1. Introduction

Sexual violence remains a crucial issue in Indonesia, with a significant rise in reported cases over recent years. According to the Ministry of Women's Empowerment and Child Protection of the Republic of Indonesia (Kementerian PPPA), the number of reported sexual violence cases increased from 8,210 in 2020 to 13,156 in 2023 [1]. Other institutions also documented a substantial number of sexual violence cases in 2023. Lembaga Layanan reported 2,363 cases, accounting for 34.80% of the total reported cases, while the National Commission on Violence against Women (Komnas Perempuan) recorded 2,078 cases, representing 24.69% of the total [2]. These figures demonstrate the prevalence of sexual violence in Indonesia and the urgent need for systematic approaches to address this issue. This rise also corresponds with the increasing availability of digital media content, including online news and social media posts, which provide valuable insights into these cases. However, much of this data is presented in unstructured formats, such as lengthy articles and anecdotal narratives, making systematic analysis challenging. For instance, manual methods of extracting information from such data often result in inefficiencies, such as delays in trend analysis or inaccuracies in identifying key patterns, ultimately limiting the ability of policymakers and advocacy groups to respond effectively. These challenges highlight the need for automated solutions to improve the consistency and efficiency of data analysis.

Event extraction (EE), a subfield of natural language processing (NLP), aims to identify and structure event-related information from unstructured text [3]. While EE has shown success in various domains and languages, such as English for biomedical and legal text analysis [4], significant gaps remain in adapting these methods to Indonesian texts. The challenges are particularly evident in sensitive domains like sexual violence, where linguistic complexity, limited annotated data, and domain-specific nuances hinder the performance of existing models. For example, Indonesian texts often include cultural idioms, contextually rich narratives, and overlapping entities, which require specialized approaches for effective event extraction. This study addresses these gaps by focusing on developing an EE system specialized for Indonesian news articles reporting sexual violence cases. Specifically, the study employs Conditional Random Fields (CRF), a probabilistic model well-suited for sequence labeling tasks. By leveraging token-level features and domain-specific annotations, the CRF model aims to capture nuanced event details such as perpetrators, victims, locations, and legal actions.

This study offers three primary contributions. First, it introduces a domain-specific annotated corpus for sexual violence in Indonesia, which serves as a valuable resource for advancing NLP research in low-resource settings. Second, it proposes a specialized labeling scheme designed to capture the linguistic and cultural characteristics of Indonesian texts, including unique patterns of expression and contextual dependencies. Third, it presents a CRF-based framework, demonstrating its potential to support policymakers, advocacy groups, and researchers in analyzing trends and driving data-informed interventions [5]. Through these contributions, the study bridges the gap between existing methods and the unique challenges of event extraction in the Indonesian context, particularly for less explored domains such as sexual violence.