INTISARI

Penggunaan Complement Naive Bayes untuk Memecahkan masalah *Imbalance Dataset* pada Twitter dengan studi kasus Jasa Transportasi Online

Oleh

Kevin Goldwin
16/394088/PA/17179

Twitter adalah sosial media dimana penggunanya dapat menulis berbagai topik dan berdiskusi masalah yang terjadi. Masalah yang muncul saat melakukan analisis sentimen menggunakan data twitter adalah *imbalance dataset*. *Imbalance dataset* dapat menimbulkan risiko kesalahan klasifikasi sehingga algoritma pengklasifikasi tidak bekerja dengan optimal.

Penelitian ini menggunakan algoritma klasifikasi *Complement Naive Bayes* (CNB) untuk mengatasi *imbalance dataset* pada data twitter. Penggunaan *Multinomial Naive Bayes* (MNB) akan dilakukan sebagai klasifikasi pembanding. Pengujian dilakukan pada dataset berjumlah 3268 tweet (24% positif dan 76% negatif serta netral) menggunakan seleksi fitur chi square 30% dan 50% fitur terbaik.

Hasil Penelitian CNB dengan seleksi fitur chi square 30% berhasil mendapatkan performa lebih baik dibandingkan dengan seleksi fitur chi square 50% dengan akurasi sebesar 88% presisi sebesar 61%, recall sebesar 72% dan F1-Score sebesar 74%. CNB memiliki nilai recall dan F1-Score yang lebih baik dibandingkan MNB pada seleksi fitur chi square 30% dan 50% fitur terbaik. Hasil Pengujian pada MNB dengan chi square 50% mendapatkan akurasi dan presisi yang lebih baik tetapi mendapatkan nilai recall dan F1-score yang lebih rendah dibandingkan dengan CNB.

Kata kunci : Twitter, Analisis Sentimen, Imbalance Dataset, Complementary Naive Bayes, Multinomial Naive Bayes

ABSTRACT

Using Complement Naive Bayes To Solve The Imbalance Dataset Problem In Twitter With Case Study Of Online Transportation Services

By

Kevin Goldwin
16/394088/PA/17179

Twitter is a social media platform where users can discuss and write about various topics. Sentiment analysis using Twitter dataset usually suffers from imbalanced dataset problem. Imbalanced dataset can increase the risk of wrong classifications, causing the classifier model to perform less optimal.

This research uses Complement Naive Bayes (CNB) classification algorithm for resolving imbalanced dataset in Twitter data. Multinomial Naive Bayes is also used for classification comparison. The test is done on the dataset of 3268 tweets (24% positive and 76% negative and neutral) using feature selection of chi-square 30% and 50% for selecting best features.

Result for CNB and chi-square 30% have the better performance compared to chi-square 50% with accuracy of 88%, precision of 61%, recall of 72%, and F1-score of 74%. CNB has better recall and F1-score using feature selection of chi-square 30% and 50% for selecting best features. Result for MNB and chi-square 50% have the better performance in accuracy and precision but worse in recall and F1-score value than CNB.

Keywords : Twitter, Sentiment Analysis, Imbalance Dataset, Complementary Naive Bayes, Multinomial Naive Bayes