

CHAPTER V

CONCLUSIONS AND SUGGESTIONS

This chapter summarizes the conclusions drawn from the research findings and presents recommendations for future work. The conclusions summarize the findings obtained from the system design, implementation, and evaluation processes. The recommendations are provided as suggestions for further development and improvement of the research in the future.

5.1. Conclusions

Based on the research results that have been obtained, several findings were identified to answer the research problems and objectives. The conclusions of this study are as follows:

1. Based on the evaluation results, the BiLSTM model trained using a bilingual dataset was successfully applied to detect bullying in both Indonesian and English texts. The best-performing model achieved an accuracy of 95.97%, precision of 95.98%, recall of 95.97%, and F1-score of 95.96%, indicating strong and balanced classification performance across both classes. This performance was obtained using the configuration of a maximum sequence length based on the 90th percentile of 22 tokens, a data split ratio of 80:10:10 for training, validation, and testing data, a batch size of 16, a learning rate of 0.01, 32 LSTM hidden units, a dropout rate of 0.4, and a threshold value of 0.5. These results indicate that training using a bilingual dataset can produce an effective model for bullying detection across two languages without requiring separate models for each language.
2. The LIME method was successfully implemented to interpret the prediction results of the BiLSTM model on six test samples, consisting of two normal-class comments, two bullying-class comments, and two misclassified comments. LIME was able to identify tokens that contributed positively and negatively along with their corresponding weights, making the prediction process more transparent. In one misclassified sample, the word “kamu” had the highest contribution with a value of 0.70, while words that were more relevant as indicators of cyberbullying, namely “congor” and “neraka”, only contributed values of 0.07 and 0.05, respectively. These results indicate that the model focused more on less relevant tokens, which led to incorrect predictions. Therefore, LIME not only provides an

explanation of the model decision but also helps identify sources of prediction errors and can serve as a basis for future model improvements.

3. The browser-based extension was successfully designed and implemented to automatically detect and filter comments containing bullying elements on the Instagram platform. The extension integrates 5 main features, namely comment extraction, comment submission to the backend for prediction processing, comment filtering based on prediction results, prediction interpretation using LIME, and manual controls for displaying or hiding comments. Furthermore, the implementation was successfully applied to 4 types of Instagram pages, namely pages accessed through direct links, profile pages, homepage feeds, and reels pages.

5.2. Suggestions

Based on the research results, several aspects can be further developed to improve model performance, interpretation quality, and system functionality. The recommendations are as follows:

1. Developing a detection model that not only analyzes comment content independently but also considers the context of the post, caption, and conversation surrounding the comment. This approach is expected to help the model understand the meaning of comments more comprehensively, allowing it to distinguish between comments that genuinely contain bullying elements and comments that only appear negative when analyzed without context.
2. Expanding the system capabilities so that it can detect bullying elements not only in text-based comments but also in comments containing GIFs. In addition, the system can be further developed to detect bullying elements in other forms of content, such as captions, messages, images, and videos. This development can broaden the scope of protection provided by the system because bullying behavior on social media is not always expressed through comments but may also appear in various other forms of content.
3. Adding support for more languages, including commonly used regional languages in social media interactions. This development can expand the applicability of the system in environments with high linguistic diversity and improve detection capabilities for comments written using mixed languages.