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

***Abstract*— Social media post such as tweet in Twitter is one of main source of text analysis. However, the large amount of data collected often contain noise, such as irrelevance post that affect system inaccuracy. One that may affect system inaccuracy is the abundance of headlines text. Headlines text is a part of the text that describes the text below it. Therefore, this study designed an effective method to collect text headlines and non-headline texts, as well as building a model to detect text headlines. The data collection method developed had low error rates, making it a good reference for obtaining text headlines and non-headline text. The model built to detect text headlines had good performance with the highest accuracy is achieved by the IndoBERTweet model at 0.9921. The CNN and fastText models built had lower accuracy of 0.9081 and 0.8793 respectively, but were 200-2000x faster than IndoBERTweet. Moreover, our findings revealed that more than half of the total data in the four COVID-19 related topics comprised text headlines. Furthermore, the observed errors in the studied models can be attributed to the complex intrinsic properties of a headline text, fastText limitations in representing non-standardized words, and the training data used to build the fastText model.**

Keywords—detection, headline, Twitter, IndoBERTweet, CNN, fastText