ABSTRAK

Tahapan *preprocessing* merupakan langkah yang sangat penting dalam analisis sentimen, yaitu berfungsi untuk mengubah data yang tidak terstruktur menjadi data yang terstruktur agar data siap diolah. Di dalam *preprocessing* terdapat beberapa langkah penting, salah satunya adalah *word normalization* yaitu proses mengubah kata tidak baku atau non standar menjadi kata baku sesuai dengan Kamus Besar Bahasa Indonesia. Pada umumnya analisis sentimen tidak melalui tahapan *word normalization* sehingga terjadi ambiguitas dan sistem tidak dapat mengklasifikasikan kelasnya dengan baik sehingga dapat mempengaruhi akurasi. Data yang telah melewati tahapan *preprocessing* selanjutnya akan diklasifikasikan menggunakan metode *machine learning.*

Terdapat beragam metode *machine learning* yang dapat digunakan untuk klasifikasi pada analisis sentimen, contohnya adalah metode *Naïve Bayes Classifier* dan metode *K-Nearest Neighbor.* Kedua metode tersebut dibandingkan karena merupakan metode yang sederhana, mudah diimplementasikan dan menghasilkan akurasi yang relatif tinggi. Penelitian ini bertujuan untuk mencari perbandingan pengaruh *word normalization* pada metode *Naïve Bayes Classifier* dan metode *K-Nearest Neighbor.*

Data yang digunakan sebanyak 1050 data berbahasa Indonesia yang di dapatkan dari media sosial Twitter dan Instagram pada tahun 2021 hingga 2022, dengan kata kunci dan postingan tentang BPJS Kesehatan, yang kemudian dibagi menjadi 840 data latih dan 210 data uji. Hasil pengujian menunjukkan bahwa dengan melakukan tahapan *word normalization* pada metode *Naïve Bayes Classifier* dan *K-Nearest Neighbor* menghasilkan akurasi yang lebih tinggi dibandingkan tanpa melakukan tahapan *word normalization.* Dan metode *Naïve Bayes Classifier* baik dengan skenario I maupun skenario II memiliki akurasi yang lebih unggul dibandingan metode *K-Nearest Neighbor* baik skenario I maupun skenario II. Pada pengujian model dengan data sejumlah 1050, akurasi yang didapatkan oleh metode *Naïve Bayes Classifier* dengan skenario I yaitu sebesar 87,14%. Sementara akurasi yang didapatkan oleh metode *Naïve Bayes Classifier* dengan skenario II yaitu sebesar 86,67%. Sedangkan untuk akurasi yang didapatkan Dari metode *K-Nearest Neighbor* skenario I sebesar 80,48% dan skenario II dihasilkan akurasi sebesar 77,14%.

Kata kunci: analisis sentimen, ujaran bpjs kesehatan, Twitter, Instagram, *word normalization*, *Naïve Bayes Classifier*, *K-Nearest Neighbor*

ABSTRACT

The preprocessing stage is a very important step in sentiment analysis, which functions to convert unstructured data into structured data so that the data is ready to be processed. In preprocessing there are several important steps, one of which is word normalization, which is the process of converting non-standard or non-standard words into standard words according to the Big Indonesian Dictionary. In general, sentiment analysis does not go through the word normalization stage so that ambiguity occurs and the system cannot classify the class properly so that it can affect accuracy. Data that has passed the preprocessing stage will then be classified using machine learning methods..

There are various machine learning methods that can be used for classification in sentiment analysis, such as the Naïve Bayes Classifier and the K-Nearest Neighbor method. Both methods are compared because they are simple, easy to implement and produce relatively high accuracy. This research aims to find a comparison of the effect of word normalization on the Naïve Bayes Classifier method and the K-Nearest Neighbor method.

The data used is 1050 Indonesian-language data obtained from Twitter and Instagram social media from 2021 to 2022, with keywords and posts about BPJS Kesehatan, which are then divided into 840 training data and 210 test data. The test results show that by performing the word normalization stage in the Naïve Bayes Classifier and K-Nearest Neighbor methods produces higher accuracy than without performing the word normalization stage. And the Naïve Bayes Classifier method both with scenario I and scenario II has superior accuracy compared to the K-Nearest Neighbor method both scenario I and scenario II. In model testing with 1050 data, the accuracy obtained by the Naïve Bayes Classifier method with scenario I is 87.14%. While the accuracy obtained by the Naïve Bayes Classifier method with scenario II is 86.67%. Meanwhile, the accuracy obtained from the K-Nearest Neighbor method in scenario I was 80.48% and scenario II resulted in an accuracy of 77.14%.

Keywords: sentiment analysis, bpjs kesehatan, Twitter, Instagram, word normalization, Naïve Bayes Classifier, K-Nearest Neighbor