Penerapan Text Mining dalam Klasifikasi Judul Skripsi

Ahmad Fathan Hidayatullah
Jurusan Teknik Informatika
Universitas Islam Indonesia
Jl Kaliurang km 14,5 Sleman Yogyakarta
fathan@uii.ac.id

Muhammad Rifqi Ma'arif
Program Studi Manajemen Informatika
STMIK Jenderal Achmad Yani Yogyakarta
Jl. Ringroad Barat, Banyuraden, Gamping, Yogyakarta
rifqi@stmikayani.ac.id

***Abstrak*—Salah satu masalah yang berkaitan dengan *text classification* yang ditemukan di perguruan tinggi yaitu proses pengelompokkan judul skripsi secara otomatis. Penelitian ini bertujuan untuk membuat model data judul skripsi di bidang informatika menggunakan *Support Vector Machine* (SVM) dan *Naïve Bayes*. Berdasarkan hasil eksperimen, model SVM memiliki akurasi yang lebih rendah dengan perbedaan yang cukup signifikan jika dibandingkan dengan model yang dihasilkan dari algoritma *Naive Bayes*.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.**

Kata Kunci—text mining; klasifikasi teks; naïve bayes; support vector machine

I. PENDAHULUAN

Perkembangan data teks saat ini telah mencapai jumlah yang cukup besar. Hal tersebut disebabkan oleh berkembangnya dunia teknologi informasi yang terdiri dari data teks di dalamnya. Saat ini, berbagai macam media *online* seperti *blog*, situs berita *online,* dan jejaring sosial menjadi sumber data teks yang sangat potensial untuk digali lebih dalam. Namun, data berbentuk teks memiliki karakteristik yang tidak terstruktur dan sangat banyak memuat *noise*. Oleh karena itu, *text mining* memiliki peran penting dalam bidang *data mining*. Dengan mengaplikasikan proses-proses dalam *text mining*, maka akan diperoleh pola-pola data, tren, dan ekstraksi dari pengetahuan-pengetahuan yang potensial dari data teks [1].

Diantara proses yang dapat dilakukan dalam *text mining* adalah klasifikasi teks. Klasifikasi teks dapat didefinisikan sebagai proses untuk menentukan suatu dokumen teks ke dalam suatu kelas tertentu. Untuk melakukan proses klasifikasi teks, ada beberapa algoritma yang dapat digunakan diantaranya *Support Vector Machine* (SVM), *Naive Bayes*, *k-Nearest Neighbor* (KNN), *Decision Tree*, dan *Artificial Neural Networks* (ANN).

Salah satu masalah yang berkaitan dengan *text classification* yang ditemukan di perguruan tinggi yaitu proses pengelompokkan judul skripsi secara otomatis. Hal tersebut diharapkan dapat membantu dalam memberikan gambaran kepada mahasiswa untuk mencari judul skripsi yang relevan dengan bidang konsentrasi mahasiswa tersebut. Oleh karena itu, dalam makalah ini akan dilakukan klasifikasi judul skripsi di bidang informatika. Dalam hal ini, data judul skripsi yang digunakan adalah judul skripsi mahasiswa jurusan teknik informatika di Universitas Islam Indonesia. Data skripsi tersebut akan diklasifikasikan berdasarkan minat studi yang ada di jurusan teknik informatika.

Penelitian ini bertujuan untuk membuat model data judul skripsi di bidang informatika. Setelah model terbentuk maka diharapkan akan dapat membantu pengklasifikasian data skripsi baru secara otomatis. Algoritma yang digunakan dalam klasifikasi adalah *Support Vector Machine* (SVM) dan *Naïve Bayes*. SVM dan *Naïve Bayes* dipilih ini karena keduanya memiliki akurasi yang cukup baik dalam klasifikasi teks [2] [3].

Penulisan makalah ini terdiri dari lima bagian. Bagian pertama merupakan pendahuluan yang memuat latar belakang dari penelitian ini. Bagian kedua membahas tentang penelitian-penelitian sebelumnya yang terkait danmendukung penelitian ini. Selanjutnya, bagian ketiga berisi tentang metode yang digunakan dalam penelitian ini. Bagian keempat memaparkan hasil eksperimen dan pembahasan dari hasil yang diperoleh. Bagian kelima adalah bagian terakhir yang berisi kesimpulan dari penelitian.

II. PENELITIAN TERKAIT

Beberapa penelitian terkait dengan klasifikasi teks telah banyak dilakukan sebelumnya. Guo, et al. [3] menggunakan algoritma *Naïve Bayes* untuk melakukan klasifikasi teks berbahasa Cina ke dalam sembilan kategori. Penelitian ini menyimpulkan bahwa algoritma *Naïve Bayes* cukup cepat dan memiliki akurasi yang baik dalam klasifikasi. Samodra, et al. [4] juga memperoleh hasil akurasi yang baikmenggunakan algoritma *Naïve Bayes* untuk mengkasifikasikan dokumen teks berbahasa Indonesia.

Dalam penelitian lain yang melakukan klasifikasi data *tweet* berbahasa Indonesia, Hidayatullah dan Azhari SN [5] menggunakan pula algoritma *Naïve Bayes*. Selain *Naïve Bayes*, algoritma SVM juga dipilih sebagai pembanding.

Dari hasil eksperimen, SVM memiliki performa akurasi yang lebih baik dibandingkan algoritma *Naïve Bayes*.

Penelitian lain menggunakan *Naïve Bayes* dan SVM telah dilakukan oleh Chandani, et al. [6]. Penelitian ini mengkomparasi antara tiga buah algoritma yaitu *Support Vector Machine* (SVM), *Naïve Bayes*, dan *Artificial Neural Network* (ANN). Berdasarkan hasil percobaan, diperoleh bahwa SVM memberikan akurasi terbaik.

Berdasarkan eksperimen yang dilakukan oleh Hmeidi et al.[7] pada klasifikasi teks berbahasa Arab, SVM memiliki hasil akurasi yang paling baik dibandingkan dengan *Naïve Bayes,* KNN, *Decision Tree*, dan *Decision Table*.

Andhika dan Widyantoro [8] juga menggunakan *Naïve Bayes*, *Support Vector Machine* (SVM),dan *Decision Tree* dalam penelitiannya. Dari ketiga metode yang digunakan dalam eksperimen, SVM memiliki akurasi yang paling baik dalam klasifikasi topik *tweet* pada Twitter. Selain itu, penelitian ini juga 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*. Dari hasil percobaan, diperoleh bahwa fitur *1-gram(unigram)* merupakan jenis fitur yang paling baik untuk klasifikasi topik.

III. DATASET

Data judul skripsi yang digunakan dalam penelitian ini terdiri dari empat kelas yaitu Sistem Informasi dan Rekayasa Perangkat Lunak (SIRPL), Komputasi dan Sistem Cerdas (KSC), Grafika dan Multimedia (GFMD), serta Sistem dan Jaringan Komputer (SJK). Sebanyak 824 data judul skripsi digunakan dalam penelitian ini. Data judul skripsi kemudian dibagi menjadi dua bagian yaitu 624 judul sebagai data *training* dan sebanyak 200 judul sebagai data *testing*. Dari ke-624 data *training*, masing-masing kategori judul memiliki jumlah data yang sama yaitu 156 data. Dari model klasifikasi yang dihasilkan, kemudian diuji dengan menggunakan 200 data dimana masing-masing kategori judul memiliki 50 data *testing*. Tabel I menjelaskan secara detail mengenai dataset yang digunakan dalam penelitian ini.

TABEL I. DATASET

Class Label	**Data *Training***	**Data *Testing***
SIRPL	156	50
SJK	156	50
KSC	156	50
GFMD	156	50
Total	624	200

IV. METODE PENELITIAN

A. Pre-processing

Mekanisme *pre-processing* pada penelitian ini cukup berbeda dengan *pre-processing* yang dilakukan pada klasifikasi data *tweet* atauSMS. Pada data *tweet* dan SMS dimungkinkan banyak terdapat kata-kata tidak baku di dalamnya. Sebaliknya, data judul skripsi memiliki karakteristik kalimat yang cukup baku dan sesuai kaidah bahasa yang benar sehingga tidak terlalu banyak *noise* dan kata-kata yang tidak baku.

Oleh karena itu, tahapan *pre-processing* dalam proses klasifikasi data judul skripsi ini terdiri dari tiga tahapan yaitu *case folding*, penghilangan *stopword*, dan tokenisasi. Proses *stemming* tidak dilakukan dalam percobaan ini. Hal ini didasarkan pada penelitian yang dilakukan oleh Hidayatullah [9]. Pada penelitian tersebut, telah dilakukan klasifikasi teks dengan membandingkan tahapan *pre-processing* dengan dan tanpa stemming. Hasilnya, *stemming* pada teks bahasa Indonesia justru menurunkan akurasi hasil klasifikasi.

B. Pemilihan Fitur

1. Fitur *N-gram*

Fitur *n-gram* digunakan dalam proses pembuatan model dengan membagi suatu kalimat menjadi beberapa bagian kata. Dalam penelitian ini, dilakukan perbandingan tiga buah fitur *n-gram* yaitu *unigram, bigram*, dan *trigram*. 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 dipecah menjadi masing-masing dua kata pada setiap bagian. Apabila terdapat kalimat *"Perancangan sistem informasi akademik"*, maka dengan fitur bigram akan dipecah menjadi sebagai berikut :

Bagian 1 : perancangan sistem
Bagian 2 : sistem informasi
Bagian 3 : informasi akademik

2. *Term Frequency*

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.

C. Metode Evaluasi

Diantara mekanisme yang dapat dilakukan untuk mengukur validitas hasil klasifikasi 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*.

V. HASIL DAN PEMBAHASAN

A. Perhitungan Akurasi pada Model Klasifikasi

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

TABEL II. HASIL PERHITUNGAN AKURASI MODEL KLASIFIKASI

Fitur	***Naïve Bayes***	**SVM**
Unigram	0.97	0.73
Bigram	0.97	0.59
Trigram	0.98	0.60

Berdasarkan Tabel II, hasil percobaan menunjukkan bahwa model yang dibangun dengan algoritma *Naive Bayes* menggunakan fitur *trigram* dan *term frequency* memiliki nilai akurasi yang paling tinggi yaitu 98%. 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*.

B. Perhitungan Precision, Recall, dan F-Score

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

TABEL III. PERHITUNGAN *PRECISION, RECALL,* DAN *F-SCORE*

Fitur	*Naive Bayes*			SVM		
	Prec	***Rec***	***F-Score***	***Prec***	***Rec***	***F-Score***
Unigram	0.97	0.97	0.97	0.78	0.73	0.75
Bigram	0.97	0.97	0.97	0.81	0.62	0.70
Trigram	0.99	0.98	0.98	0.83	0.64	0.72

Berdasarkan hasil yang diperlihatkan pada Tabel III, 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 97% untuk fitur *unigram* dan *bigram*. Sedangkan pada fitur *trigram,* nilai *f-score* mencapai 98%. Perolehan *f-score* dengan algoritma SVM dengan fitur *unigram* diketahui memiliki nilai tertinggi yaitu 75%. Adapun nilai *f-score* menggunakan fitur *bigram* yaitu 70% dam fitur *trigram* diperoleh sebesar 72%.

Pada pembahasan selanjutnya, pembahasan difokuskan pada hasil perhitungan *precision* dan *recall* untuk masing-masing fitur dan label. Lebih lanjut, pembahasan juga hanya difokuskan pada hasil perhitungan untuk model klasifikasi yang dibangun dengan algoritma *Naïve Bayes* yang memiliki hasil pengujian yang lebih baik dibandingkan dengan SVM.

C. Perhitungan Precision dan Recall untuk Tiap Fitur dan Label

Tabel IV sampai dengan Tabel VI masing-masing menyajikan hasil perhitungan *precision* dan *recall* untuk masing-masing *class label*. Secara berturut-turut, Tabel IV, V, dan VI menunjukan hasil perhitungan untuk fitur *unigram, bigram* dan *trigram* dari model klasifikasi yang dibangun dengan algoritma *Naïve Bayes*.

TABEL IV. HASIL PERHITUNGAN PRECISION DAN RECALL UNTUK FITUR UNIGRAM

Class Label	***Precision***	***Recall***
SIRPL	0.967	0.989
SJK	1.0	0.936
KSC	0.948	0.978
GFMD	0.989	1.0

Pada Tabel IV, didapatkan tiga kategori *(class label)*dengan menggunakan fitur *unigram*yang memiliki nilai *recall* lebih besar dari nilai *precision*, yaitu SIRPL (Sistem Informasi dan Rekayasa Perangkat Lunak), KSC (Komputasi dan Sistem Cerdas) serta GFMD (Grafika dan Multimedia). Hanya ada satu kategori yang memiliki nilai *precision* yang lebih besar dari nilai *recall* yaitu SJK (Sistem dan Jaringan Komputer).

Dalam tabel tersebut, juga didapatkan nilai *precision* yang sempurna untuk SJK dan nilai *recall* yang sempurna untuk kategori GFMD. Hal pertama menunjukkan bahwa model klasifikasi yang dibangun tidak membuat kesalahan satupun dalam mengklasifikan judul skripsi yang masuk kategori SJK. Hal kedua menunjukkan bahwa model klasifikasi yang dibangun bisa dengan tepat memberikan label GFMD kepada data yang sesuai.

TABEL V. HASIL PERHITUNGAN PRECISION DAN RECALL UNTUK FITUR BIGRAM

Class Label	***Precision***	***Recall***
SIRPL	0.978	0.947
SJK	0.959	0.989
KSC	0.959	0.979
GFMD	1.0	0.979

Berdasarkan hasil pada Tabel V, didapatkan dua kategori dengan fitur *bigram*yang memiliki nilai *recall* lebih besar dari nilai *precision*, yaitu SJK dan KSC dan juga dua kategori dengan nilai *precision* yang lebih besar yakni SIRPL dan GFMD. Dalam tabel tersebut, didapatkan pula nilai *precision* yang sempurna untuk GFMD. Hal ini menunjukkan bahwa model klasifikasi yang dibangun dengan menggunakan fitur *bigram* tidak membuat kesalahan satupun dalam mengklasifikan judul skripsi yang masuk kategori GFMD.

TABEL VI. HASIL PERHITUNGAN PRECISION DAN RECALL UNTUK FITUR TRIGRAM

Class Label	***Precision***	***Recall***
SIRPL	0.989	0.968
SJK	1.0	0.978
KSC	0.979	0.978
GFMD	1.0	1.0

Tabel VI menunjukkan bahwa dengan menggunakan fitur *trigram* didapatkan hampir semua kategori memiliki nilai *precision* yang lebih besar dari nilai *recall*. Dalam tabel tersebut diperoleh nilai *precision* yang sempurna untuk GFMD dan SJK dan nilai *recall* yang sempurna untuk GFMD. Penggunaan fitur *trigram* dalam studi kasus judul skripsi memberikan model klasifikasi akurasi yang sempurna dalam mengklasifikasikan judul skripsi yang termasuk kategori Grafika dan Multimedia. Dengan nilai *precision* dan *recall* yang sempurna, maka tidak satupun data yang seharusnya termasuk kategori tersebut diklasifikasikan ke kategori yang lain. Begitu juga dengan label GFMD tidak diberikan ke selain judul skripsi yang termasuk kategori Grafika dan Multimedia.

VI. KESIMPULAN

Dalam percobaan ini, telah berhasil dibangun model untuk melakukan klasifikasi judul skripsi bidang Teknik Informatika. Secara garis besar, terdapat dua jenis model yang dibangun dengan dua pendekatan yang berbeda yaitu *Naive Bayes*dan *Support Vector Machine*.

Berdasarkan hasil eksperimen, model SVM memiliki akurasi yang lebih rendah dengan perbedaan yang cukup signifikan jika dibandingkan dengan model yang dihasilkan dari algoritma *Naive Bayes*. Hal tersebut kemungkinan disebabkan algoritma SVM yang digunakan dalam percobaan ini adalah algoritma SVM dengan linear kernel. Linear kernel pada SVM hanya berjalan optimal untuk kasus klasifikasi biner (*binary classification*), sedangkan data skripsi yang digunakan dalam percobaan ini memiliki lebih dari satu jenis *class* label sehingga dikategorikan sebagai *multiclass classification*. SVM akan berjalan optimal pada *multiclass classification* 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

[1] A. Kao dan S. Poteet, "Text mining and natural language processing: introduction for the special issue," *SIGKDD Explor. Newsl,* vol. 7, no. 1, pp. 1-2, 2005.

[2] P. Y. Zhang, "A HowNet-based Semantic Relatedness Kernel for Text Classification," *TELKOMNIKA Indonesian Journal of Electrical Engineering,* vol. 11, no. 4, pp. 1909-1915, 1 Apr 2013.

[3] D. Li Guo, D. Peng dan L. Ai Ping, "A New Naive Bayes Text Classification Algorithm," *TELKOMNIKA Indonesian Journal of Electrical Engineering,* vol. 12, no. 2, pp. 947-952, 1 Feb 2014.

[4] J. Samodra, S. Sumpeno dan M. Hariadi, "Klasifikasi dokumen teks berbahasa Indonesia dengan menggunakan naive bayes," dalam *Seminar Nasional Electrical, Informatics, and IT's Education*, 2009.

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

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

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

[8] F. R. Andhika dan D. H. Widyantoro, "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.

[9] A. F. Hidayatullah, "The Influence of Indonesian Stemming on Indonesian Tweet Sentiment Analysis," dalam *Proceeding of International Conference on Electrical Engineering, Computer Science and Informatics (EECSI 2015)*, Palembang, Indonesia, 2015.