

CHAPTER I

INTRODUCTION

This chapter introduces the background of the study by outlining the current context, identifying the research gap, and explaining the rationale behind the research. It also presents the research problem, objectives, significance, and limitations to clearly define the focus and boundaries of the study.

1.1. Background

The rapid advancement of information technology has significantly changed how people communicate, with traditional face-to-face interactions increasingly being replaced by internet-based communication through social media platforms [1]. This transformation has facilitated the rapid dissemination of information, expanded public participation in online discussions, and enabled users to share content that encourages interaction through comment sections. However, these advantages have also created opportunities for negative behaviors such as bullying. Bullying refers to repeated aggressive behavior directed toward individuals who are in a weaker position or unable to defend themselves. When such behavior occurs in an online environment, it is commonly referred to as cyberbullying [2].

Bullying has significant psychological and social consequences for its victims. Individuals subjected to bullying often experience severe emotional distress, including anxiety, prolonged stress, and reduced self-esteem. In more severe cases, bullying may lead to depression, social withdrawal, and even self-harm or suicidal ideation [3]. One of the most well-known examples is the case of Amanda Todd, who died by suicide after experiencing prolonged cyberbullying. She was subjected to online extortion and the unauthorized distribution of her private photographs, which resulted in continuous harassment across multiple social media platforms. The persistent psychological pressure ultimately had devastating effects on her mental health, leading to her suicide [4]. This case illustrates the severe risks associated with cyberbullying, which can occur at any time while allowing perpetrators to conceal their identities, making it potentially more harmful than traditional face-to-face bullying [5].

Numerous surveys have consistently demonstrated the prevalence of cyberbullying over the years. A 2019 UNICEF Indonesia survey reported that approximately 45% of respondents, representing 1,249 individuals, had experienced

cyberbullying [6]. Similarly, a 2021 study conducted by the Center for Digital Society found that 45.35% of adolescents aged 13 to 18 had been victims of cyberbullying, with Instagram identified as one of the platforms with the highest incidence rates [7]. The consistently high proportion of victims across these studies indicates that cyberbullying remains a persistent issue over time. This finding is further supported by the 2024 National Survey on the Life Experiences of Children and Adolescents, which reported that 10.71% of males and 9.33% of females aged 13 to 17 experienced cyberbullying within the previous 12 months [8]. These findings confirm that cyberbullying continues to be a significant social problem.

The high incidence of cyberbullying on Instagram is consistent with its large user base in Indonesia, which reached approximately 103 million active users, accounting for about 58% of Instagram users in Southeast Asia [9]. The 2025 Internet Penetration and User Behavior Survey also reported that Instagram is the second most widely used social media platform among Generation Z users aged 13 to 18 [10]. This demographic is particularly vulnerable because adolescents are still undergoing emotional and identity development [11], making them more susceptible to social pressure and online validation. In response, Meta introduced the Teen Account feature to provide additional safety measures, including anti-bullying protections [12]. However, Arturo Béjar and researchers from New York University and Northeastern University found that approximately 64% of these additional safety features were either unavailable or no longer functioning effectively, including anti-bullying mechanisms [13]. These findings suggest that Instagram's built-in protections are still insufficient to fully safeguard users from cyberbullying.

Given Instagram's massive user base, manual content moderation has become increasingly impractical, highlighting the need for automated detection systems powered by artificial intelligence. Previous research has shown that deep learning models, including Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), and Bidirectional Long Short-Term Memory (BiLSTM), are highly effective for text classification tasks [14]. Other studies have further improved performance through hybrid deep learning architectures [15][16][17][18]. Transformer-based models have also achieved strong results in text classification. Maulana and Aditya compared IndoBERT and BiLSTM for Indonesian cyberbullying detection and reported that both models achieved an identical accuracy of 97%, demonstrating that BiLSTM remains competitive with transfer learning approaches [19]. Furthermore,

several studies have shown that BiLSTM can outperform certain hybrid architectures in text classification tasks [20][21][22]. These findings indicate that BiLSTM remains an effective architecture under specific conditions. Therefore, it was chosen as the classification model for this study.

Despite considerable progress in text-based cyberbullying detection, most existing studies are limited to monolingual datasets. In contrast, bilingual and code-mixed language usage has become increasingly common across social media platforms [23]. Moreover, practical implementations that can be directly utilized by everyday users remain limited, as most studies primarily focus on model development and evaluation rather than deployment. This study addresses the challenge of bilingual cyberbullying detection involving both Indonesian and English, which introduces additional complexity in word representation, sentence processing, and semantic understanding. Furthermore, conventional deep learning models generally provide little explanation regarding the reasoning behind their predictions, making it challenging for users to understand why a particular comment is classified as cyberbullying. Local Interpretable Model-agnostic Explanations (LIME) is an explainable artificial intelligence (XAI) technique capable of interpreting model predictions by revealing the contribution of individual words to the final classification outcome [24]. However, previous studies have primarily employed LIME only during the evaluation phase rather than integrating it directly into an operational detection system [25].

To overcome these limitations, this study proposes the development of a browser extension for Instagram that automatically detects and filters cyberbullying comments using a BiLSTM model integrated with LIME. The browser extension was selected as the deployment platform because desktop devices currently account for approximately 43.68% of social media access, demonstrating that browser-based access remains a significant means of interacting with Instagram [26]. The primary contributions of this research include the development of a BiLSTM model for bilingual cyberbullying detection, the implementation of a browser extension capable of performing real-time detection on Instagram, and the direct integration of LIME to provide transparent explanations for model predictions. The proposed system is expected to provide an additional proactive and practical layer of protection for users, complement the moderation mechanisms provided by social media platforms, and reduce user's exposure to cyberbullying comments in online environments.

1.2. Problem Statements

Based on the background that have been explained earlier, the formulation of the problem in this study is as follows:

1. How does the BiLSTM model trained on a bilingual dataset perform when evaluated separately on Indonesian and English language data?
2. How can the LIME method be applied to provide interpretable explanations for the predictions generated by the BiLSTM model in cyberbullying detection?
3. How can a browser-based extension be designed and implemented to automatically detect and filter cyberbullying comments on Instagram while providing interpretable prediction explanations using LIME?

1.3. Research Objectives

Based on the research problems that have been explained earlier, the objectives of this study are as follows:

1. To evaluate the performance of a BiLSTM model trained on a bilingual dataset by testing it separately on Indonesian and English language data.
2. To apply the LIME method to provide interpretable explanations for the predictions generated by the BiLSTM model in cyberbullying detection.
3. To design and implement a browser extension capable of automatically detecting and filtering cyberbullying comments on Instagram while providing interpretable prediction explanations using LIME.

1.4. Research Contributions

Based on the research objectives that has been explained earlier, this study is expected to provide the following contributions:

1. To develop a browser extension that helps reduce user's exposure to cyberbullying comments on Instagram.
2. To facilitate faster detection and filtering of cyberbullying comments without relying solely on the platform's built-in moderation system.
3. To demonstrate a practical application of deep learning-based text classification and explainable artificial intelligence for enhancing digital safety and improving the quality of online social interactions.

1.5. Research Limitations

To clearly define the scope of this study and avoid an overly broad discussion, the following limitations are established:

1. The proposed browser extension is developed exclusively for Chromium-based browsers, such as Google Chrome and Microsoft Edge.
2. The detection system is limited to text-based comments posted on public Instagram posts.
3. The dataset consists of bilingual data in Indonesian and English and is limited to the harassment category of cyberbullying.
4. Model evaluation is conducted incrementally, where the best-performing configuration from each experimental stage is used as the basis for the subsequent stage.
5. The system operates under the assumption that only one Instagram browser tab is active at any given time.