Abstract

Event extraction in low-resource languages, such as Indonesian, poses significant challenges due to limited annotated data, linguistic complexities, and class imbalances. This study aims to develop an effective and interpretable event extraction system tailored for Indonesian news articles related to sexual violence cases. Addressing sexual violence is critical due to its prevalence and severe societal impact in Indonesia, where cases are often publicly shared through anecdotal narratives or digital media, making systematic analysis challenging. To achieve this, the study proposes a Conditional Random Fields (CRF) model specialized for event extraction tasks. The model leverages feature-rich input, including lexical, contextual, and semantic features, to capture nuanced patterns and dependencies in the data. To address class imbalance, data augmentation techniques like synonym replacement were implemented. The CRF model was evaluated against baseline approaches, including a Rule-Based system and CNN-BiLSTM, and achieved a competitive F1-score of 0.730, highlighting the importance of feature engineering in low-resource settings. While data augmentation introduced variability, it often introduced noise, limiting its impact on overall performance. The findings emphasize the balance of efficiency and effectiveness offered by CRF in addressing challenges in low-resource NLP. Additionally, this study lays the groundwork for future research, supporting hybrid approaches integrating statistical models with advanced deep learning frameworks to enhance event extraction capabilities in underrepresented languages and domains.

Keywords: event extraction, conditional random fields, data augmentation, sexual violence