

PERBANDINGAN KINERJA EKSTRAKSI FITUR TF-IDF & BOW DALAM SENTIMEN ANALISIS BPJS KESEHATAN DENGAN ALGORITMA XGBOOST

Oleh:

Made Donita Maharani, NIM 2115091065

Jurusan Teknik Informatika

Program Studi Sistem Informasi

ABSTRAK

BPJS Kesehatan merupakan salah satu layanan publik terbesar di Indonesia dan sering menjadi topik pembahasan masyarakat di platform media sosial. Berbagai opini yang muncul, baik berupa apresiasi maupun keluhan, mencerminkan persepsi pengguna terhadap kualitas layanan yang diberikan. Oleh karena itu, analisis sentimen diperlukan untuk memahami kecenderungan opini masyarakat tersebut. Penelitian ini bertujuan untuk melakukan klasifikasi sentimen terhadap opini pengguna BPJS Kesehatan yang diambil dari platform X (Twitter), serta membandingkan kinerja dua metode ekstraksi fitur, yaitu *Term Frequency-Inverse Document Frequency* (TF-IDF) dan *Bag of Words* (BoW) menggunakan algoritma XGBoost. Data dikumpulkan melalui proses *scraping data*, dilanjutkan dengan pelabelan manual oleh validator ahli, serta *preprocessing*. Pembobotan kata dilakukan menggunakan TF-IDF dan BoW sebelum data dianalisis menggunakan model XGBoost pada tujuh skenario pengujian. Hasil pengujian menunjukkan bahwa metode TF-IDF unggul pada metrik akurasi dan presisi dengan nilai masing-masing 78,63% dan 70,30%, sedangkan BoW lebih unggul pada recall dan F1-score dengan nilai 70,23% dan 69,14%. Hal ini menunjukkan bahwa TF-IDF lebih efektif dalam meningkatkan ketepatan prediksi melalui pembobotan kata yang selektif, sementara BoW lebih mampu menangkap variasi data sentimen secara lebih luas. Dengan demikian, kedua metode memiliki keunggulan masing-masing sesuai dengan karakteristik evaluasi yang digunakan.

Kata Kunci: Analisis sentimen, BPJS Kesehatan, TF-IDF, BoW, XGBoost

**COMPARISON OF TF-IDF AND BOW FEATURE EXTRACTION
PERFORMANCE IN SENTIMENT ANALYSIS OF BPJS KESEHATAN
USING THE XGBOOST ALGORITHM**

By

Made Donita Maharani, NIM 2115091065

Informatic Engineering

Information System Study Program

ABSTRACT

BPJS Kesehatan is one of the largest public health insurance services in Indonesia and frequently becomes a topic of public discussion on social media platforms. The various opinions expressed, including both appreciation and complaints, reflect users' perceptions of the quality of services provided. Therefore, sentiment analysis is required to understand the tendency of public opinion. This study aims to classify user sentiment toward BPJS Kesehatan based on opinions collected from the X (Twitter) platform and to compare the performance of two feature extraction methods, namely Term Frequency-Inverse Document Frequency (TF-IDF) and Bag of Words (BoW), using the XGBoost algorithm. The data were collected through a data scraping process, followed by manual labeling by expert validators and preprocessing. Word weighting was performed using TF-IDF and BoW before the data were analyzed using the XGBoost model across seven testing scenarios. The experimental results show that the TF-IDF method outperforms BoW in terms of accuracy and precision, achieving values of 78.63% and 70.30%, respectively, while BoW performs better in recall and F1-score with values of 70.23% and 69.14%. These findings indicate that TF-IDF is more effective in improving prediction accuracy through selective word weighting, whereas BoW is better at capturing a wider variation of sentiment data. Thus, both methods have their respective advantages depending on the evaluation characteristics applied.

Keywords: *Sentiment analysis, BPJS Kesehatan, TF-IDF, BoW, XGBoost*