Topic Classification Using the Long Short-Term Memory (LSTM) Method with FastText Feature Expansion on Twitter

Bella Adriani Putri
School of Computing
Telkom University
Bandung, Indonesia
bellaadrp@student.telkomuniversity.ac.id

Erwin Budi Setiawan
School of Computing
Telkom University
Bandung, Indonesia
erwinbudisetiawan@telkomuniversity.ac.id

I. INTRODUCTION

The rapid advancement of technology has made social media the primary means of communication worldwide [1]. Twitter is among the most widely used and highly effective social media platforms for information dissemination [2]. Twitter is a popular social network where individuals can communicate by posting concise messages called tweets. The user count for Twitter in 2018 stood at 152 million users who actively use the platform daily, and there are 330 million users who engage with it monthly [2]. Users can send and read tweets with a character limit of up to 280 characters [3].

The limited character count often leads to tweets being written concisely and not always adhering to proper grammar [4]. The restricted tweet length also increases the use of emoticons [4], abbreviations, spelling errors [5], and slang terms [6] by users. This results in the use of various word variations, making it challenging to understand tweets without proper classification [6]. In this research, feature expansion is employed to address these issues. Feature expansion is a semantic procedure that enhances the original text, giving the impression that it possesses a greater size [7].

The main goal of this research is to evaluate the impact of FastText feature expansion in the context of classifying Indonesian-language tweets based on their topic classification. FastText can reduce vocabulary mismatches by calculating the similarity between words in the corpus [8]. This research also utilized Long Short-Term Memory (LSTM) as a classification technique. This research used LSTM because it performs better using memory cells compared to recurrent neural networks in general [9].

The primary contribution of this research is to examine and combine the LSTM method and feature expansion with FastText for topic classification. This research explores several scenarios, starting with data splitting on the tweet dataset and testing the LSTM method as the baseline model. Furthermore, the best max feature is determined during the feature extraction stage with TF-IDF before testing the feature expansion. Finally, the scenario testing involves the LSTM method with adding feature expansion using FastText, which is built using News corpus, Tweet corpus, and Tweet+News corpus. As a result, the anticipated outcome of this research is that FastText could increase the performance of topic classification.

This research's structure is as follows: The literature review will be provided in Section 2. The system design will be concentrated in Section 3. The results and discussions will be provided in Section 4. Finally, the research's conclusion will be presented in Section 5.