JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 7, Nomor X, Bulan 2023, Page 999-999
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI 10.30865/mib.v5i1.2293

Misogyny Text Detection on Tiktok Social Media in Indonesian Using the Pre-trained Language Model IndoBERTweet

Abstract− Social media is a popular communication and information platform due to its ease and speed of access. By using social media, one can express himself freely. This triggers irresponsible individuals to utter hate speech with the aim of bringing down a person or group of people. Misogyny is a form of hate speech directed at women. The problem of misogyny should not be underestimated because misogyny can be one of the main reasons women feel miserable. In this study, a model will be built to detect misogyny text on the Indonesian language TikTok social media using the IndoBERTweet pre-trained model. IndoBERTweet is a pre-trained model based on the BERT model, which has been trained using Indonesian language datasets taken from the previous Twitter social media, resulting in a good performance for detecting misogynous texts on social media by classifying them. The dataset used is in the form of text data taken from misogyny comments by focusing on forms of misogyny in the form of stereotypes, dominance, sexual harassment, and discredit in short video content on women's TikTok social media accounts. The performance of built model performs hyperparameter settings which include batch size 16, epochs 10, and learning rate 7e-5 and is evaluated using a confusion matrix with the best accuracy results of 76.89%.

Keywords: Misogyny; BERT; Pre-trained Model; IndoBERTweet

1. INTRODUCTION

Social media is one of the most popular communication and information platforms today. The popularity of social media cannot be separated from the visuals, convenience, and speed of access, which are its main attractions. By using social media, a person can express himself freely without any restrictions. This freedom of expression triggers many irresponsible individuals to commit crimes in the form of hate speech online that aim to bully or bring down a person or group of people [1].

One of the social media that is becoming a trend in Indonesia is TikTok. According to Statista, a website that collects statistical data on social media in the world, recorded until July 2022, Indonesia ranks second in terms of the number of TikTok users, with a total of 99 million active users.

Hate speech is carried out online through social media regardless of gender. According to OXIS, an Oxford survey website, both a man and a woman can still be the target of online hate speech by irresponsible persons, but women are still more likely to receive hate speech when compared to men.

Misogyny is one form of negativity that is often shared or given on social media. Misogyny is a form of hate speech against women, whether it is directed at individuals or groups. Furthermore, misogyny is a problem that cannot be underestimated. This is because misogyny is the main reason women around the world feel miserable [1], [2]. Misogyny can be categorised into several forms of behaviour, such as Stereotype, Dominance, Sexual Harassment, and Discredit [3]. Misogyny behaviour is increasingly common along with the development of social media as a forum for expressing free and anonymous opinions, especially during the COVID-19 pandemic [4]. The continued increase in misogyny behaviour every year is a concern that should not be underestimated, and a solution must be sought to resolve it.

Misogyny can be categorized into several forms of behaviour such as humiliating, sexual harassment, domination, and discrediting. Shaming is a form of misogyny that restricts or belittles women because of some physical characteristic. Sexual harassment is a form of misogyny in the form of requests or statements to take actions that are sexually directed, such as sexual comments, crude jokes, and constant invitations to have extramarital affairs that can make you uncomfortable. Discredit and domination are forms of misogyny in the form of harsh expressions and/or statements that men are superior to women [1], [5], [6].

Misogyny text detection is a task that is done to detect whether a text or sentence is an expression of misogyny or not. In terms of text detection or recognition, the task of misogyny text detection has some comfort with sentiment analysis. In sentiment analysis, we detect whether a text has a positive or negative meaning towards a target label, while in misogyny text detection, we search whether a text has misogyny or not [3].

Research related to misogyny text detection has been done before. As in research [1], [3], [5], analysis of the problem of misogyny text detection was carried out using datasets from social media Twitter using 2 main tasks, where the first task aims to detect misogyny behaviour using 2 labels namely misogyny and non-misogyny, while the second task focuses more on detecting misogyny behaviour more specifically, such as incriminating, thwarting, discrediting, domination, sexual harassment, stereotyping & objectification, threats of violence, and so on. In the first study [1], the detection of misogyny texts was carried out using an Arabic language dataset taken from social media Twitter. The model used in this research is BERT using the pre-trained MARBERT model with the best performance results from the model being 91.74% accuracy on task 1 and 80.81% accuracy on task 2. Then in the second study [3], the detection on misogyny was carried out using English and Spanish datasets taken by social media Twitter using the Support Vector Machine or SVM model. The best performance results obtained in this study were 81.47% accuracy using the English dataset for task 1 and 54.22% F-target using the Spanish language dataset for task 2. In the third study [5], misogyny dedication was carried out using the dataset is in the form of a meme image embedded using GloVe to retrieve the text contained in the meme image. The main model used is BERT with the best performance results of 66.24% score for task 1 and 66.76% score for task 2.

Submitted: **dd/mm/yyyy**; Accepted: **dd/mm/yyyy**; Published: **dd/mm/yyyy**

In research [7], the problem of detecting misogyny was analysed using an Indonesian-language dataset taken from social media Twitter using only 2 labels, namely misogyny and non-misogyny. In this study, an experiment was conducted by comparing the effect of BERT Embedding on LR, CNN, and LSTM. The best performance results in this study were 86.15% accuracy and 81.37% F1-score obtained using the LSTM model.

The most visible shortcoming in previous studies is that most of the data used were in English and it was rare to find research using data sets in Indonesian. In addition, research on misogyny detection to classify using 2 labels, has an average performance of around 81%. However, this performance is better when compared to the performance of classification using more than 2 labels, which is equal to 67%. This is because the composition of certain labels in the dataset used indicates an imbalance condition [1], [3].

In previous studies, the Natural Language Processing or NLP approach and machine learning have been used to detect misogyny text on various types of social media [3], [8], [9]. The best performance result achieved in detecting misogyny text is 90%, which is achieved in detecting English content containing misogyny on social media Twitter using the BERTweet pre-trained method [9]. The BERTweet pre-trained model has been trained using datasets from social media Twitter with the limitation of only being trained using English datasets [8]. Similar to BERTweet, IndoBERTweet is a pre-trained model based on the BERT model that has been trained using datasets from Twitter in Indonesian [10]. Thus, using pre-trained IndoBERTweet can overcome the limitations of the pre-trained BERTweet model, which is only trained using English datasets.

The drawback of the previous misogyny text detection research is that most of the datasets used come from social media with English text [11] and Spanish [3]. In addition, research on the detection of misogyny on social media is mostly carried out using datasets from social media Twitter and Instagram.

This research will focus on detecting text misogyny on TikTok's Indonesian social media using the NLP approach and the BERT model. The BERT approach method is transfer learning by using the IndoBERTweet pre-trained model.

2. RESEARCH METHODOLOGY

2.1 General System

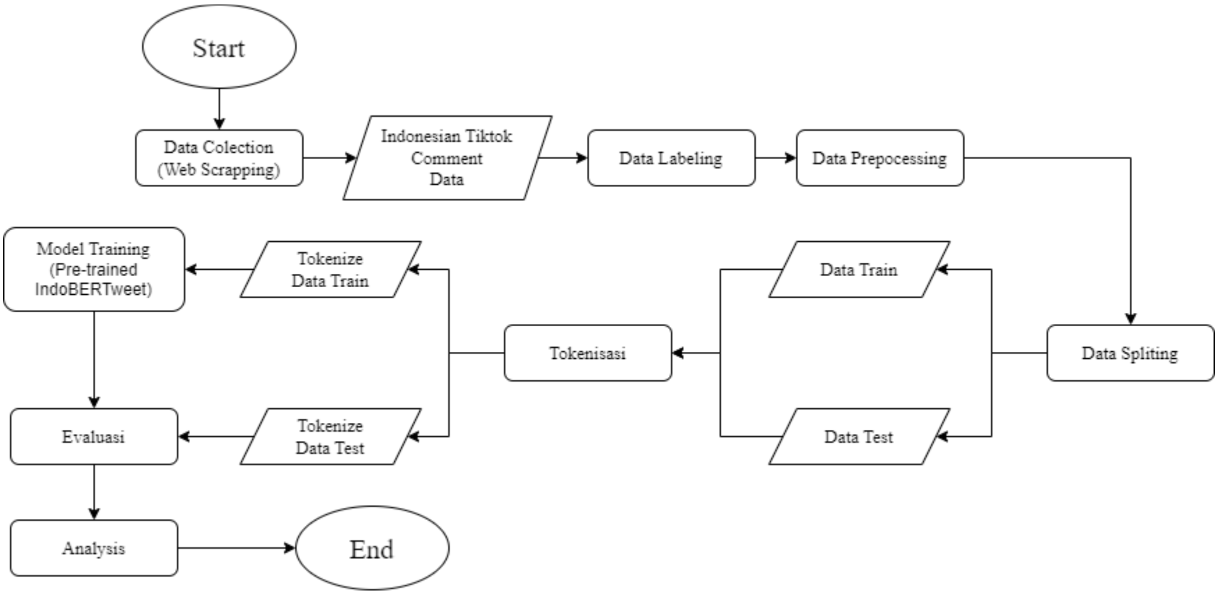

Figure 1. General System Design

In Figure 1, you can see the stages of developing a misogyny detection system starting from collecting misogyny data which focuses on forms of misogyny in the form of stereotypes, dominance, sexual harassment, and discredit on the Indonesian social media TikTok using the Web scraper method. The misogyny data collected is in the form of comments with the intention of insulting and/or dropping short video content on women's TikTok social media accounts. After collecting data, the process continues with data labelling, pre-processing, and data splitting. Data splitting is done by dividing the dataset into two parts, namely training data and test data. Before conducting training on the model, the tokenization process will be carried out first on the train data and test data, to adjust the shape of the data in such a way that it can be accepted by the BERT model. After training the model, the process continues with evaluation and analysis to draw conclusions from this research.

2.2 Data Collection (Web Scrapping)

The dataset used in this study is in the form of text data taken from comments with the intention of insulting or dropping on accounts on the TikTok social media belonging to women, which are limited to Indonesian as many