ABSTRACT
COMPARATIVE STUDY OF DATA PREPROCESSING METHODS FOR CODE-MIXED SENTIMENT ANALYSIS IN BAHASA INDONESIA

Ghifari Nugraha Pradana
20/457770/PA/19808

This research jumps into the importance of preprocessing in code-mixed sentiment analysis, finding the best type and combination of the method specifically focusing on the Indonesian-English code-mixed tweets. The Indonesian-English code mixed is commonly used for Indonesians in Twitter due to the era of globalization of the English language and to fully express what the Indonesians want to share. Due to the similarity of the structure between the two languages, both being Subject-Verb-Object (SVO) sentence structured.

This research aims to compare the various types of major preprocessing methods and combinations of preprocessing methods such as Language Identification, Lexical Normalization, Stop Word Removal, and Translation. There will be a total of 16 scenarios which consists of a combination of mixing and matching these 4 main preprocessing methods to find the best type and/or the best combination of these preprocessing methods for code-mixed Indonesian-English sentiment analysis. The research will go through all the scenarios with the same data, preprocessing cleaning method, same 80-20 data splitting and random state of 42, and the same machine learning model, which is SVM and RF, and a deep learning model LSTM.

All three models are capable of code-mixed sentiment analysis, but their performance and optimal preprocessing strategies vary; notably, approaches using Lexical Normalization and only using Translation generally boost performance, while combining threee or more methods often leads to inferior results.

Keywords: Sentiment Analysis, Code-Mixed Data, SVM, RF, LSTM

INTISARI

COMPARATIVE STUDY OF DATA PREPROCESSING METHODS FOR CODE-MIXED SENTIMENT ANALYSIS IN BAHASA INDONESIA

Ghifari Nugraha Pradana
20/457770/PA/19808

Code-Mixed Indonesia-Inggris biasa digunakan orang Indonesia di Twitter karena era globalisasi bahasa Inggris dan untuk mengungkapkan secara utuh apa yang ingin disampaikan oleh orang Indonesia. Karena kesamaan struktur antara kedua bahasa, keduanya merupakan struktur kalimat Subjek-Verba-Objek (SVO), pencampuran kode Indonesia-Inggris merupakan proses yang cukup gampang dilakukan.

Penelitian ini bertujuan untuk membandingkan berbagai jenis metode *preprocessing* utama dan kombinasi metode *preprocessing* seperti Identifikasi Bahasa, Normalisasi Leksikal, *Stop Word Removal*, dan Terjemahan. Akan ada total 16 skenario yang terdiri dari kombinasi campuran dan pencocokan 4 metode *preprocessing* utama ini untuk menemukan tipe terbaik dan/atau kombinasi terbaik dari metode *preprocessing* tersebut untuk analisis sentimen campuran kode Indonesia-Inggris. Penelitian akan melalui semua skenario dengan data yang sama, metode pembersihan preprocessing yang sama, *data splitting* yang sama, 80-20 split dengan random state=42, dan model pembelajaran mesin yang sama, yaitu Model SVM, RF, dan juga Model pemelajaran dalam LSTM.

Ketiga model mampu melakukan *code-mixed sentiment analysis*. Namun, hasil prapemrosesan optimal mereka sangat bervariasi; terutama, skenario yang menggunakan normalisasi leksikal dan terjemahan umumnya selalu meningkatkan kinerja code-mixed sentiment analysis, sementara penggabungan lebih dari tiga metode sering kali menghasilkan hasil yang lebih buruk.

Kata Kunci: Analisis Sentimen, Data Campuran *Code-Mixed*, SVM, RF, LSTM