INTISARI

Analisis sentimen untuk mengetahui reputasi suatu produk melalui media sosial terutama Twitter telah banyak dilakukan. Pada umumnya, analisis dilakukan dengan menggunakan seluruh *tweet* sebagai data latih. Padahal beberapa aspek dari kalimat seperti panjang kalimat, konten, subyektifitas dan posisi sebuah kata yang mengandung opini dapat mempengaruhi nilai akurasi dari sebuah model klasifikasi. Tujuan dari penelitian ini adalah untuk mengetahui pengaruh adanya batas minimal dan maksimal jumlah kata per *tweet* pada data latih dalam pembuatan model analisis sentimen terhadap operator telekomunikasi serta merancang sistem tersebut dengan menggunakan mode dengan performa terbaik.

Penelitian ini menggunakan data latih berupa *tweet* sejumlah 4927 yang telah melalui proses pelabelan secara manual dan menghilangkan *tweet* dengan sentimen normal. Nilai akurasi, presisi, *recall* dan *F-measure* dari model yang dibangun dengan data latih dihitung dengan Weka *Experimenter* untuk empat buah algoritme yaitu *Naïve Bayes, Support vector machine, K-Nearest Neighbor dan Decission Tree.* Data latih dengan performa terbaik digunakan untuk membangun model klasifikasi dengan Weka *Explorer*. Model klasifikasi kemudian digunakan untuk mengembangan aplikasi monitoring kualitas layanan telekomunikasi berdasarkan analisis sentimen pada *Twitter* secara waktu nyata.

Hasil penelitian menunjukkan bahwa pada data tidak berimbang, pembatasan maksimal jumlah kata per *tweet* dapat meningkatkan akurasi secara signifikan untuk setiap algoritme, namun nilai akurasi tersebut tidak menggambarkan performa dari algoritme dikarenakan semakin sedikit batas maksimal kata, data latih semakin tidak berimbang. Sementara itu pada data berimbang, akurasi model meningkat untuk semua algoritme meskipun tidak signifikan. Peningkatan akurasi tertinggi terdapat pada Algoritme KNN yaitu sebesar 12.04%, disusul Algoritme NB dengan 2.59%, SVM sebesar 2.17% dan DT sebesar 2.04%. Secara umum Algoritme SVM masih memberikan nilai akurasi tertinggi yaitu 87,10%. Di sisi lain, pemberian batas minimal jumlah kata per *tweet* tidak berpengaruh hanya mingkatkan akurasi jika data training terdiri lebih dari satu kalimat. Untuk data training yang hanya terdiri dari satu kalimat, pemberian batas minimum jumlah kata per tweet justru akan mengurangi nilai akurasi. Aplikasi monitoring kualitas jaringan telekomunikasi secara waktu nyata dibangun dengan model terbaik yang dilatih menggunakan Algoritme SVM dengan data latih berimbang dengan batas maksimal kata pada *tweet* sebanyak 18 dan jumlah kata minimal sebanyak 1 yang merupakan model dengan performa terbaik.

Kata Kunci: Analisis Sentimen, Twitter, Peningkatan Akurasi, Algoritme

ABSTRACT

For the last decade, it's very common to use sentiment analysis to dive into brand perception on social media, particularly Twitter. Usually the model for this analysis is built from all Twitter data acquired from the query, meanwhile several aspects of the sentence such as length, purity, subjectivity and position within the opinionated text result to different model accuracy. This research is aimed to find out the effect of limitation on number of word per tweet towards model accuracy when classifying sentiment on telecommunication provider Twitter mention in Indonesia.

4927 of training data in a form of tweet was used in this research after being manually labelled with positive and negative, and cleaned from unopiniated data. Accuracy, precision, recall and F-measure of all the models from four algorithms (Naïve Bayes, Support vector machine, K-Nearest Neighbor and Decision Tree) built from the dataset is then calculated and compared using Weka Experimenter. The best dataset is later used to build a model for a real time classification tools to calculate and visualize sentiment on telecommunication provider in Indonesia.

The research shows that in unbalance data, the limitation on maximum number of word per tweet can significantly increase model accuracy. However this accuracy can't precisely illustrate performance of the model. This due to the fact that the smaller number of limitation the more unbalance the data is. On the other hand, in balance data, the limitation on maximum number of word per tweet can also increase the accuracy of the model, although it's not as significant as the accuracy acquired from the unbalance data. However this accuracy is more valid then the unbalance data and can illustrate model performance very well. The biggest increase in accuracy found in KNN Algorithm with 12,04% increment, followed by NB with 2,59% increment, SVM with 2,17% increment and DT with 2,04% increment. Overall SVM still result in higher performance compared to other algorithms with 87,10% of accuracy. Meanwhile the limitation of minimum number of words per tweet can only increase the accuracy when the training data is consist of more than two sentences, usually above 15 words per tweet. On the training data that only consist of one sentence, the limitation on minimum words per tweet will only result in decrement of the model accuracy. The real time classification tools for sentiment analysis on telecommunication provider is then built using SVM Algorithm with balance training data which is filtered to maximum of 18 words per tweet and minimum of 1 words per tweet, that is considered to have the best performance compared to another model.

Keywords: Sentiment Analysis, Twitter, Increase in Accuracy, Algorithm