ABSTRACT

The sentence classification is one of the major research fields in the natural language processing. Especially sentiment analysis is one of the well-known application of text classification research. Sentiment analysis (SA) or opinion mining is analysis of polarity of a given sentence, e.g., negative, neutral or positive. It is one of the active research areas in Natural Language Processing (NLP). Machine learning based sentiment analysis have been proven to be successful in finding people's opinion in various products and services. However, in many cases the data being used to train such machine learning models could be highly imbalanced. In order to overcome this problem, data augmentation is introduced. Data augmentation is a sampling algorithm to create diverse data representations and address class imbalance in training datasets.

This thesis combines *Naïve Bayes* and ADASYN to identify classes of imbalanced data and leverage existing data to address the problem. We have evaluated the model performance on both original and augmented data, and found that the augmented data has higher accuracy compared to the original data. Using ADASYN with parameter `n_neighbors = 7` increases model accuracy from 79% to 84%. The proposed data augmentation approach seems to be very useful to address the class imbalance problem.

Keywords : ADASYN, data augmentation, natural language processing, sentiment analysis, *Naïve Bayes*.

INTISARI

Klasifikasi teks adalah salah satu bidang penelitian utama dalam *natural language processing*, khususnya *sentiment analysis* yang merupakan salah satu penerapan penelitian klasifikasi teks yang terkenal. *Sentiment Analysis* (SA) atau penambangan opini adalah analisis polaritas suatu kalimat, misalnya negatif, netral, atau positif. *Sentiment analysis* berbasis *machine learning* terbukti berhasil menemukan opini masyarakat terhadap berbagai produk dan layanan. Namun, dalam banyak kasus, data yang digunakan untuk melatih model *machine learning* tersebut bisa jadi sangat tidak seimbang. Untuk mengatasi masalah ini, augmentasi data diperkenalkan. Augmentasi data adalah algoritme pengambilan sampel untuk menciptakan representasi data yang beragam dan mengatasi ketidakseimbangan kelas dalam set data pelatihan.

Tesis ini menggabungkan pengklasifikasi *Naïve Bayes* dan ADASYN untuk mengidentifikasi kelas data yang tidak seimbang dan memanfaatkan data yang sudah ada untuk mengatasi masalah tersebut. Evaluasi performa model pada data asli dan data yang diaugmentasi menunjukkan bahwa data yang diaugmentasi memiliki akurasi yang lebih tinggi dibandingkan dengan data asli. Penggunaan ADASYN dengan parameter `n_neighbors = 7` meningkatkan akurasi model dari 79% menjadi 84%. Pendekatan augmentasi data yang diusulkan tampaknya sangat berguna untuk mengatasi masalah ketidakseimbangan kelas.

Kata Kunci : ADASYN, augmentasi data, *natural language processing*, analisis sentiment, *Naïve Bayes*.