CHAPTER 1

INTRODUCTION

1.1 Background

X, previously known as Twitter, is a widely used social media platform for communication and expressing opinions through text. It contains many personal messages, opinions, and people's views on various topics [1]. Prior to important days such as general elections in Indonesia, X is often used for political activities such as campaigning, criticizing political figures, or specific political parties [2].

The year 2024 marks a pivotal moment for Indonesia as the nation prepares for its presidential election. The atmosphere is already charged with political activities and discussions, as candidates and their supporters gear up for a significant democratic event. The General Election Commission (KPU) has officially set the timeline for the presidential election, with the campaign period intensifying as the election date approaches.

This election sees prominent figures vying for the highest offices in the country. Among the notable candidates are Prabowo Subianto, who has received support from former President Joko Widodo, Anies Baswedan, and Ganjar Pranowo. The political landscape is further enriched by emerging leaders from various parties, reflecting the dynamic and competitive nature of Indonesian democracy.

The influence of social media has become more pronounced in this election cycle. Candidates and their supporters utilize platforms like X to shape public opinion, mobilize support, and engage with voters. This digital campaign strategy is crucial, given the high internet penetration and the increasing reliance on social media for information among Indonesians [3]. The intense activity on these platforms has led to a plethora of opinions, both positive and negative, regarding the presidential candidates and their policies.

In this context, sentiment analysis becomes a valuable tool for understanding public opinion, voter preferences and behavior. However, X posts are limited to 280 characters and the texts are usually unstructured and even very informal. This makes tweets on X distinctive for their irregularity of language and lack of clear context, making the classification of these texts a significant challenge in the field of sentiment analysis [4].

1.2 Theoretical Framework

Several machine learning-based text classification methods already exist, including Naïve Bayes, K-Nearest Neighbor, and Support Vector Machines. Additionally, convolutional neural networks have been used for text classification. Recurrent neural networks have also been employed, showing promising results in effectively conveying contextual infor-

mation in the text. Moreover, text classification mainly consists of two phases which are feature extraction phase and the classification phase[5]. Term weighting is one of the common ways to extract term features [6]. For instance, a sentimet analysis on US Airlines tweets dataset was conducted using several ML techniques and shows promising results with SVM outperforming others[7]. Other research involving deep learning architecture, combining TF-IDF weighted Glove word embedding with CNN-LSTM architecture, outperforms conventional methods in sentiment analysis on product reviews from Twitter[8].

However, the slowed testing and training speed of these models has limited their application. Table 1.1 [6] provides a summary of the advantages and disadvantages of several text classification methods.

Table 1.1: Summary of Text Classification Methods

Methods	**Advantage**	**Disadvantage**
Naïve Bayes	Requires less training time and data than other approaches	Limited by class imbalance, as a probability value must be estimated for each conceivable value
SVM	Yields high accuracy with large datasets	Fine tuning of model is very difficult and tedious. Long training time for large datasets.
Decision Tree	Less time for training	The model is susceptible to overfitting.
KNN	Data can be continuously incorporated over time without requiring explicit retraining.	Higher prediction complexity for large dataset and dimensions. Equal importance is given to all features.
RNN	Capture of sequential data, which is critical for sentiment text categorization	Train more slowly than other models. Complicated and computationally costly.
LSTM	More efficient compared to RNNs; can effectively capture long-term dependencies	The model is highly complex, resulting in extended training times.

1.3 Conceptual Framework/Paradigm

In recent years, fastText has become quite popular in text classification tasks [9]. Its main advantages are fast, low-cost, and efficient model training. There have been many studies utilizing fastText for word embedding. For instance, a study on detecting deepfakes on social media utilized deep learning techniques and fastText embeddings to discern tweets generated by machines [10]. The experimental findings indicate that using a CNN architecture with fastText embeddings achieves an impressive classification accuracy of 0.93 for tweet data. However, Facebook AI Researchers state in their paper [9] that fastText often

performs comparably to deep learning classifiers (CNN in this case) in terms of accuracy, but with much shorter training and evaluation times. Furthermore, some studies have used fastText as the text classification method. Among the studies that use fastText is the text classification on a Portuguese online news corpus for telecommunication companies [11]. The findings reveal F1-scores exceeding 0.78 for each of the three topics. Another research also concluded that FastText Supervised Classifier model effectively classifies Covid-19 tweets into emotions, providing a better understanding of people's mental health and how sentiments have changed over time [12].

1.4 Statement of the Problem

Although the fastText engineering model performs well compared to traditional text classifiers and is on par with deep learning classifiers, previous studies have not explored the use of feature weighting. A robust term weighting can improve the performance of text classification by considering the relation between words in the document [13]. Further research revealed that n-gram processing in fastText results in a vector representation formed from words that have a low frequency of occurrence and are meaningless in a dataset. The number of such words increases as the number of characters in the text increases, which in turn affects the performance of text classification [14].

To date, although there has been already some research on text classification using fastText as the classifier on the topic of the Indonesian Election 2024, previous research has not explored the effect of using different term weightings. This gap highlights the need for a study that focuses on this specific context.

1.5 Objective and Hypotheses

The primary objective of this research is to enhance the precision and efficiency of text classification for sentiment analysis in the context of the Indonesian Election 2024 and obtaining the most optimal combination of feature weighting with the fastText model[15]. This will be attained by leveraging various term weighting techniques namely TF-IDF, TF-RF, TF-IDFC-RF to obtained high quality word vector that contains words that are significant and relevant which are than fed to the fastText model. Furthermore, the research will assess and contrast the performance of models that incorporate the fastText approach with these various term weighting, fastText based model, and other common text classification method.

This combined approach is expected to exceed the performance of both previous research and other classification methods. It is hypothesized that by optimizing the configuration of the fastText input layer, removing irrelevant words from the n-gram processed word vectors, and thoughtfully weighting terms based on their occurrence and discriminating capacity, the classification model will improve its ability to distinguish and categorize

text more accurately. This approach differs from previous methods by introducing an optimized combination of fastText and term weighting, offering a novel perspective on text classification better than based model fastText and other common models.

1.6 Assumption

The study assumes that fastText can be optimized through modifications to its input layer and n-gram processing, and that term weighting techniques will significantly impact text classification performance.

1.7 Scope and Delimitation

This research focuses on optimizing text classification using fastText for social media data, specifically tweets related to the Indonesian Election 2024. The study will not cover other social media platforms or topics outside the election context.