TITLE

KLASIFIKASI DOKUMEN BERITA KECELAKAAN TRANSPORTASI BERBAHASA INDONESIA MENGGUNAKAN METODE SUPPORT VECTOR MACHINE

ABSTRACT

ABSTRAK
Menurut data Badan Inteligen Negara (BIN), kecelakaan transportasi di Indonesia merupakan pembunuh terbesar ketiga setelah penyakit jantung koroner dan Tuberculosis. Tingginya tingkat kecelakaan tersebut membuat informasi terkait topik kecelakaan transportasi sering ditemukan pada portal berita online berbahasa Indonesia. Saat ini, penentuan kategori berita yang dipublikasi masih dilakukan secara manual, sehingga diperlukan sistem yang dapat mengklasifikasikan berita secara otomatis. Pada penelitian ini, metode klasifikasi dalam teks mining akan diterapkan untuk mengolah dan menganalisis data web bertopik kecelakaan transportasi. Metode klasifikasi yang digunakan adalah Support Vector Machine (SVM) dengan pendekatan one-against-all, one-against-one dan Directed Acyclic Graph Support Vector Machine (DAGSVM). Ketiga pendekatan tersebut akan dibandingkan dalam proses klasifikasi, pendekatan terbaik akan ditentukan berdasarkan nilai f-measure yang dihasilkan mendekati 1. Klasifikasi dilakukan dengan mengkategorikan data web menjadi empat kategori yaitu kecelakaan darat, kecelakaan laut, kecelakaan udara dan lainnya. Tahapan penelitian ini terdiri dari pengumpulan data, pembersihan data, pembuatan kamus n-gram, pembangkitan fitur, pembangunan model dan klasifikasi. Data pembelajaran yang digunakan berjumlah 10.948 halaman web dan data pengujian berjumlah 550 halaman web yang terdiri dari 500 data berlabel dan 50 data tidak berlabel. Terdapat tiga pola fitur yang masing-masing dibangkitkan berdasarkan dua jenis kamus yaitu kamus tanpa stopwords dan kamus menggunakan stopwords sehingga dihasilkan 6 dataset yang berbeda. Hasil klasifikasi pada pengujian model menunjukkan bahwa pendekatan DAGSVM merupakan pendekatan terbaik dibandingkan dua pendekatan lainnya dengan nilai f-measure tertinggi yaitu 0,958. Hal yang sama juga diperlihatkan pada pengujian data berlabel dimana pendekatan DAGSVM memiliki f-measure terbaik sehingga pendekatan ini selanjutnya diterapkan pada klasifikasi data tidak berlabel untuk menentukan kategori berita tersebut.

Kata kunci: klasifikasi web, SVM multiclass, one-against-all, one-against-one, DAGSVM

ABSTRACT
According to Indonesian State Intelligence Agency data, the transportation accident in Indonesia is the third most killer after coronary heart and Tuberculosis. The high rates of transportation accidents make information related to these topics often found in Indonesian online news portal. Currently, the determination of news categories published is still done manually, so that it needs a system that could classify news automatically. In this research, the classification method in text mining is used to process and analyze web data by the topic of transportation accident categories. Support Vector Machine (SVM) with one-against-all, one-against-one and Directed Acyclic Graph Support Vector Machine (DAGSVM) approaches are used as classification methods. These approaches would be compared in the classification process. The approach with the highest accuracy value (closer to 1) will be the best approach. The classification is done by categorize web data into four categories, namely crash land, sea accidents, air crash and others. The steps of this research consist of data collecting, data cleaning, dictionaries n-gram building, features generating, models developing, and classifying. The data training which are used are 10,948 web pages, whereas the data testing are 550 web pages, consist of 500 labeled data and 50 unlabeled data. There are three features pattern that each of them is generated based on two kinds of dictionaries (with and without stopwords) then create six different features. The classification results on model testing show that DAGSVM is the best SVM approach compared to the other two approaches with the f-measure value is 0.958. DAGSVM also is the best approach on the testing of labeled data. So this approach applied to the classification of unlabeled data to determination of news categories.

Keywords: web classification, SVM multiclass, one-against-all, one-against-one, DAGSVM