Improving Indonesian Text Classification Using Multilingual Language Model

Ilham Firdausi Putra  
*Sekolah Teknik Elektro dan Informatika*  
*Institut Teknologi Bandung*Bandung, Indonesia  
ilhamfputra31@gmail.comAyu Purwarianti  
*Sekolah Teknik Elektro dan Informatika*   
*Institut Teknologi Bandung*Bandung, Indonesia  
ayu@stei.itb.ac.id

*Abstract*—Compared to English, the amount of labeled data for Indonesian text classification task is very small. Recently developed language model called multilingual language model have shown its ability to create multilingual representations. In this paper, we investigate the effect of combining English and Indonesian data on building Indonesian sentiment analysis and hate-speech classifier using multilingual language model. Using feature-based approach, we observe its performance on various data size and total added English data. The experiment showed that the addition of English dataset, especially if the amount of Indonesian data is small, can improve the performance. Using fine-tuning approach, we further showed its effectiveness at utilizing English language on building Indonesian text classification model.

Keywords—multilingual language model, text classification, sentiment analysis, hate-speech classification, Indonesian text

# Introduction

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vestibulum et posuere ante.

# Related Works

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vestibulum et posuere ante.

# Indonesian Text Classification

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vestibulum et posuere ante.

# Experiment

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vestibulum et posuere ante.

# Results and Dicussions

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vestibulum et posuere ante.

1. Average F1-Score Gains

| Total Data | Average Gain | |
| --- | --- | --- |
| *XLM-R* | *mBERT* |
| 500 | 0.176221 | 0.176221 |
| 1000 | 0.129394 | 0.129394 |
| 2500 | 0.165718 | 0.165718 |
| 5000 | 0.109215 | 0.109215 |
| 7500 | 0.118456 | 0.118456 |
| Max | 0.051226 | 0.051226 |

# Conclusion

In this work, we investigate the use of Indonesian and English text data with multilingual language model to solve Indonesian sentiment analysis and hate-speech classification. Employing two different technique, we investigate the lorem ipsum dolor sit amet, consectetur adipiscing elit. Vestibulum et posuere ante

##### References

1. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*