

INTISARI

KOMPARASI METODE-METODE ANALISIS SENTIMEN DAN EKSTRAKSI FITUR PADA DATASET ULASAN ONLINE HOTEL

Kevin Susilo

(18/427583/PA/18543)

Ulasan-ulasan tentang suatu hotel menjadi pertimbangan bagi wisatawan atau calon pelanggan dalam memilih hotel yang tepat. *TripAdvisor* merupakan salah satu laman web di mana wisatawan dapat memesan kamar hotel secara *online*, yang memiliki jumlah ulasan-ulasan sebesar satu miliar ulasan. Banyaknya jumlah ulasan tersebut mengakibatkan perlunya metode-metode yang lebih cepat dan tidak menguras tenaga untuk menelaah ulasan-ulasan tersebut, salah satunya dengan melakukan analisis sentimen.

Di dalam penelitian ini, dilakukan analisis sentimen pada data ulasan hotel yang di-*scrape* dari laman *TripAdvisor*, dengan menggunakan *Naïve Bayes*, *Decision Tree*, *Random Forest*, dan *K-Nearest Neighbors*. Teknik *oversampling* SMOTE digunakan untuk menangani jumlah data tiap kelas yang tidak seimbang pada *dataset*. Metode ekstraksi fitur *Bag-of-Words* dan *TF-IDF* digunakan dan diuji, untuk mendapatkan kombinasi metode analisis sentimen dan ekstraksi fitur yang menghasilkan kinerja terbaik. Kinerja kombinasi tersebut dievaluasi menggunakan *k-fold cross validation* dan beberapa metrik, yaitu *accuracy*, *precision*, *recall*, dan *f1-score*.

Kinerja terbaik, berdasarkan model, dihasilkan oleh *Naïve Bayes*, dengan *accuracy* sebesar 54.87%, *precision* sebesar 49.73%, *recall* sebesar 51.75%, dan *f1-score* sebesar 50.25%. Kinerja yang dapat terbilang kurang memuaskan tersebut disebabkan oleh beberapa hal. Penggunaan SMOTE untuk memproses data untuk KNN adalah salah satunya, sementara yang lainnya adalah keserupaan distribusi frekuensi kemunculan kata pada kelas-kelas yang hampir serupa. Agregasi kelas-kelas dengan distribusi frekuensi kemunculan kata yang serupa, dan tidak menggunakan SMOTE untuk KNN, direkomendasikan, beserta dengan dipergunakannya model dan metode ekstraksi fitur yang lebih *robust* untuk penelitian di masa depan.

Kata Kunci: Analisis Sentimen, *Naïve Bayes*, *Decision Tree*, *Random Forest*, *K-Nearest Neighbors*, *Bag-of-Words*, *TF-IDF*, *TripAdvisor*

ABSTRACT

COMPARISON OF SENTIMENT ANALYSIS AND FEATURE EXTRACTION METHODS ON ONLINE HOTEL REVIEW DATASET

Kevin Susilo
(18/427583/PA/18543)

Reviews regarding a hotel are a consideration for potential tourists or customers when choosing the right hotel for them. TripAdvisor is one of the web pages where tourists can rent a hotel room online, and it has accumulated over one billion reviews. The large number of reviews available causes the need for faster, less taxing methods to swiftly gather insight from said reviews. One such method is sentiment analysis.

In this research, sentiment analysis is done on hotel review data that was scraped from the TripAdvisor web page, using Naïve Bayes, Decision Tree, Random Forest, and K-Nearest Neighbors methods. The oversampling technique, SMOTE, is used to handle the severely imbalanced data between the classes in the dataset. The feature extraction methods Bag-of-Words and TF-IDF are used and tested to find a combination of sentiment analysis and feature extraction methods that will yield the best performance. Said performance is evaluated using k-fold cross-validation and a few metrics, which are accuracy, precision, recall, and f1-score.

The best performance, model-wise, is produced by Naïve Bayes, with an accuracy of 54.87%, precision of 49.73%, recall of 51.75%, and f1-score of 50.25%. The performance, which could be said to be less than satisfactory, stems from a few issues. The usage of SMOTE to process data for KNN is one of them, while the other is the similarity of frequency distributions of word occurrences on similar classes. Aggregating classes with similar frequency distributions of word occurrences, and refraining from using SMOTE for KNN, are recommended, alongside with using other, more robust models and feature extraction methods for future research.

Keywords: Sentiment Analysis, Naïve Bayes, Decision Tree, Random Forest, K-Nearest Neighbors, Bag-of-Words, TF-IDF, TripAdvisor