Abstract

Twitter is a microblogging service that allows users to send and receive tweets or messages with a limit of only 280 characters per tweet. This issue causes tweets to be very short in content, not always using correct grammar, and often using slang. Generally, due to these factors, the accuracy of topic classification experiments in tweets is low. Therefore, this study implements a feature expansion to reduce vocabulary mismatch and a feature with a value of 0 to the value of its word similarity if its word similarity appears in the tweet. The feature expansion process can create a vector representation of tweets whose high dimensions and sparse can make the model get the semantic information and produce good accuracy. This feature expansion method looks for the similarity of the word using *fastText*. The results showed that the topic classification system using the Gradient Boosted Decision Tree method and with feature expansion had the highest accuracy of 91,39%.