INTISARI

PERBANDINGAN EFISIENSI ALGORITMA TWCNB DAN K-NEAREST NEIGHBOR PADA KLASIFIKASI SENTIMEN TWITTER

FEBRIAL WINARTA PRATAMA

12/334687/PA/14920

"Apa yang orang lain pikirkan" telah menjadi sesuatu yang penting untuk menjadi pertimbangan dalam pengambilan keputusan yang biasa diutarakan dalam bentuk opini atau sentimen. Adanya internet dan media sosial Twitter mendukung ketersediaan data dalam jumlah besar. Hal ini menjadi faktor pendukung adanya penelitian di bidang analisis sentimen dan klasifikasi teks. Berbagai penelitian dengan berbagai kasus klasifikasi dan implementasi algoritma telah dilakukan, seperti algoritma *Transformed Weight-normalized Naïve-Bayes (TWCNB)* dan *K-Nearest Neighbor (KNN)* telah dikembangkan dalam domain penelitian sama yaitu klasifikasi teks. Namun di antara kedua algoritma tersebut belum diketahui secara pasti mana yang memiliki performa yang lebih baik dalam pengklasifikasian sentimen.

Penelitian ini mengimplementasikan algoritma TWCNB dan KNN dalam mengklasifikasi teks *tweet* berbahasa Indonesia pada suatu topik yang ditentukan. *Tweet* diekstraksi sebagai data mentah dilewatkan pada berbagai *preprocessing* dan digunakan sebagai fitur pada pemodelan dua algoritma tersebut. Skenario pengujian yang dilakukan meliputi pengaruh peningkatan jumlah data dan keseimbangan persebaran data pada performa pengklasifikasian. Performa diukur dari segi akurasi, *f-measure*, dan waktu pengujian.

Kedua skenario pengujian memiliki pengaruh pada perfoma klasifikasi kedua algoritma. Adanya peningkatan akurasi, *f-measure*, dan waktu pengujian seiring dengan bertambahnya jumlah data. Juga pada skenario perbedaan persebaran data di mana klasifikasi memiliki akurasi dan waktu pengujian lebih baik pada *dataset* yang tidak seimbang dibandingkan *dataset* yang seimbang, namun memberikan *f-measure* yang kurang baik. Hasil evaluasi performa pengklasifikasian pada penelitian ini menyimpulkan bahwa algoritma KNN memiliki performa pengklasifikasian lebih baik dibandingkan algoritma TWCNB. Antara algoritma TWCNB dan KNN memiliki gap akurasi dan *f-measure* rata-rata sebesar 1,15% dan 2,22% namun TWCNB memiliki waktu pengujian 2-3 kali lipat dari algoritma KNN seiring meningkatnya jumlah *dataset* pengujian.

Kata kunci: algoritma TWCNB, algoritma KNN, klasifikasi teks, analisis sentimen

ABSTRACT

EFFICENCY COMPARISON BETWEEN TWCNB AND K-NEAREST NEIGHBOR ALGORITHM ON TWITTER SENTIMENT CLASSIFICATION

FEBRIAL WINARTA PRATAMA

12/334687/PA/14920

"What the other people think" is an important thing for consideration in decision making which is usually expressed in the form of opinion or sentiment. The internet and social media such Twitter provide data with great availability. These factors support many researches in field of sentiment analysis and text classification. Several researches with any kind of case and implementation of algorithm, like Transformed Weight-normalized Naïve-Bayes (TWCNB) and K-Nearest Neighbor (KNN) that had been developed beforehand in same research domain. However it is not certain yet which one of them that has better performance in sentiment classification.

This research implements TWCNB and KNN algorithm in classifying Indonesian tweets on decided topic. The extracted tweets was processed through some preprocessing procedure and used as modeling feature of algorithms. Testing scenario used in this research divide into two scenarios. First scenario is testing on different amount of dataset and second scenario is testing on different dataset class proportion. Classification performance is measured by three metrics. Those metrics are accuracy, f-measure, and testing time.

These scenarios definitely give different effect on both algorithm classification performance. Accuracy, f-measure value, and testing time increase along with increase size of dataset. Also in second scenario which results better accuracy and testing time on imbalance dataset than balanced dataset., but results worse f-measure value. This research performance evalution concludes that KNN algorithm has better classification performance than TWCNB algorithm. There are gaps between TWCNB and KNN accuracy about 1.12% in average and f-measure about 2.22% in average, but TWCNB has testing time which is relatively two to three times longer than testing time of KNN algorithm along with the increase of dataset size.

Keywords: TWCNB algorithm, KNN algorithm, text classification, sentiment analysis