1. Introduction

Background

Social media has evolved into a forum for the public to articulate their views on diverse societal issues. Social media has undergone swift expansion due to its significant impact. Twitter, now referred to as X, is a social media medium frequently utilized by the public to express their opinions. X is considered an engaging and uncomplicated conversation or forum platform. Prior sentiment analysis study [1], [2], [3] performed sentiment analysis on prevalent public topics, gathering data from X, which is regarded as offering a broader range of information expressed in comparatively straightforward prose. Nonetheless, despite the diversity of the acquired data, researchers frequently encounter obstacles, like restricted data availability for analysis and disproportionate opinion labeling within the dataset [4].

The development of huge language models is one area where the advancement of natural language processing (NLP) is becoming more and more obvious. [5]. The efficacy of an NLP model is significantly impacted by various supporting elements, among which data augmentation is crucial. Data augmentation is typically employed to enhance model training efficacy in diverse NLP tasks, such as sentiment analysis.

Data augmentation denotes techniques employed to enhance the quantity of data by incorporating significantly altered replicas of current data or by generating new synthetic data derived from existing datasets. Historically, data augmentation techniques were extensively utilized in computer vision using methods such as rotation, cropping, and various visual modifications to produce new data from images[6], [7]. Nonetheless, in NLP, a barrier exists in formulating universal principles for text transformations that yield novel linguistic patterns. The efficacy of a model is significantly influenced by the quality and quantity of training data. A primary impediment to enhancing the efficacy of an NLP model is the restricted availability of training data[8], [9]. Challenges often arise from factors like privacy policies and elevated operating costs, including the time-consuming and resource-intensive data annotation process. The expense of compensating people for manual labeling jobs might represent a substantial cost in model development, affecting time efficiency[9].

A previous study [10], [11] performed an extensive review of data augmentation in NLP, classifying augmentation methods according to particular methodologies; nevertheless, these classifications often appear overly restricted or broad, including categories like back-translation and model-based approaches. Another study [12] conducted a targeted survey on data augmentation for text classification. This study will utilize data augmentation through the generation of unique samples employing the paraphrase technique.

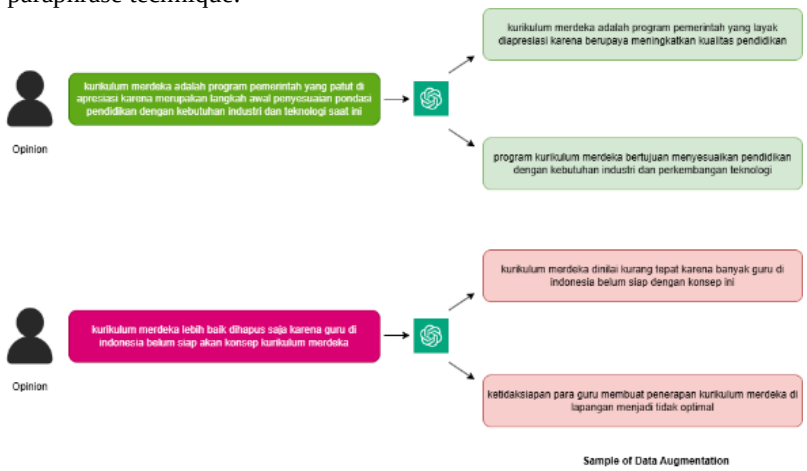

Fig. 1 Data Augmentation using Paraphrase Method

Topics and Limitations

This study seeks to assess the influence of result correctness in sentiment analysis tasks when comparing restricted data to augmented data. The data augmentation methodology is executed by prompting techniques and the application of a multi-turn dialogue approach to acquire the dataset.

Objective

The purpose of this research is to evaluate the performance of text classification with data augmentation and without data augmentation. Augmentation data is obtained from where the focus of the research is to compare the performance of text classification on sentiment analysis tasks.