ABSTRACT

Crime Information Extraction is a task to extract important information from unstructured text in crime domain. Several prior studies on crime information extraction applied rules to perform crime information extraction. Defining rules manually, however, is a very daunting tasks and the performance is usually not quiet good, in which some misclassified instances could be found.

Most of crime information extraction researches has been performed in English dataset. There are few prior studies of crime information extraction conducted in Indonesian text, one of them used rule-based information extraction and applied Part-of-Speech (POS) tagging and Dependency parsing as features. Nonetheless, the performance of the prior study should be improved, especially on location and date/time extraction. The misclassification was caused by the system which could not determine the named-entity with the method proposed in the previous research.

This study proposes a system capable of extracting criminal information on Indonesian online news using a combination method of named-entity recognition and Support Vector Machine (SVM) to extract the location and time of the crime. After the location and time were extracted, crime type claasification was performed using SVM method. The proposed system outputs are: crime type, crime location, and crime date/ the time of the incident. The evaluation was conducted by comparing the extraction result of the proposed system against gold label extraction and the prior study result. The results show that the proposed system outperformed the prior system significantly. Despite its performance improvement, the results show the proposed method still needs to be improved in several areas, especially in the classification of Crime Scene sentences and also date and time format recognition in online news articles. The crime type classification has a F1-score of 92%, while the F1 score for Crime Location Extraction is 90.8%, with precision 92.5%, and recall 89.2%, and for Crime Date Extraction, the F1-score is 94.1%, with 100% precision, and 89% recall.