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.

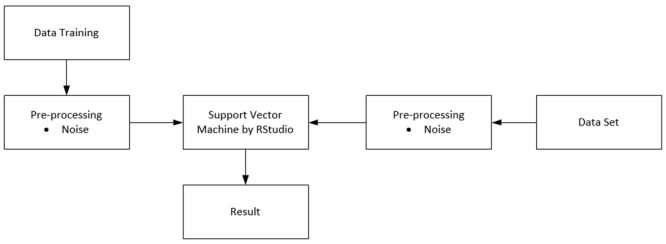

Figure 1. Cyberbullying Comment Classification Process Using Support Vector Machine Method

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].