments and 175 non-bullying comments. The study shows the KNNFilter algorithm are able to reduce training data sizes by more than 50% and still obtain a high performing model. ost of the hybrid models had a cross-validated accuracy between 70% and 80% that built on a reduced data set, compared with the SVM model constructed on the full training data. Therefore the hybrid model of the study seems to indicate that the approach works well even in cases that use the real world data [5].

Michele et al. (2016) conduct research by adopt an unsupervised approach to detect cyber bully traces over social networks, based on Growing Hierarchical Self Organizing Map. The model of the study comprises several hand crafted features that are used to catch semantic and syntactic communicational behavior of potential cyber bullies. The study conducted some experiments on datasets taken from literature, like those coming from FormSpring and YouTube platforms, and also on a real data stream, collected from Twitter. The result of the study indicate that the model achieves reasonable performance and could be usefully applied to build concrete monitoring applications to mitigate the heavy social problem of cyberbullying [6]. Nahal et al. (2014) conducted research by applying SVM fuzzy algorithm to detect cyberbullying. In the study used 3 features, that is lexical features (eg number of swearwords and capitalized words), feature sentiments, and features based on metadata (eg user age and gender), these features are used to perform cyberbullying detection [7] [3].

Zhong et al. (2016) conducted a research to detect cyberbullying contained in the post social media Instagram. In the study, cyberbullying detection was performed using the development of EarlyWarning mechanisms to identify vulnerable posted images of attack. The study approached using more than 3000 images in the post of the Instagram photo-sharing network along with comments contained. The study utilizing new features of the determination of the topics obtained from the image description and pretrained convolutional neural network of the pixel image, in addition to standard images and text. The study resulted the potential targets for cyberbullying are on the classification of images and captions [8]. The main implication arise from the present findings is that the young internet The main implication arise from the present findings is that the young internet users are not aware of the level of endamagement of their online behaviors and its effects on other peoples' lives [9].

METHODOLOGY

In this article, the object of this research is the Instagram account belongs to selebgrams that popular in Indonesia, namely Karin Novilda and Samuel Alexander. The reason that we used this two selebgrams is that because they often uploading controversial posts. Karin Novilda or more popular with name *Awkarin* often upload inappropriate photos, hedonism lifestyle, and disrespectful words. This makes *Komisi Perlindangan Anak Indonesia* (KPAI) act because they assume Awkarin's behaviour is not exemplary [13]. In another side, Samuel Alexander or more popular with name *Young Lex* is a rap singer who had a duet with Awkarin. He also has similar character with Awkarin, even he once cursed his concert-goers while singing on the stage [14]. They also have many Instagram followers. Awkarin has two million followers, while Young Lex has 800 thousand, according to the data from Social Blade [11]. Therefore, in every photo they upload, there are almost always bullying comments, although there also many supporting comments.

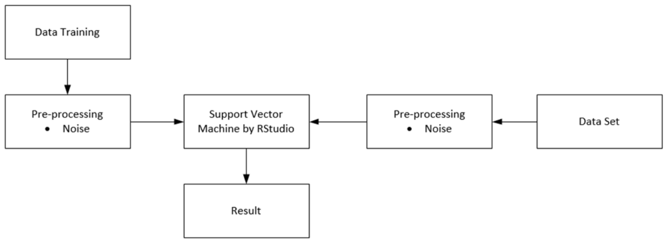

Figure 1. Cyberbullying Comment Classification Process Using Support Vector Machine Method

As in Figure 1, the processing of comment classification begins with the collection of comments obtained from the photos they uploaded on Instagram. In this article taken several photos, where each photo is taken comments randomly, so that obtained 1053 comments used as training set. Most commented languages still are not standard and contain noise like emoticons. Therefore, the comments are standardized, the abbreviations are outlined, and the emoticons are omitted. The comments are then classified manually whether they are bullying comments or not.

After all the comments have been classified and labeled, the comments are incorporated into the SVM model processed by using R. After obtaining the SVM model with the training set as many as 1053 data, the test is tested with 34 comments. Test sets are also obtained by randomly selecting from the comments that exist. This is done to predict how much percentage accuracy SVM in testing a comment, whether included in the category of bullying or not.

Text classification begins by classifying some text that you want to classify. SVM requires text data that has been labeled. In this experiment the data has been labeled manually with the correct text data and label value. The SVM model can be formed from into the data already collected. The SVM model can also predict data that has no label. The data collected for the SVM model is 1053. Data is randomly taken from the account comments column https://www.instagram.com/awkarin/ and https://www.instagram.com/young_lex18/. The Cyberbullying Comment Classification was shown in Table 1. Comments that labeled -1 were classified as comments of cyberbullying and comments labeled 1 were not.

Table 1. Data obtained from 'awkarin' social media accounts that have been standardized and labeled

Real Comment	Standardized comments	Label *Bullying*
motivasi nya apansi lu post beginian?	Motivasinya apa sih lu post yang kayak gini?	-1
baguss buat nakutin tikus drmh ni	Bagus buat nakutin tikus di rumah nih	-1
Sensor dong kak	Sensor dong kak	1
Makin langgeng yo ka karin	Makin langgeng ya kak karin	1

The language used to create a text classification program with SVM is the R language using the RStudio IDE. The R language already provides a library that can be used to handle SVM and text classification, the 'e1071' library used to handle SVM and RTextTools libraries to handle text classification, RTextTools requires the e1071 library.

```
library(e1071)

library(RTextTools)


# Retrieve Data From Directory

dataDirectory <- "/home/nico/contoh-svm/svmtutorial/ClassifyTextWithR/" # folder

data <- read.csv(paste(dataDirectory, 'bullyingData.csv', sep=""), header = TRUE)
```

Table 2. Data set used

Text	*Category Bullying*
Motivasinya apa sih lu post yang kayak gini?	-1
Bagus buat nakutin tikus di rumah nih	-1

Sensor dong kak	1
Makin langgeng ya kak karin	1

The data set uses only standardized comments along with its labels. Table 2 shows that the data has two columns of text and catBull. The text field contains predefined comments while the catBull column contains label information. Labels with captions -1 if comments include negative / bullying comments and 1 if comments are not bullying

IMPLEMENTATION

The data that has been collected as table 2 will be changed into Document Term Matrix form. By using the library from 'RTextTools', the Document Term Matrix can be generated with the create_matrix function.

```
# Create a document term matrix

dtMatrix <- create_matrix(data["text"])
```

A container is required to hold the Document Term Matrix for use for creating an SVM model using the 'RTextTools' library. The container configuration shows that all data sets will be a training set.

```
# Configure the training data

container <- create_container(dtMatrix,
data$catBull, trainSize=1:1052, virgin=FALSE)



# train a SVM Model

model <- train_model(container, "SVM",
kernel="linear", cost=1)
```

Once the SVM model is formed, the model can be used to predict a comment.

```
# comments to be predicted

predictionData <- list(

"Cie bang foto nih",

"Dasar engga punya otak",

"Sehat selalu",

"Mukanya kayak jamban",

"Abg kece")



# Create A Predefined Document Term Matrix

predMatrix <- create_matrix(predictionData,
originalMatrix=dtMatrix)



# Create Appropriate Containers

predSize = length(predictionData);

predictionContainer <-
create_container(predMatrix,
labels=rep(0,predSize), testSize=1:predSize,
virgin=FALSE)



#Prediction

results <- classify_model(predictionContainer,
model)

results
```

Table 3. Results from predicted classification using svm

No	SVM_WORDS	SVW_LABEL	SVM_PROB
1	Cie bang foto nih	1	0.6160739
2	Dasar engga punya otak	-1	0.5856106
3	Sehat selalu	1	0.6246462
4	Mukanya kayak jamban	-1	0.8222283
5	Abg kece	1	0.5826695

As expected in table 3, sentence number 1 is classified as a comment that contains no bullying elements and sentence number 4 is classified as a bullying comment. In sentence 2 has been classified as a comment containing bullying elements and sentence number 5 is classified as a comment that does not contain bullying elements, but with a low probability. This low probability indicates that the model is not too sure of this prediction.

In this article, we tested the SVM model with a test set containing 34 data. From 34 comments that tested, 27 comments were true according to the manual classification. So, the accuracy level of this SVM model is:

$$\frac{27}{34} * 100\% = 79{,}412\%$$

CONCLUSION

From this research, tested SVM models on test set counted 34 data. The SVM model is able to classify the test set with an accuracy of 79.412%. Of the 34 comments tested, 27 comments resulting in the same with the manual classification, either classified into positive about bullying or classified into non-bullying.

The level of accuracy may be enhanced by increasing the training set and developing the semantics of the comment. Semantics can improve accuracy because sometimes they are satire. The language of the comment is not classified into bullying, but in a sense, the comment belongs to bullying. For further research, it is possible to develop a model to detect cyberbullying, but taking into account the semantics of the commentary. For further implementation, this model can be applied to Instagram, for example if a comment in an Instagram post is indicated as a bullying comment, it can be omitted from that post.

ACKNOWLEDGEMENT

This research was fully supported by Universitas Gunadarma, Jakarta, Indonesia. The authors gratefully acknowledge Universitas Gunadarma for providing research funding and for permission in using the research facilities.

REFERENCES

[1] P. Galán-García, J.G. de la Puerta, C.L. Gómez, I. Santos, P.G. Bringas, "Supervised Machine Learning for the Detection of Troll Profiles in Twitter Social Network: Application to a Real Case of Cyberbullying," in Logic Journal of the IGPL, 2016.

[2] R.M. Kowalski, S. Limber, S.P. Limber, & P.W. Agatston, Cyberbullying: Bullying in the Digital Age, John Wiley & Sons, 2012.

[3] C.V. Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G.D. Pauw, W. Daelemans, V. Hoste, "Detection and Fine-Grained Classification of Cyberbullying Events," in RANLP 2015 (Recent Advances in Natural Language Processing), 2015.

[4] Divyashree, Vinutha H., Deepashree N. S., "An Effective Approach for Cyberbullying Detection and avoidance," in International Journal of Innovative Research in Computer and Communication Engineering, 2016.

[5] D. Ducharme, L. Costa, L. DiPippo, L. Hamel, "SVM Constraint Discovery using KNN applied to the Identification of Cyberbullying," The 13th International Conference on Data Mining, 2017.

[6] Michele Di Capua, Emanuel Di Nardo, Alfredo Petrosino, "Unsupervised Cyber Bullying Detection in Social Networks," in 23rd International Conference on Pattern Recognition (ICPR), 2016.

[7] V. Nahar, S. Al-Maskari, X. Li, C. Pang. "Semi-supervised Learning for Cyberbullying Detection in Social Networks," in ADC Databases Theory and Applications, pages 160-171, Springer International Publishing, 2014.

[8] H. Zhong, H. Li, A. Squicciarini, S. Rajtmajer, C. Griffin, D. Miller, C. Caragea, "Content-Driven Detection of Cyberbullying on the Instagram Social Network," in Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16), 2016.

[9] Andriansyah, Miftah & Suhendra, Adang & Wicaksana, Iwayan & Wicaksana, Simri. (2014). Comparative Study: The Implementation of Machine Learning Method for Sentiment Analysis in Social Media. A Recommendation for Future Research. Advanced Science Letters. 20. . 10.1166/asl.2014.5631..

[10] M.S. Ozden, S. Icellioglu, "The Perception of Cyberbullying and Cybervictimization by University Students in Terms of Their Personality Factors," World Conference on Education Science, Elsevier, 2013.

[11] Social Blade, "Youtube, Twitch, Twitter, Instagram Statistics," 2017, [Online], Available: https://socialblade.com, [Accessed: 20 July 2017].

[12] DitchTheLabel, "The Annual Bullying Survey 2017," Ditch The Label, 2017.

[13] S. Muslimah, "Mengenal Awkarin: Seleb Instagram dan YouTube yang Disorot KPAI," 2017, [Online], Available: https://news.detik.com/berita/d-3302012/mengenal-awkarin-seleb-instagram-dan-youtube-yang-disorot-kpai, [Accessed: 18 March 2017].

[14] F. Amanah, "Dilempari Penonton Ketika Manggung, Young Lex Murka dan Keluarkan Kata-Kata Kotor Ini!," 2017, [Online], Available: http://style.tribunnews.com/2017/01/02/dilempari-penonton-ketika-manggung-young-lex-murka-dan-keluarkan-kata-kata-kotor-ini, [Accessed: 18 March 2017].

[15] M. Ricky, "4 Hal yang Bikin Warganet Indonesia Jadi Juara Cyberbullying Dunia," 2017, [Online], Available: http://m.solopos.com/2017/07/27/4-hal-yang-bikin-warganet-indonesia-jadi-juara-cyberbullying-dunia-837581, [Accessed: 28 July 2017].

[16] F.K. Bohang, "Instagram Jadi Media Cyber-Bullying Nomor 1," 2017, [Online], Available: http://tekno.kompas.com/read/2017/07/21/12520067/instagram-jadi-media-cyber-bullying-nomor-1, [Accessed: 28 July 2017].