CHAPTER 1

INTRODUCTION

This chapter includes the following subtopics, namely: (1) Rationale; (2) Theoretical Framework; (3) Conceptual Framework/Paradigm; (4) Statement of the problem; (5) Hypothesis (Optional); (6) Assumption (Optional); (7) Scope and Delimitation; and (8) Importance of the study.

1.1 Rationale

The increased use of social media is happening so rapidly from day to day that it makes social media a very important source of information. Every user always expresses their opinions on social media. Twitter, as a popular microblogging platform, produces a large amount of user-made content in the form of short messages called tweets. With over 500 million tweets posted every day, the platform generates an immense amount of data that can be analyzed for various purposes [1]. Based on this, a system that can perform sentiment analysis that has the best performance is needed. Analyzing the sentiment in these tweets can provide valuable insights into public opinion, customer feedback, and trends [2]. Sentiment analysis, also known as opinion mining, is a solution to problems that is the process of collecting and analyzing individual opinions, thoughts, and influences on a variety of topics, products, and subjects. Sentiment classification is the task of sentiment analysis that determines whether a text is positive, negative, or neutral [3, 4].

CNN method has the advantage of capturing local relationships between word neighbors in a sentence but less understanding of long-distance dependencies between words [5]. The LSTM method is the result of the improvisation of the RNN that addresses the lost gradient problem. According to the study conducted, the use of LSTM can address the shortcomings of CNN by memorizing information for a long period of time [5–7]. The shortcomings of LSTM are easy to overfitting because of complex model [5].

Word embedding has different ways of embedding sentences. In Table 1.1 represent ways of word embedding work. For example, Fasttext learns word representation by considering sub-words [8]. However, it provides a high amount of computation. Glove uses global statistics or word co-occurrence to derive the semantic relationship of words in the corpus, but it cannot capture sub-word information [9]. It may cause out-of-vocabulary. Combining word embedding with multi-channel methods is a solution to overcome the limitations each word embedding method. Multi-channel embedding can also view the vector representation of the word embedding extensively, thus improving the performance of sentiment classification over using only a single word embedding. In the example provided, Fasttext can be used to get sub-words which can handle out-of-vocabulary, while Glove can be used to obtain semantics between words. Nevertheless, it can improve computation cost of the method.

Table 1.1: Comparison of Word Embedding Models

Word Embedding	**Work**	**Advantage**	**Disadvantage**
Word2Vec [8]	Learn linguistic patterns as linear relationships between words.	Captures local word relationship.	- Did not learn the subword representation. - Unable to overcome out-of-vocabulary (OOV) -Did not learn global semantic.
GloVe [8, 9]	Focus on learning words using global statistics and co-occurrence of words.	Better at representing global semantic relationships	Unable to overcome out-of-vocabulary (OOV).
FastText [8, 10]	Improvement of Word2Vec to learn words based on sub-words.	- learn languages morphologically. - Ability to handle out-of-vocabulary (OOV).	-Did not capture global semantic relationships - More n-grams increase model computation complexity.

So far, multi-channel word embedding research has been conducted by Dahou et al [3] using Multi Channel Word Embedding CNN (MCE-CNN) on Arabic data sets. MCE-CNN consists of two channels of CNN embedding model: the primary WEV (Word Embedding Vector) to address common sentiment features on layer embedding channel 1, and the domain-specific to address features on Twitter data and review on Layer Embedding channel 2. The proposed method can go beyond the previous method. The research conducted by Lin et al [12] using twitter datasets, the architecture used for the multi embedding of CNN is the same as the research [3] but uses several different word embeddings, namely Glove and Lexicon2vec. However, one of the challenges in this research is the construction of the lexicon used in Lexicon2vec. The quality and coverage of the lexicon significantly impact the performance of the model, as an incomplete or biased lexicon can lead to suboptimal word representations. The two channels are then merged into a feature matrix filtered with a layer of attention and forwarded to the next layer using CNN and LSTM. The results obtained by Multi-Channel CNN-LSTM (MCECNN-LSTM) have accuracy of 81.25% and loss of 0.442 better than MCE-CNN with Twitter dataset. The quality and coverage of the lexicon significantly impact the performance of the model, as an incomplete or biased lexicon can lead to suboptimal word representations.

Research related to multi-channel LSTM has been conducted by Azwar et al [13]. The architecture used is multi-channel based BLSTM (MCAB-BLSTM) in sarcasm detection on news headlines. It is related to multi-channel embedding because it uses two embeddings contained in each channel, namely Glove and Fasttext. Then the word embedding feature on each channel will be forwarded to the BiLSTM layer, attention layer, max polling layer, and relu layer. MCAB-BLSTM accuracy is 96.64% in news headline dataset.

1.2 Theoretical Framework

The theoretical foundation of this study are using multi-channel word embedding and LSTM networks.

1. LSTM networks: Introduced as an improvement to Recurrent Neural Networks (RNNs), LSTMs address the vanishing gradient problem by preserving information over long sequences, making them well-suited for sequential data analysis.

2. Multi-channel word embeddings: This method using multiple embedding to view the vector representation of the word embedding extensively. GloVe and FastText are used in this study. GloVe uses global co-occurrence statistics to learn semantic relationships between words, while FastText considers subword information, addressing out-of-vocabulary issues. The combination of these embeddings in a multi-channel framework strengthens the model's ability to understand text.

1.3 Conceptual Framework/Paradigm

The proposed research explores the relationship between word embeddings, sequential modeling, and sentiment classification performance. The conceptual framework consists of three main components: input, classification process, and output. The input comprises preprocessed tweets that are converted into numerical representations using GloVe and FastText embeddings, ensuring rich semantic representation. In the processing stage, sentiment analysis is conducted using an LSTM-based architecture, which integrates both embeddings in a multi-channel setup to capture diverse linguistic patterns effectively. Finally, the output consists of sentiment classification, where the model predicts whether a given tweet expresses a positive or negative sentiment.

1.4 Statement of the Problem

Word embedding is an important component in sentiment classification due to its ability to capture contextual information from text. However, single embedding methods have limitations in capturing the full meaning of words. FastText is able to look at words morphologically and address OOV in words, but is less effective in capturing global semantic relationships between words. In contrast, GloVe excels at understanding global semantic relationships through co-occurrence statistics, but is unable to handle OOV words and morphological information. Most of the previous studies use a single embedding method, so the contextual information obtained is still limited. In addition, CNN (Convolutional Neural Network) architecture is often used in multi-channel embedding but has the disadvantage of capturing text sequential dependencies due to its focus on local relationships between words. This research identifies that the combination of FastText and GloVe in a multi-channel LSTM architecture can overcome the limitations of single embedding models by improving model performance.

1.5 Objective and Hypotheses

The aim of this study was to improve the performance of sentiment analysis on the Twitter review dataset by using the LSTM method as a layer multi-channel embedding (MCE) rather than using the method on the previous research using CNN. The word embedding used for multi-channel embedding is Fasttext and Glove. The metrics used to see the performance of the proposed method are accuracy and loss. There are two hypotesis determined in this research. First, multi-channel embedding with Fasttext and Glove can improve performance over using single-channel embedding. Second, multi-channel embedding using LSTM method as a layer word embedding will give better metrics performances than using the multi-channel embedding CNN method.

1.6 Assumption

As for the hypothesis determined in this study, multi-channel embedding with Fasttext and Glove can improve performance over using single-channel embedding. Multi-channel embedding can view the vector representation of the word embedding extensively, thus improving the performance of sentiment classification over using only a single word embedding. Fasttext learns the word representation by considering sub-words, but has a high computational cost. Glove uses global statistics or word co-occurrence to infer the semantic relationship of words in the corpus, but it cannot capture sub-word information, which can lead to out-of-vocabulary. Combining word embedding with multi-channel methods is a solution to overcome the limitations of each word embedding method. So far, multi-channel word embedding research has been conducted by Dahou et al [4] using Multi Channel Word Embedding CNN (MCE-CNN) on Arabic dataset. MCE-CNN consists of two channels of CNN embedding model: the primary WEV (Word Embedding Vector) to address common sentiment features on layer embedding channel 1, and the domain-specific to address features on Twitter data and review on Layer Embedding channel 2.

The proposed method can go beyond the previous method. The research conducted by Lin et al [11] using twitter datasets, the architecture used for the multi embedding of CNN is the same as the research [4] but uses several different word embeddings, namely Glove and Lexicon2vec. In the study [4, 11], it was proven that multi-channel embedding on the twitter dataset can improve accuracy by 12% compared to using single embedding.

The next hypothesis determined in this study, multi-channel embedding using LSTM method as a layer word embedding will give better metrics performances than using the multi-channel embedding CNN method. CNN can extract local features [6, 7]. It is more suitable in image processing. LSTM effectively retains contextual information on features across long sequences [5–7]. It is suitable for word processing such as sentiment analysis. However, the disadvantage of LSTM is that it is a complex model that is vulnerable to overfitting [5]. To overcome this problem, it is possible to train with a large amount of data and reduce the complexity of the model [5].

The LSTM method is the result of the improvisation of the RNN that addresses the lost gradient problem. According to the study conducted, the use of LSTM can overcome the deficiency of CNN, i.e. lack of memory for a long period of time. In the study [12] it was seen that the use of LSTM methods was superior to CNN. In the study [5, 13], it was proven that the use of LSTM methods is better than the application of Logistic Regression, SVM, MLP, DNN, CNN, and RNN. On the implementation of hybrid deep learning in the study [7], implementation LSTM after layer embedding and CNN on the subsequent layer (LSTM-CNN) had better accuracy than implementation CNN-LSTM.

1.7 Scope and Delimitation

This reasearch focus on sentiment classification using Twitter Dataset. This dataset are from Lin et al. [12] contains 1.6 million tweets with positive and negative label. The dataset used english topics from various domains. The purpose of using topics from various domains is to focus on methods in Twitter dataset. Pretrained word embeddings are used in this research because they help reduce training time and are suitable for scenarios with limited computational resources.

1.8 Significance of the Study

The study contributes to the field of sentiment analysis by introducing a novel MCE-LSTM model in Twitter dataset. The goal is to improve accuracy and loss performance in Twitter dataset based on multiple embedding features and better contextual information based long-sequences. This significance study explores potential for applications in real-world sentiment classification task in Twitter dataset.