JISKa, Vol. 1, No. 2, September, 2016, Pp.90 – 93
ISSN 2527 - 5836

PERBANDINGAN NAÏVE BAYES CLASSIFIER DAN SUPPORT VECTOR MACHINE UNTUK KLASIFIKASI JUDUL ARTIKEL

Muhammad Rifqi Ma'arif
Program Studi Manajemen Informatika
STMIK Jenderal Achmad Yani Yogyakarta
e-mail : rifqi@stmikayani.ac.id

Abstract

Support Vector Machine (SVM) and Naïve Bayes Classifier (NBC) is a two popular algorithm to solve text mining problem specifically for text classification. In previous work, a lot of researches stated that SVM has outperforms NBC in term of classification accuracy. One of the interesting thing that we found in previous researches is they use almost same kind of text dataset. They used tweet data from Twitter which is kind of informal text. In this work we will try those two algorithms to classify more formal text. We use data of article titles. Surprisingly we get pretty much different results. In our experiment NBC have better performance than SVM.

Keywords : text classification, naïve bayes classifier, support vector machine

Support Vector Machine (SVM) dan *Naïve Bayes Classifier* (NBC) merupakan dua algoritma yang sangat polpuler untuk text mining, khususnya untuk klasifikasi teks. Pada penelitian-penelitian sebelumnya SVM cenerung menghasilkan performa yang lebih baik dari NBC pada segi akurasi hasil klasifikasi. Salah satu hal yang menarik dari penelitian-penelitian sebelumnya adalah penggunaan jenis data yang hamper sama antara satu dengan lainnya. Penelitian-penelitian sebelumnya kebanyakan menggunakan data tweet dari situs Twitter. Data tweet merupakan jenis teks yang informal dengan banyak sekali noise dan tidak mengindahkan aturan tata bahasa. Pada penelitian kali ini, akan algoritma SVM dan NBC akan diujicobakan kedalam data teks yang lebih formal, yakni data dari judul-judul artikel. Dalam percobaan yang sudah dilakukan, didapatkan hasil yang berbeda dengan penelitian sebelumnya.Pada klasifikasi teks judul artikel NBC memiliki performa akurasi yang lebih baik jika dibandingkan dengan SVM.

Kata Kunci : klasifikasi teks, naïve bayes classifier, support vector machine

1. PENDAHULUAN

Teks merupakan format yang cukup sederhana dan mudah diimplementasikan untuk merepresentasikan data.Karena kemudahannya tersebut saat ini teks menjadi format yang cukup dominan untuk menyimpan ataupun merepresentasikan data dan informasi.Media online seperti blog dan website banyak yang menggunakan format teks untuk menyajikan konten yang mereka miliki kepada pembaca.Namun disamping kesederhanaannya, format data teks yang tidak terstruktur membuatnya sulit untuk diolah oleh sistem komputer. Tidak seperti format tabular pada basis data relasional dan format *hierarchical* pada XML (eXtensible Markup Language) yang lebih terstruktur, format data teks yang tidak terstruktur juga mengandung banyak sekali noise yang membuat proses *information retrieval* lebih sulit untuk dilakukan.

Semakin banyaknya data yang disimpan atau direpresentasikan dalam format teks, mendorong para peneliti untuk memperoleh informasi yang terkandung dalam teks secara otomatis. Cabang keilmuan yang focus pada pengolahan data teks dikenal dengan nama *text mining*. Salah satu kategori dalam text mining adalah klasifikasi teks (text classification). Klasifikasi teks adalah sebuah proses yang bertujuan untuk menentukan kelas atau kategori dari suatu teks. Teks disini bias berupa frase, kalimat, paragraph, atau bahkan dokumen teks. Proses klasifikas teks pada umumnya melibatkan algoritma data mining. Dua algoritma data mining yang sering digunakan untuk klasifikasi teks adalah adalah *Naive Bayes Classifier* (NBC) dan *Support Vector Machine* (SVM).

Penelitian-penelitian sebelumnya yang mengambil tema klasifikasi teks diantaranya dilakukan oleh Hidayatullah dan Azhari SN (2014).Pada penelitian tersebut, digunakan algoritma NBC

dan SVM untuk melakukan analsisi sentiment terhadap data tweet berbahasa Indonesia yang terkait dengan pemilihan presiden tahun 2014. Analisis sentiment merupakan sub bagian dari klasifikasi teks yang bertujuan untuk mengetahui preferensi seseorang terhadap suatu objek, apakah postif, negative, atau netral. Dari hasil percobaan yang dilakukan penelitian tersebut menyimpulkan bahwa algoritma SVM memiliki performa yang lebih baik dibandingkan dengan NBC dari segi akurasi.Penelitian lainnya yang menghasilkan SVM sebagai algoritma dengan performa akurasi terbaik dilakukan oleh Chandani, et.al (2015), Hmeidi et.al (2014) serta Andika dan Widiyantoro (2012).

Dari ketiga penelitian diatas, SVM memiliki tingkat akurasi yang lebih baik apabila dibandingkan dengan NBC ketika digunakan untuk melakukan klasifikasi pada data teks yang berupa tweet pengguna platform microblogging Twiteer[1]. Penelitian-penelitian tersebut menggunakan *word level n-gram* dalam pemilihan fiturnya. Tiga jenis fitur *word level n-gram* diuji coba, yaitu *1-gram*, *2-gram*, serta kombinasi antara *1-gram* dan *2-gram*.

Pada penelitian ini, algoritma NBC dan SVM akan dikomparasikan performanya untuk melakukan klasifikasi pada judul artikel. Judul artikel memiliki karakteristik yang sedikit berbeda dengan data tweet. Judul artikel merupakan serangkaian kata yang disusun secara sistematis sehingga meminimalisir terjadinya ambiguitas.Kata-kata yang digunakan dalam judul artikel biasanya juga lebih sedikit mengandung *noise* apabila dibandingkan dengan kata-kata dalam *tweet* yang cenderung informal.

Data yang digunakan dalam penelitian ini adalah data judul untuk artikel-artikel tentang agama Islam yang diperoleh dari berbagai website. Dataset terdiri dari 1827 judul artikel *unique* yang dibagi menjadi 1600 data untuk *training* dan 274 data untuk *testing*. Data judul artikel yang digunakan terbagi menjadi empat kelas/kategori yaitu Adab, Aqidah, Fiqh dan Al-Qur'an Hadist. Masing-masing kategori memiliki jumlah data *training* yang sama yakni 400 data.

2. METODE PENELITIAN

Tahapan-tahapan dalam penelitian ini merupakan tahapan yang sudah berlaku secara umum dalam melakukan text mining.Tahapan-tahapan dalam *text mining* secara umum adalah *text preprocessing*, *feature selection*dan *evaluation* (Berry &Kogan , 2010). Dimana penjelasan dari tahap-tahap tersebut adalah sebagai berikut :

1. *Text preprocessing*. Dalam melakukan *text mining,* teks dokumen yang digunakan harus dipersiapkan terlebih dahulu, setelah itu baru dapat digunakan untuk proses utama. Proses mempersiapkan teks dokumen atau dataset mentah disebut juga dengan proses *text preprocessing. Text preprocessing* berfungsi untuk mengubah data teks yang tidak terstruktur atau sembarang menjadi data yang terstruktur. Dalam penelitian ini ada tiga tahapan *text preprocessing* yang dilakukan yaitu tokenisasi, *stopword removal,* dan *n*ormalisasi.
2. *Feature selection*. Tahapan ini merupakan tahapan penting dalam *text mining.* Salah satu fungsi penting yang disediakan oleh proses ini adalah untuk dapat memilih term atau kata apa saja yang dapat dijadikan sebagai wakil penting untuk kumpulan dokumen yang akan kita analisis.Dalam penelitian ini, metode feature selection yang digunakan adalah metode *n-gram dan term frequency. Penjelasan mengenai kedua metode tersebut adalah sebagai berikut:*
    a. Fitur *n-gram* digunakan dalam proses pembuatan model dengan membagi suatu kalimat menjadi beberapa bagian kata. Dalam *n-gram*, 'n' menunjukkan jumlah kata yang akan dikelompokkan menjadi satu bagian. Misalnya, apabila n=2 atau biasa disebut dengan bigram, maka sebuah kalimat akan dibagi kedalam masing-masing dua kata pada setiap bagian. Contoh pada kalimat *"Hukum kredit perumahan dengan KPR"*, maka dengan fitur bigram akan dipecah menjadi "*Hukum kredit*", "*kredit perumahan*", " *perumahan dengan*", dan "*dengan KPR*".

[1] http://www.twitter.com

b. *Term frequency* merupakan salah satu metode yang digunakan untuk melakukan perhitungan pembobotan *term.* Fitur *term frequency* dilakukan dengan menghitung frekuensi kemunculan term tertentu pada suatu dokumen.

3. *Evaluation*atau evaluasi dilakukan untuk mengukur validitas hasil klasifikasi. Tiga metode evaluas yang umum digunakan adalah dengan menghitung nilai *precision, recall,* dan *f-score*. Perhitungan nilai *precision* akan mengukur tingkat kepastian *(exactness)* atau jumlah data *testing* yang diklasifikasikan dengan benar oleh model klasifikasi yang dibangun. Perhitungan *recall* merupakan kebalikan dari *precision*. *Recall* mengukur sensitifitas atau rasio dari data untuk setiap label yang diklasifikasikan dengan benar terhadap data yang salah diklasifikasikan ke label lainnya (*missclassified*). *F-score* merupakan *trade-off* antara *precision* dan *recall*. Nilai *f-score* didapat dengan menghitung *harmonic mean* antara *precision* dan *recall.*

3. HASIL DAN PEMBAHASAN

3.1. Perhitungan Akurasi

Proses training dan testing dilakukan dengan fitur *term frequency* dikombinasikan dengan *n-gram*. Hasil perhitungan akurasi dapat dilihat pada tabel 1.

Tabel 1. Hasil Pengukuran Akurasi SVM dan NBC

Fitur	Naive Bayes	SVM
Unigram	0.79	0.73
Bigram	0.79	0.66
Trigram	0.81	0.68

Berdasarkan data pada tabel 1, hasil percobaan menunjukkan bahwa model yang dibangun dengan algoritma *Naive Bayes* menggunakan fitur *trigram* dan *term frequency* memiliki nilai akurasi yang paling tinggi yaitu 81%.Sementara itu, model yang dibangun dengan algoritma SVM terbaik 73% dengan menggunakan fitur *unigram*.Pada penelitian-penelitian sebelumnya, model yang dihasilkan dari algoritma SVM pada umumnya menghasilkan akurasi klasifikasi yang lebih baik dari *Naive Bayes*.Lebih lanjut, pada penggunaan algoritma *Naïve Bayes*, fitur *trigram* memberikan hasil akurasi yang paling baik dibanding fitur *unigram* maupun *bigram*.Sementara itu, pada model klasifikasi yang dibangun dengan menggunakan SVM, fitur *unigram* memberikan hasil yang paling bagus diantara *bigram* maupun *trigram*.

3.2. Perhitungan Precision, Recall dan f-Score

Hasil perhitungan evaluasi menggunakan *precision, recall*, dan *f-score* dalam penelitian ini secara lebih rinci ditunjukkan oleh Tabel II.

Tabel 1. Hasil Deteksi Robot dan Bola.

Fitur	Naive Bayes			SVM		
	Precision	Recall	F-Score	Precision	Recall	F-Score
Uinigram	0.77	0.77	0.77	0.58	0.63	0.65

Bigram	0.77	0.77	0.77	0.61	0.52	0.60
Trigram	0.79	0.78	0.78	0.63	0.54	0.62

Berdasarkan hasil yang diperlihatkan pada table 2, diketahui bahwa perolehan *f-score* dengan algoritma *Naive Bayes* memberikan hasil yang lebih tinggi dibandingkan dengan algoritma SVM. Nilai *f-score* dengan algoritma *Naive Bayes* diketahui mencapai 77% untuk fitur *unigram* dan *bigram*.Sedangkan pada fitur *trigram,* nilai *f-score* mencapai 78%.Perolehan *f-score* dengan algoritma SVM dengan fitur *unigram* diketahui memiliki nilai tertinggi yaitu 65%. Adapun nilai *f-score* menggunakan fitur *bigram* yaitu 60% dam fitur *trigram* diperoleh sebesar 62%.

4. KESIMPULAN

Berdasarkan hasil eksperimen, model klasifikasi yang dihasilkan oleh Naïve Bayes Classifier memiliki performa yang lebih tinggi dengan perbedaan dibandingkan dengan model yang dihasilkan dari yang dihasilkan dari algoritm SVM.Hal tersebut disebabkan penggunaan kernel *linear* dalam algoritma SVM yang digunakan, sedangakankernel *linear* hanya berjalan optimal untuk kasus klasifikasi biner (*binary classification*).SVM akan berjalan optimal pada kasus*multiclass classification*seperti *text mining*dengan menggunakan kernel yang didesain untuk data multidimensi. Hal tersebut diluar cakupan dari percobaan ini dan akan menjadi bagian dalam penelitian selanjutnya.Pada perhitungan *precision, recall,* dan *f-score* diketahui bahwa hasil perhitungan *ketiganya* memiliki pola yang sama dengan perhitungan akurasi. Secara keseluruhan, hasil perolehan *f-score* dengan algoritma *Naive Bayes* memberikan hasil yang lebih tinggi dibandingkan dengan algoritma SVM.Model klasifikasi yang dibangun dengan *Naïve Bayes* memiliki nilai tertinggi ketika menggunakan fitur *trigram*, sementara model klasifikasi yang dibangun dengan SVM memiliki nilai tertinggi ketika menggunakan fitur *unigram*.

DAFTAR PUSTAKA

Berry, M.W. and Kogan, J., 2010. Text Mining.Applications and Theory. West Sussex, PO19 8SQ, UK: John Wiley & Sons.

V. Chandani, R. S. Wahono dan Purwanto., 2015. Komparasi algoritma klasifikasi machine learning dan feature selection pada analisis sentimen review film, Journal of Intelligent Systems*,* vol. 1, no. 1, Februari 2015.

I. Hmeidi, M. Al-Ayyoub, N. A. Abdulla, A. A. Almodawar, R. Abooraig dan N. A. Mahyoub., 2014. Automatic Arabic text categorization: A comprehensive comparative study, Journal of Information Science*,* vol. 41, no. 1, January 2014.

F. R. Andhika dan D. H. Widyantoro., 2012, Klasifikasi topik terhadap teks pendek pada jejaring sosial Twitter,Jurnal Sarjana Institut Teknologi Bandung bidang Teknik Elektro dan Informatika*,* vol. 1, no. 3, Oktober 2012.

A. F. Hidayatullah dan Azhari SN., 2014.Analisis sentimen dan klasifikasi kategori terhadap tokoh publik pada Twitter, dalam Seminar Nasional Informatika 2014 (semnasIF 2014), Yogyakarta, 2014.