1. Introduction

Cyberbullying is the act of threatening or endangering others by posting text or images that humiliate or harass people through the internet or other communication devices, such as cell phone or computer [1]. In Indonesia, a survey about cyberbullying has been conducted in 2019 by the Polling Indonesia and *Asosiasi Penyelenggara Jasa Internet Indonesia* (APJII)[2]. There are 5900 participants from all over Indonesia, and the obtained results are 49% participants claimed they have been bullied. 7.6% of 49% participants admitted to doing the same action in order to repay the bullying action.

One of the social media, that may let cyberbully acts occur is Twitter. There is quite a lot of user who posted a cyberbully sentence or picture with the intention to harras other people. Examples of cyberbullying sentences on Twitter are "*Dasar Belle gobl#k* (You stupid Belle)", "*Ahahaha, dasar anak mony#t lu, ibuk lo lont# ya* (Ahahaha, you son of a monkey, your mom is a sl#t, right)"

The bully itself can hide with a fake account, making it difficult for victims to find out information about the bully and the motives of the bully [3]. This of course will have a negative impact on victims such as lack of self-confidence, fear, changes of behavior to depression [4]. Therefore, an action is needed to prevent and reduce cyberbullying. One of them is by building a cyberbully detection system.

Previously, there was several research focusing on cyberbullying detection in Indonesian [5] -[7]. Adriansyah et al.[5] used a Support Vector Machine (SVM) to classify Instagram comment on Selebgram account. Noviantho et al.[6] used SVM and Naive Bayes to classify a cyberbully text. Nurrahmi et al.[7] used eight general rules from Sarna [8] for features extraction process and used SVM and KNN to classify the cyberbully text from Twitter. All of the previous research above using machine learning as the classifier where in a machine learning classifiers, there are no stages available for performing the features extraction process, so it must be done in different stages [9]. Also it is still have a few flaws such as lack of rules for features extraction process [7] and lack of data with class cyberbully.

But there was another research proposed by Banerjee et al.[10] used Convolutional Neural Network (CNN) to classify the Hindi cyberbully text and succesfully become the highest accuracy score compared to previous Hindi cyberbully detection system. Other research proposed by Sanguansat [11] used Doc2Vec as the features extraction method for sentiment analysis on social media, then compared the result with TF-IDF (Term Frequency-Inverse Document Frequency). [11] shows that the accuracy rate of Doc2Vec has higher score then TF-IDF.

Therefore, in this research, we proposed another experiments on cyberbully detection using another method such as deep neural network, CNN, as the classifier and Doc2Vec as the features extractions methods. Deep neural network are mathematically complex evolution of machine learning. CNN contains a convolutional and pooling layer which can be used to extract the important features from data. Doc2Vec had the ability to identify the semantics of texts and can be used to obtain the characteristic of cyberbully class.

Major contributions in this research are analyze the effect of Doc2Vec as features extractions method and analyze the performances of the CNN model in classifying the given dataset and compared it to baseline classification method namely Support Vector Machine (SVM) and Random Forest (RF). We also conducted two scenarios of experiment, using original dataset which is imbalance, and balance dataset as a result of downsampling.