ABSTRACT

High public activity in discussing governor elections through social media generates large volumes of comment data, but these comments often use informal language, colloquial expressions, abbreviations, and are mixed with regional languages and local dialects that are difficult to understand. This hinders comment data processing for analysis or other purposes.

Manual normalization require significant time and resources, especially for large datasets. Manual normalization is also prone to inconsistencies and human errors. The continuously increasing volume of social media comment data makes manual normalization inefficient, thus requiring automation solutions.

An automatic text normalization system was developed using a Phrase-Based Statistical Machine Translation approach utilizing Moses framework. A parallel corpus dataset was built from 31,889 informal-formal sentence pairs, while the monolingual corpus consists of 1,613,381 sentences extracted from Wikipedia. The model was evaluated using BLEU metrics to measure the quality of normalization results.

The best model achieved a BLEU score of 82.16 on test data and 81.04 on validation data, successfully recognizing various informal language patterns such as non-standard abbreviations, repeated words with numbers, and slang terms. However, the system has limitations in handling Out-Of-Vocabulary terms.

Keywords: *text normalization, PBSMT, moses, KenLM, social media, governor election*