

ABSTRAK

Adanya persepsi masyarakat luar terhadap orang Madura yang tersebar di berbagai *platform* media sosial menjadi tantangan dalam mengidentifikasi dan mengukur opini tersebut terlebih secara manual. Selain itu, kesalahan ejaan yang sering muncul dapat memengaruhi tingkat akurasi. Penggunaan teknologi Teknologi *Natural Language Processing* dengan kombinasi *Naïve Bayes Classifier*, *Damerau-Levenshtein Distance* untuk koreksi ejaan, serta fitur *N-Gram* dan pembobotan TF-IDF akan dipakai untuk meningkatkan akurasi analisis sentimen. Data 766 komentar dari Instagram, TikTok, dan YouTube, yang memuat opini tentang masyarakat Madura, dibedakan ke dalam dua kelas dalam proses manual labelling, menjadi 383 positif dan 383 negatif. Setelah melalui proses *preprocessing*, perbaikan ejaan dengan *Damerau-Levenshtein Distance*, ekstraksi fitur *N-Gram* dan pembobotan TF-IDF, data dibagi menjadi 80% data latih dan 20% data uji. Pengujian dilakukan menggunakan 5 parameter *alpha* dari 1.0 hingga 3.0 dengan empat skenario yakni model dasar *Naïve Bayes Classifier*, dengan koreksi ejaan *Damerau-Levenshtein Distance*, serta kombinasi fitur *bigram* dan *trigram*. Berdasarkan hasil pengujian menggunakan *confusion matrix* dan *K-Fold cross validation* $K=5$, akurasi tertinggi terdapat pada kombinasi *Naïve Bayes Classifier*, *Damerau-Levenshtein Distance*, dan *N-Gram* dengan $n=2$ atau *bigram* yang memiliki peningkatan 1.31% dari model dasar *Naïve Bayes* dengan akurasi terendah yakni 81.70% menjadi 83.01%.

Kata Kunci: Analisis Sentimen, Madura, *Naïve Bayes Classifier*, *Damerau-Levenshtein Distance*, *N-Gram*

ABSTRACT

The existence of outside perceptions of Madurese spread across various social media platforms is a challenge in identifying and measuring these opinions, especially manually. In addition, spelling errors that often appear can affect the level of accuracy. The use of Natural Language Processing technology with a combination of Naïve Bayes Classifier, Damerau-Levenshtein Distance for spelling correction, and N-Gram features and TF-IDF weighting will be used to improve the accuracy of sentiment analysis. The data of 766 comments from Instagram, TikTok, and YouTube, which contain opinions about Madurese people, are divided into two classes in the manual labelling process, into 383 positive and 383 negative. After going through preprocessing, spelling correction with Damerau-Levenshtein Distance, N-Gram feature extraction and TF-IDF weighting, the data was divided into 80% training data and 20% test data. Tests were conducted using 5 alpha parameters from 1.0 to 3.0 with four scenarios, namely the basic Naïve Bayes Classifier model, with Damerau-Levenshtein Distance spelling correction, and a combination of bigram and trigram features. Based on the test results using confusion matrix and K-Fold cross validation K=5, the highest accuracy is found in the combination of Naïve Bayes Classifier, Damerau-Levenshtein Distance, and N-Gram with n=2 or bigram which has an increase of 1.31% from the basic Naïve Bayes model with the lowest accuracy of 81.70% to 83.01%.

Keywords: Sentiment Analysis, Madura, Naïve Bayes Classifier, Damerau-Levenshtein Distance, N-Gram