Indonesian Headline Detection on Twitter Social Media

Kaenova Mahendra Auditama
School of Computing
Telkom Univeristy
Bandung, Indonesia
kaenova@student.telkomuniversity.ac.id

Mahendra Dwifebri Purbolaksono
School of Computing
Telkom University
Bandung, Indonesia
mahendradp@telkomuniversity.ac.id

Ade Romadhony
School of Computing
Telkom University
Bandung, Indonesia
aderomadhony@telkomuniversity.ac.id

I. Introduction

Social media is a versatile tool that can serve many purposes. One such purpose is to gather a vast quantity of information and data in a relatively short period of time as compared to other methods. Among the various social media platforms available, Twitter is a popular text-based option [1]. The information obtained from a large amount of text can be used for news classification [2], classification of social media user personality [3], and one of the most influential research forms is sentiment analysis on public opinion [4]–[8]. With easy data collection on Twitter, this has become one of the factors driving the rapid development of text-based research.

The ease of collecting data on Twitter can lead to irrelevant data on the desired topic. This can happen because the sender of a text cannot be determined. One form of irrelevance that often occurs is analyzing public opinion using data that has many headlines. When collecting data using certain keywords on Twitter, many headlines are gathered from journalistic organization accounts such as *@detikcom* or *@CNNIndonesia*. Many studies collect data without considering this issue, resulting in data that contains headlines [6]–[8]. This can affect the quality of the processed information and the irrelevance towards the formation of information in some domain, especially on public opinion.

In order to improve the quality of research that processes Twitter data, where the presence of headline texts in the collected data is not desirable, detection of headline texts can be performed. Detection of a specific type of text has been carried out in several previous studies. Examples of text detection that are often performed to help reduce crime and disturbances are spam text detection [9]. Another aspect that can help reduce social crimes is the detection of abusive language on Twitter [10]. Such studies have shown good performance using a model that utilizes Multilayer Perceptron (MLP). Performance improvement using MLP models has occurred in conjunction with the discovery and utilization of transformer-based models that are able to perform text type detection better [10].

Based on these issues, this research develop an accurate and relevant model for processing data taken from Twitter by detecting headline and non-headline texts. In this study, data was obtained from Twitter, which includes headlines from journalism organization accounts and individual accounts, which were then validated by journalism organization to ensure data quality. This data then used as training data. Furthermore, fastText classification model (fastText), Convolutional Neural Network with fastText representation (CNN), and IndoBERTweet models were trained to detect headline and non-headline texts. With the trained models, the headline text found in data collection on Twitter social media can be analyzed.

We found IndoBERTweet, CNN, and fastText model have great performance on detecting headline text with highest average accuracy of 0.93. In addition, the proposed data collection method used in this study resulted in high-quality data with low error rates, establishing it as a reliable approach for collecting technique of headline and non-headline text data. Furthermore, it was found that more than half of the data obtained from the four COVID-19 related topics Twitter queries were predicted as text headlines, indicating the need for proper processing to provide accurate and relevant information.