

1. INTRODUCTION

Twitter is a social media platform that enables its users to send and read short text-based messages, limited to a maximum of 280 characters tweet [1]. Over time, Twitter has evolved into one of the largest sources of information, providing a convenient, rapid, and reliable platform for users to share anything happening around them with friends and followers. Consequently, Twitter can be utilized for conducting research on topic detection. Topic detection has been previously explored using techniques such as Word Embedding [2].

Topic detection is the process of assigning subjects to unstructured data or text. Classifying topics for documents (text), it allows for searches, statistical analysis, and meaningful classification [3]. In this research, topic detection is employed to retrieve a series of text documents (tweet corpus) and provide a set of topics for analysis, describing the content of each tweet in the corpus. There are challenges in categorizing tweets based on topics due to the character limit of 280 characters, leading to issues with grammar, slang, and low-quality tweets [2]. The utilization of slang terms in tweets can cause variations in vocabulary, making it difficult to comprehend the intended meaning [4]. To address this problem, one approach is using feature expansion. Feature expansion is a technique that enriches the original text by adding a semantic component to obtain a broader perspective of the text document [5].

Related research on topic detection using expansion features, Yahya et al [3], conducted research on detecting topics on Twitter using the Gradient Boosted Decision Tree method. In this study, the authors used the extraction of the TF-IDF (Term Frequency Inverse Document Frequency) feature and the expansion of the fastText feature, with an accuracy of 91.39% and an F1 score of 91.44%. Twitter used data of 30360 data and a news corpus of 97.794 data.

In addition, Ramadhy et al [4], did some research to expand the feature set. Using FastText and Logistic Regression, this study recognizes Trending Topic Analysis on Twitter social media. This study has an accuracy rate of 76.39%. However, compared to the baseline, the accuracy of this study increased by 14%. From a basic accuracy of 76.25%, after implementing feature expansion.

In the case of hybrid methods, Xu B et al [5], High Performance Web Attack Detection Method based on the CNN-GRU Model, a hybrid experiment between CNN and GRU using word2vec feature extraction, This hybrid layer is created to increase the accuracy of a single method, with CNN as the input layer and GRU as the output layer, the accuracy results reach 97.56%, compared to CNN only 93.44% and GRU has a yield of 97.17%.

Cao B et al [6], it was shown that the CNN-GRU model with the CSIC 2010 dataset conducted an experiment on the Network Intrusion Detection Model Based on CNN and GRU showing increased accuracy compared to the single model. Using the NSL-KDD dataset, the CNN-GRU hybrid method has an accuracy of 99.69%, superior to CNN and GRU as a single model.

Feature expansion for sentiment analysis in Twitter [7]. in this study various tests were carried out on feature expansion with satisfactory results to increase the level of accuracy as well as the function of feature expansion to expand text documents which results in the system being more accurate in checking data, in this study This was tested using TF-IDF as a feature expansion method with logistic regression [8], NB, and SVM as the model, and the accuracy results reached 82.02%.

Based on the explanation previously described, so far no research has been found that has developed a Topic Detection system using the extended GloVe feature with the CNN-GRU hybrid method using datasets obtained from Twitter. The tweet data obtained generally contains comments in text form. Then, the tweet data will go through a feature expansion process to make it more accurate in categorizing tweets on topics based on the words loaded in the tweet.

The purpose of this research is to build and analyze a topic detection system using GloVe feature expansion with CNN-GRU which is tested by comparing single models, hybrid models, single-model feature expansion, and hybrid-feature expansion to find out how much the accuracy improvement is using a dataset in the form of text obtained from Indonesian Twitter.

The organizational structure of this research paper is as follows: section 2 describes the methods used in this research, section 3 describes the results and discussion, and section 4 contains conclusions.