

1. INTRODUCTION

Twitter is one of the most popular social networks online as a micro-blogging system [1]. Twitter can be used as a medium of information transmission and a place to freely express opinions on a topic, and it has widespread access [2]. Users can share articles or upload about anything happening in the world, so users can freely follow the topics being discussed. It becomes important because it can be used as a source of information in evaluating user responses to discussed topics, so it has a wide range of topics [3]–[5]. Word embedding was previously used in research classification [6]–[8]. Topic classification effectively explores data, links similar documents, and meaningful classification [9].

Twitter is unique, as this social media platform can only write text-based messages that do not exceed 280 characters [10]. Unlike other social media that have longer writing boundaries. Because tweets have short written limitations, tweets are often written in an unstructured grammar, using slang language, or even the content of tweets is irrelevant to a topic [5]. The use of unexplained grammar and varied vocabulary makes tweets difficult to detect, so it's a unique challenge to classify the topic on a tweet so that the information or discussion is relevant. Therefore, the extension of features in this research is used to address the problem.

Research on the classification of topics is multilabel, so the method used is machine learning or deep learning. The study [5] performed topic detection on Twitter using the Gradient Boosted Decision Tree method. In this study, the authors used the extraction of the TF-IDF 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.

Similar research was carried out in the study [8] employing Word2vec for feature expansion and TF-IDF for feature extraction to extend topic detection using the Gradient Boosted Decision Tree. The data used in this research are tweets and news data. Twitter data consists of 35,605 data and news data of 142,544 data. The results of this study obtained an accuracy of 85.44%.

Research [11] conducted testing by classifying the text as an essay. This study showed that the accuracy of the results of the RNN algorithm obtained a higher accuracy compared to the CNN algorithm. RNN gets 55% accuracy, while CNN receives 50%. The study used 2000 data essays. There is no such thing as a shortage or a scale of the lowest scale.

The main contribution to this research is to apply hybrid deep learning methods with the expansion of features to classify topics. As far as the researchers searched, no study of the classification of topics using Hybrid Deep Learning was found. Therefore, the study aims to analyze the influence of the hybrid deep learning merger of the CNN and RNN algorithms as topic classification by adding the GloVe feature expansion.

The combination of the CNN and RNN algorithms is chosen because the combinations of these two methods can handle text with different lengths so that the classification process is not limited to letters alone, as well as capable of dealing with complex and non-explicit features so that better decisions can be made [5]. The purpose of expanding the GloVe feature is to address word discrepancies in tweets. GloVe employs matrix factoring to capture word co-occurrence information and global statistics, enabling semantic relationships between words in documents [12], [13]. In this study, the data used is social media Twitter using Indonesian. A total of 55,411 datasets are used and then further processed using the Term Frequency - Inverse Document Frequency (TF-IDF) extraction feature and the expansion feature using the word embedding method of GloVe.