

Pengaruh *Text Preprocessing* terhadap Analisis Sentimen Komentar Masyarakat pada Media Sosial Twitter (Studi Kasus Pandemi COVID-19)

Syifa Khairunnisa¹, Prof. Dr. Adiwijaya, S.Si., M.Si.², Said Al Faraby, S.T., M.Sc.³

^{1,2,3}Fakultas Informatika, Universitas Telkom, Bandung

¹syifakhnsa@students.telkomuniversity.ac.id, ²adiwijaya@telkomuniversity.ac.id,

³saidalfaraby@telkomuniversity.ac.id

Abstract

COVID-19 is a pandemic that is troubling many people. This has led to a lot of public comments on Twitter social media. The comments are used for sentiment analysis so that we know the polarity of the sentiment that appears, whether it is positive, negative, or neutral. The problem when using twitter data is that the tweet data still contains many non-standard words such as abbreviated writing due to the maximum limitation of characters that can be used in one tweet. Preprocessing is the most important initial stage in sentiment analysis when using Twitter data, because it affects the classification performance results. This study specifically discusses the preproceesing technique by performing several test scenarios for the combination of preprocessing techniques to determine which preprocessing technique produces the most optimal accuracy and its effect on sentiment analysis. Feature extraction using N-Gram and word weighting using TF-IDF. Mutual Information as a feature selection method. The classification method used is SVM because it is able to classify high-dimensional data according to the data used in this study, namely text data. The results of this study indicate that the best performance is obtained by using a combination of cleaning and stemming; and normalization of words, cleaning, and stemming with the same accuracy of 77.77%. the use of unigram results in higher accuracy compared to bigram. Mutual Information is able to reduce overfitting problems by reducing irrelevant features so that train and test accuracy is quite stable.

Keywords: COVID-19, twitter, sentiment analysis, preprocessing, support vector machine
