JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 8, No. 3, September 2019, Pp. 110-118

Detection of Indonesian Hate Speech in the Comments Column of Indonesian Artists' Instagram Using the RoBERTa Method

Adhe Akram Azhari[1)], Yuliant Sibaroni[2)], Sri Suryani Prasetiyowati[3)]
1, 2, 3) Informatics, School of Computing, Telkom University, Bandung, Indonesia

Article history:
Received 17 August 2018
Revised 15 February 2019
Accepted 4 April 2019
Available online 4 April 2019

DOI :
https://doi.org/10.29100/jipi.v4i1.781

* Corresponding author.
Corresponding Author
E-mail address:
adheakramazhari@students.telkomuniversity.ac.id
yuliant@telkomuniversity.ac.id
srisuryani@telkomuniversity.ac.id

I. *Introduction*

HATE speech is an act that shows hatred towards one person or group. This action aims to have a certain impact either directly or indirectly [1]. Most Indonesian people take this action through social media. Because social media provides a comment feature where everyone is free to give their opinion but not a few people misuse this feature. The impact of this action is in the form of violence, depression, violence, and social conflict [1]. For example, Indonesian artists often get hate speech from the public. This action can make the artist become depressed and not rule out the possibility of taking his life. Therefore this action must be watched out for.

Instagram is one of the social media that is often used by the community, especially the people of Indonesia today. It has been noted that Instagram users are growing faster than Facebook. Instagram users have increased by 4.5 percent since the start of 2020[2]. This indicates that there are more Instagram users than on Facebook. There are several features provided by Instagram to upload photos, and videos and make comments. Users can take advantage of the comments feature to give their opinions regarding photos or videos that have been uploaded. Not a few Instagram users abuse the function of the comments feature. For example, spreading hate speech through the Instagram comment column to the person who uploaded the photo or video.

In this increasingly rapid technological development, many Indonesian people use the internet to carry out their daily activities. Especially in 2020, Indonesia was affected by COVID-19 where people were forced to do work at home. According to research conducted by We Are Social on April 23, 2020, Instagram users are growing faster than Facebook[2]. Instagram recorded an increase of 4.5 percent since the beginning of 2020. Globally Instagram is ranked fifth as a social media that has active users [2].

Artists in Indonesia mostly use Instagram in carrying out their work such as doing product promotions and uploading their works. Of course, after uploading content there will be a comment feature where people are given the opportunity to give praise, suggestions, and input to the artist. But not a few people blame this feature by giving comments in the form of hate. Most people feel free to give comments so they can be abused. Therefore there are often hate comments. Freedom in terms of commenting makes most Indonesian people not afraid to comment on their social media accounts.

So that Indonesia formed a Virtual Police program that aims to find social media accounts that provide comments in the form of hate speech[3]. In 2021 the Virtual Police program has been formed and has found 415 social media accounts that often display hate speech[3]. Comments are often found under the guise of criticism but these comments contain hate speech[3].

Based on the problems that occur, there are many studies that take the topic of detecting hate speech. To support the formation of a hate speech detection tool, machine learning is needed. Many studies use machine learning in detecting hate speech. In 2018 a hate speech detection machine was made using the CNN method[5]. This study obtained an average accuracy of 99.8%, a precision of 99.46%, and a recall of 97.99%[5]. This research get 2 types of output, namely comments that include hate speech or non-hate speech. In previous research, CNN serves to form patterns that will present classifications, and also previous researchers hoped that by using the Deep Learning method with the CNN algorithm, it would be able to detect an image that contains elements of hate speech. In 2021, research will be conducted on social media Twitter to detect hate speech [6]. This study uses a support vector machine to perform classification. In this study compared three kernels namely the RBF, linear, and sigmoid kernels. The results of this study RBF kernel has the highest accuracy value.

Next is research conducted by Yunita Suryani, et al. who once conducted a study related to hate speech on the Instagram social media of one Indonesian artist in 2021[7]. This study uses descriptive and qualitative methods [7]. This study aims to describe hate speech made by people who hate the artist. In 2020 a study was conducted regarding the analysis of comment sentiment on Instagram social media[8]. This study used the TF-IDF method and the naïve Bayes classifier in carrying out the classification. The results of this study are that this method can detect hate speech on Instagram social media by 92%[8].In 2019 Sakti Putra Perdana B.B once made a hate speech detection machine in the Instagram comment column[9]. This study used a deep neural network classification method. The results of this study are the average precision, recall, and F1 values of 97% and 97.19% accuracy and an average classification time of 5.22 seconds[9]. Annisa Briliani has also made a hate speech detection machine in the Instagram comment column using the k-nearest neighbor classification method [10]. The results of this study showed an average value of precision, recall, and F1 of 96% and 96.22% accuracy [10]. In 2019 Elvira Erizal also made a machine for detecting hate speech in the Instagram comment column using the maximum entropy classification method[11]. The results of this study obtained a precision value of 92.18%, a recall of 90.38%, an F1 of 90.44%, and an accuracy of 90.56% [11].

Yinhan Liu, et al conducted a study related to RoBERTa in 2019 [12].There are, 160GB of data was used to be examined. This study uses three different benchmarks to measure the performance of the RoBERTa model, namely GLUE results, SQuAD results, and RACE results. In the GLUE results, the performance of RoBERTa is superior compared to the large XLNET and large BERT models [12]. Furthermore, the results of SQuAD results RoBERTa get the highest accuracy value compared to XLNET large and BERT large [12]. In the last benchmark, RoBERTa's RACE results again outperformed the other models because RoBERTa's accuracy reached 86.5% [12].

In 2021 there will be research related to the uses of the RoBERTa method in detecting English [4]. In this study to compare the accuracy of several machine learning in detecting English, there are SVM, Logistic Regression, RoBERTa, and Meta-Classifier [4]. Of the four models, Roberta has the highest level of accuracy, namely 0.695 [4]. This proves that RobertTa is able to detect language more accurately than the fourth machine learning that has been compared [4]. Based on the latest research in 2021 proves the accuracy level of RoBERTa is higher compared to other models in detecting English. This shows that RoBERTa excels in detecting the English language. So this research will conduct trials on the RoBERTa method in detecting Indonesian.

This study detects hate speech comments in Indonesian. This study uses the RoBERTa method because the RoBERTa method has a high accuracy value in detecting language compared to other machine learning. This study aims to determine whether the use of full preprocessing can affect the accuracy level of the Roberta model in detecting hate speech in Indonesian and in general to see whether the Roberta model is good enough in detecting hate speech in Indonesian language social media.