

ABSTRACT

Indonesian stopwords that are available today have not given positive impact in sentiment analysis accuracy because the word list has not been made especially for sentiment analysis. Chi-square in this case could be a method for making stopwords by sorting the lowest value of it. The trend of sentiment analysis grows to handle Big Data such as social media Twitter data as a corpus. Twitter data are real-time data that keeps increasing every second, but the past researches about sentiment analysis have not study about the development of a real-time sentiment analysis system, that covers the data crawling form Twitter Streaming API, preprocessing, classification until visualization.

This research aims to develop a stopwords maker system using the lowest chi-square, test the performance with SVM and Naïve Bayes classifier using produced stopwords, and then compared with other Indonesian stopwords available and by using no stopwords at all. The object used in this research is customer satisfactory of 7 mobile telecommunication operator in Indonesia.

The result of this research shows that using lowest chi-square features as stopwords in Twitter sentiment analysis gave positive impact on accuracy compared with using no stopwords at all; especially with Naïve Bayes Classifier. The accuracy of Naïve Bayes increased 1.9% with produced stopwords using 1000 lowest features, but on SVM stopwords usage is not giving meaningful impact on accuracy, presicion, neither recall. Nevertheless, SVM classification algorithm always giving accuracy above 99%, by using stopwords or not. According to this result, in this research a real-time system of Twitter sentiment analysis for customer satisfactory of mobile telecommunication operator in Indonesia with SVM has been built without stopwords dan stemmer.

Keywords– Real-Time Sentiment Analysis, Stopwords, Twitter

INTISARI

Stopwords Bahasa Indonesia yang tersedia pada saat ini belum mampu memberikan pengaruh positif pada akurasi dalam hasil analisis sentimen karena daftar kata yang tersedia di dalamnya belum dikhususkan untuk analisis sentimen. *Chi-square* dalam hal ini dapat digunakan sebagai metode untuk membuat *stopwords* berdasarkan nilai terendahnya. Tren analisis sentimen berkembang untuk menangani Big Data, seperti menggunakan data Twitter sebagai *corpus*. Data Twitter adalah data *real-time* yang terus bertambah setiap detik, namun dalam penelitian-penelitian analisis sentimen sebelumnya belum diteliti mengenai pembangunan analisis sentimen otomatis secara *real-time* yang meliputi pengambilan data Twitter Streaming API, *preprocessing*, klasifikasi sentimen hingga visualisasi.

Penelitian ini bertujuan untuk membangun sistem pembuatan *stopwords* dengan memanfaatkan *chi-square* terendah, menguji performanya dengan algoritme klasifikasi SVM dan Naïve Bayes, kemudian membangun sistem *real-time* berdasarkan sistem dengan performa terbaik. Objek penelitian adalah kepuasan pelanggan 7 operator telekomunikasi seluler di Indonesia.

Hasil penelitian menunjukkan bahwa penggunaan *chi-square* terendah sebagai *stopwords* berhasil memberikan dampak positif pada akurasi terutama pada *classifier* Naïve Bayes. Akurasi Naïve Bayes meningkat hingga 1,9% dengan menggunakan *stopwords chi-square* dengan 1000 fitur terendah, namun penggunaan *stopwords* pada SVM tidak memberikan pengaruh yang berarti baik pada akurasi, presisi maupun *recall*. Meski demikian, algoritme klasifikasi SVM memberikan akurasi di atas 99% baik menggunakan *stopwords* maupun tidak menggunakan *stopwords*. Berdasarkan hasil pengujian tersebut, dalam penelitian ini juga telah dibangun sebuah sistem analisis sentimen Twitter *real-time* untuk kepuasan pelanggan operator telekomunikasi seluler di Indonesia dengan algoritme SVM tanpa *stopwords* dan *stemmer*.

Kata kunci – Analisis Sentimen *Real-Time*, *Stopwords*, Twitter