## 1. Introduction

Social media, an online form of social networking, has experienced remarkably rapid growth. Today, most internet users around the world use social media [1], and their frequent use has a profound effect on many aspects of everyday life. Among the various social media platforms, X (formerly Twitter) remains exceptionally popular and serves a variety of purposes, from advertising to sharing personal thoughts or factual information [2].

Emotions are a fundamental component of human intelligence and play a crucial role in our daily decision-making. Although emotional connections are vital in human interactions, they are often overlooked. People typically experience a range of core emotions, such as happiness, sadness, anger, and fear [3]. On X, users express a wide range of emotions from joy and praise to criticism and anger, highlighting the platform's role as a space for emotionally charged communication shaped by temporal patterns [4]. For example, posts expressing happy often include positive affirmations or joyful language, encouraging likes, shares, or celebratory comments, which contribute to vibrant interactions. Meanwhile, posts indicating sad tend to have reflective or melancholic tones, often inviting supportive or empathetic responses. Posts conveying angry typically feature strong language or repeated words, frequently attracting debates or heated exchanges, thereby increasing engagement. In contrast, fear is usually shown through worry or apprehension, often prompting empathetic interactions or expressions of solidarity.

This research uses the Robustly Optimised BERT Pretraining Approach (RoBERTa) to classify emotions expressed by X users, focusing on the textual content of their posts. RoBERTa builds on the Bidirectional Encoder Representations from Transformers (BERT) architecture, demonstrating superior model performance due to improved training methods, larger datasets, and greater computational resources. Notably, RoBERTa removes BERT's Next Sentence Prediction (NSP) mechanism and replaces it with Dynamic Masking [1,5]. Its effectiveness has been validated across a broad range of tasks. For instance, F. I. Kurniadi et al. [6] applied RoBERTa to detect depression in social media posts, achieving 98% accuracy and an F1 score of 0.98. Similarly, Murarka et al. [7] used RoBERTa to classify mental illness-related posts on Reddit, obtaining F1 scores of 0.89 for combined inputs and 0.86 for post-only inputs, both demonstrating robust performance. Furthermore, M. A. A. Yani [1] utilized RoBERTa to analyze cyberbullying in Indonesian X posts, achieving 86.9% accuracy with an F1 score of 0.77, underscoring RoBERTa's adaptability to various linguistic contexts.

The primary focus of this research lies in addressing the challenges posed by slang and informal expressions in Indonesian social media. Slang, while rich in emotional cues, introduces significant complexity due to its informal nature. To address this, two preprocessing methods are examined: full preprocessing, which removes slang through normalization, stemming, and stopword elimination, and half preprocessing, which retains slang to preserve its informal linguistic nuances. By contrasting these methods, the research aims to provide insights into how preprocessing impacts the performance of RoBERTa in classifying emotions effectively in noisy and informal datasets.