Scientific Journal of Informatics
Vol. 3, No. 1, Mei 20xx

p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

Metode *K-Means* untuk Optimasi Klasifikasi Tema Tugas Akhir Mahasiswa Menggunakan *Support Vector Machine* (SVM)

Oman Somantri[1], Slamet Wiyono[2], Dairoh[3]

[1,2,3]Jurusan Teknik Informatika, Politeknik Harapan Bersama Tegal
Email: [1]oman_mantri@yahoo.com, [2]slamet2wiyono@gmail.com, [3]dairoh@poltektegal.ac.id

Abstrak

Masih sulitnya dalam menentukan klasifikasi tema tugas akhir mahasiswa sering dialami oleh setiap perguruan tinggi. Algoritma SVM digunakan untuk mengklasifikasi jenis tema tugas akhir mahasiswa. SVM merupakan metode yang banyak digunakan untuk klasifikasi. *K-Means Clustering* merupakan metode pengelompokan paling sederhana yang mengelompokkan data kedalam *k* kelompok berdasar pada *centroid* masing-masing kelompok. Optimasi klasifikasi tema tugas akhir mahasiswa menggunakan SVM dan *K-Means* untuk meningkatkan tingkat akurasi. Hasil yang diperoleh memiliki tingkat akurasi yang lebih baik yaitu 86,21%.

Kata Kunci: *Text mining*, *Support Vector Machine*, *K-Means*, Tugas Akhir

1. PENDAHULUAN

Mengambil mata kuliah tugas akhir atau skripsi merupakan kewajiban bagi setiap mahasiswa Diploma tingkat akhir, karena dengan tugas akhir itulah menentukan apakah mahasiswa tersebut lulus atau tidak dalam ujian sidang tugas akhir. Menentukan sebuah tema tugas akhir dan skripsi untuk mencari masalah penelitian menjadi salah satu kesulitan utama bagi mahasiswa, hal ini tentunya akan berpengaruh kepada tepat atau tidaknya mahasiswa tersebut lulus kuliah. Judul tugas akhir yang harus sesuai dengan tema untuk menjawab masalah penelitian yang telah ditentukan oleh setiap perguruan tinggi sesuai dengan jurusan yang diambil memberikan sebuah ketetapan bahwa judul tugas akhir yang diajukan oleh mahasiswa haruslah sesuai dengan jenis tema yang ditentukan.

Kesulitan yang dialami oleh banyak para pengambil kebijakan di perguruan tinggi dalam hal ini pada jurusan program studi adalah dalam menentukan klasifikasi tema dari judul tugas akhir yang diajukan oleh mahasiswa masih hanya berdasarkan intuisi, karena selama ini dalam penentukan klasifikasi jenis tema tugas akhir hanya berdasarkan pada perkiraan terhadap isi konten yang akan diteliti oleh mahasiswa sehingga kesesuaian antara judul dan tema berdasarkan pada teks judul terkadang diabaikan bahkan tidak sesuai.

Text mining merupakan pengembangan dari metode *data mining* yang dapat diterapkan untuk mengatasi permasalah terkait dengan pengklasifikasian tema judul tugas akhir mahasiswa [1]. Algoritma-algoritma dalam *text mining* dibuat untuk dapat mengenali data yang sifatnya semi terstruktur seperti sinopsis, abstrak maupun isi dari dokumen-dokumen. Kategori teks atau klasifikasi teks adalah suatu proses yang mengelompokan suatu teks kedalam suatu kategori tertentu [2]. Kategorisasi teks

membuat pengelolaan informasi tersebut menjadi efektif dan efisien, sehingga dapat digunakan seperti untuk penyaringan terhadap email spam, melakukan penggalian opini (*opinion mining*), dan analisis sentimen. Algoritma kategorisasi teks saat ini telah banyak berkembang, antara lain *Support Vector Machines* (SVM), *Naïve Bayessian* (NB), pohon keputusan, *K-Nearest Neighbour* (KNN), dan lainnya. Dari berbagai macam algoritma yang telah dikembangkan tersebut, KNN dan SVM telah diakui lebih handal dibandingkan dengan algoritma yang lainnya [3]. Pada penelitian yang dilakukan oleh Wulandini & Nugroho (2009), membandingkan metode klasifikasi teks NBC dengan metode *Support Vector machine* (SVM), C4.5 dan *K-Nearest Neighbour* (K-NN), hasil penelitian menunjukkan akurasi masing-masing metode dari yang terbaik adalah SVM akurasi 92%, NBC akurasi 90% C4.5 akurasi 77.5% dan yang terendah K-NN akurasi 50% [4].

Algoritma *Support Vector Machines* (SVM) digunakan untuk klasifikasi penentuan jenis tema tugas akhir mahasiswa. SVM adalah metode yang banyak digunakan untuk klasifikasi data berupa *text* dengan tingkat akurasi yang lebih baik. Tetapi dalam hal ini untuk proses klasifikasi dokumen, seringkali ditemukan hasil yang kurang baik dikarenakan jumlah data dokumen yang besar dan bervariasi sehingga harus dikelompokkan terlebih dahulu [5]. Pada paper ini *K-Means Clustering* digunakan untuk memperbaiki proses klasifikasi data teks yang dilakukan yaitu dengan terlebih dahulu dilakukannya klasterisasi data agar tingkat akurasi model yang diusulkan menjadi lebih baik.

Penelitian terkait dengan klasifikasi tugas akhir dan skripsi telah dilakukan oleh beberapa peneliti, diantaranya Prilianti & Wijaya (2014) meneliti mengenai pengembangan aplikasi berbasis *text mining* untuk automasi penentuan *trend* topik skripsi dengan metode *K-Means Clustering* [6]. Klasifikasi tugas akhir dan skripsi dengan algoritma *K-Means clustering* digunakan dalam proses penemun pola terbukti dapat membantu proses pengelompokan berbagai topik skripsi yang ada sehingga diperoleh informasi yang bermakna dalam menentukan tren penelitian Universitas dari tahun ke tahun.

Jiang, S., dkk., meneliti mengenai *improve* atau pengembangan model algoritma *K-Nearest Neighbor* untuk klasifikasi *text*. Dalam penelitian ini untuk mengoptimalisasi K-NN digunakan *one pass clustering algorithm* sehingga tingkat akurasi klasifikasi *text* menjadi lebih baik [7]. Penelitian selanjutnya dilakukan oleh Wangsa B.K, dkk., meneliti mengenai pembuatan sistem peringkat berita otomatis berbasis *text mining* menggunakan *Generalized Vector Space Model* dengan studi kasus berita diambil dari media masa *online*. Pada penelitian ini dihasilkan dengan menggunakan metode GVSM dapat diketahui kalimat mana yang lebiih berbobot terhadap suatu dokumen sehingga dapat dilakukan peringkasan dengan memperhatikan tingkat kemiripan kalimat dengan tingkat akurasi 94%[8].

Dari ketiga penelitian sejenis tersebut diatas, berbeda dengan penelitian sebelumnya, yaitu untuk dapat meningkatkan tingkat akurasi pada SVM maka diterapkan algoritma *K-Means* sehingga tingkat akurasi untuk klasifikasi dalam penentuan tema tugas akhir mahasiswa sesuai dengan judul yang diajukan oleh mahasiswa terjadi peningkatan akurasi.

Penggunaan *text mining* untuk mengklasifikasi jenis tema tugas akhir mahasiswa dengan menggunakan algoritma *Support Vector Machines* (SVM) dan dioptimalisasi dengan menggunakan *K-Means Clustering* untuk menghasilkan tingkat akurasi yang lebih baik.

2. METODE

2.1. *Text mining*

Text mining secara umum adalah teori tentang pengolahan koleksi dokumen dalam jumlah besar yang ada dari waktu ke waktu dengan menggunakan beberapa analisis, tujuan pengolahan teks tersebut adalah mengetahui dan mengekstrak informasi yang berguna dari sumber data dengan identifikasi dan eksplorasi pola menarik dalam kasus *text mining*, sumber data yang dipergunakan adalah kumpulan atau koleksi dokumen tidak terstuktur dan memerlukan adanya pengelompokan untuk diketahui informasi sejenis.

Text mining menurut Han & Kamber, adalah satu langkah dari analisis teks yang dilakukan secara otomatis oleh komputer untuk menggali informasi yang berkualitas dari suatu rangkaian teks yang terangkum dalam sebuah dokumen. Prosedur utama dalam metode ini terkait dengan menemukan kata-kata yang dapat mewakili isi dari dokumen untuk selanjutnya dilakukan analisis keterhubungan antar dokumen dengan menggunakan metode statistik tertentu seperti analisis kelompok, klasifikasi dan asosiasi[9]. Tahapan dalam *text mining* secara umum diantaranya adalah *tokenizing*, *filtering*, *stemming*, *tagging*, dan *analyzing* [10].

2.2. *Support Vector Machine* (SVM)

Support Vector Machine (SVM) adalah metode klasifikasi yang bekerja dengan cara mencari *hyperplane* dengan margin terbesar *hyperplane* adalah garis batas pemisah data antar-kelas. Margin adalah jarak antara *hyperplane* dengan data terdekat pada masing-masing kelas. Adapun data terdekat dengan *hyperplane* pada masing-masing kelas inilah yang disebut *support vector* [11]. Pada dasarnya, SVM merupakan metode yang digunakan untuk klasifikasi dua kelas (*binary classification*). Pada perkembangannya, beberapa metode diusulkan agar SVM bisa digunakan untuk klasifikasi *multi-class* dengan cara mengombinasikan beberapa *binary classifier* [12]. Metode yang pernah diusulkan adalah metode *One-against-one*. Adapun untuk metode *One-against-one*, akan dikonstuksi sejumlah *k(k-1)/2* model klasifikasi SVM dengan masing-masing model dilatih menggunakan data dari dua kelas yang berbeda.

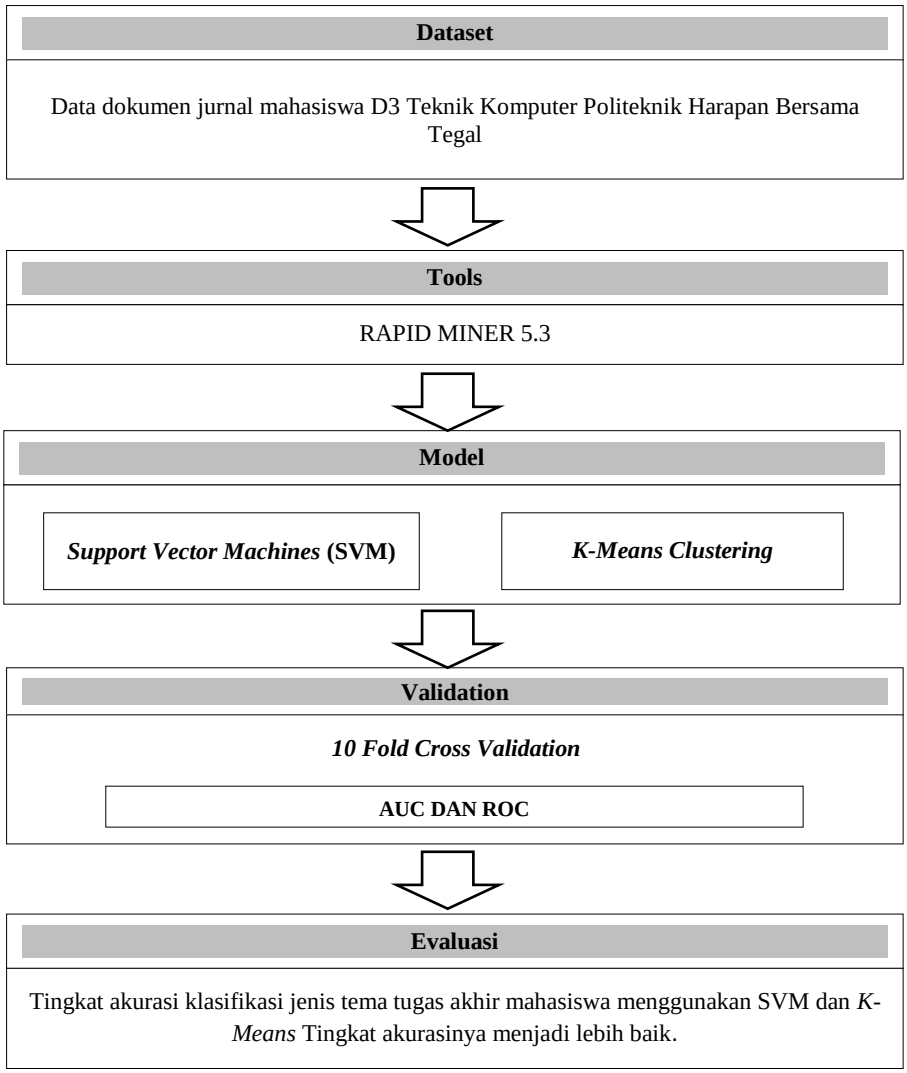
2.3. *K-Means Clustering*

Salah satu metode dalam pengelompokan dokumen adalah *K-Means Clustering*. *K-Means Clustering* merupakan metode pengelompokan paling sederhana yang mengelompokkan data kedalam *k* kelompok berdasar pada *centroid* masing-masing kelompok. Hanya saja hasil dari *K-Means* sangat dipengaruhi parameter *k* dan inisialisasi *centroid*. Pada umumnya *K-Means* menginisialisasi *centroid* secara acak.

Namun metode yang diusulkan akan memodifikasi *K-Means* dalam inisialisasi *centroid* khususnya dalam memperbaiki performa dalam pengelompokan dokumen.

2.4. Metode yang Digunakan

Metode yang digunakan adalah penerapan *K-Means* sebagai model untuk meningkatkan tingkat akurasi untuk klasifikasi jenis tema tugas akhir mahasiswa dengan menggunakan *Support Vector Machine* (SVM), digambarkan seperti pada Gambar 1.



Gambar 1. Tahapan penelitian

2.5. *Dataset* dan Alat Penelitian

Dataset diperoleh dari data jurnal mahasiswa D3 Tenik Komputer Politeknik Harapan Bersama Tegal tahun akademik 2014/2015, sebanyak 131 jurnal mahasiswa dengan berbagai macam jenis tema terdiri dari multimedia, pemrograman *desktop*, dan pemrograman *web*. Alat atau *tools* yang digunakan menggunakan *Software Rapid Miner 5.0*, sebagai pendukung pengolahan data menggunakan Ms. Excel 2007.

2.6. *Preprocessing Data*

Tahap awal sebelum melakukan proses pengelompokan dokumen adalah dengan mempersiapkan teks yang ada didalam dokumen. Pada tahap pra-proses ini dilakukan beberapa subproses agar dokumen dapat dipakai untuk melakukan proses pengelompokan. *Subproses* diantaranya yaitu dilakukan sebagai berikut:

a. *Tokenizer*, yakni proses yang bertujuan untuk memisah teks menjadi beberapa *token* berdasarkan pembatas berupa spasi atau tanda baca.
b. Proses selanjutnya adalah menghilangkan teks yang bersesuaian dengan teks yang terdapat pada daftar *stopword*, karena teks tersebut dianggap tidak dapat mewakili konten dokumen.
c. Kemudian pada teks yang masih tersisa dilakukan proses *stemming*, yaitu proses pengubahan teks menjadi bentuk dasarnya.
d. Selanjutnya, setiap kata tersebut disebut sebagai *term*. Nantinya setiap *term* akan didaftar dan diberi bobot.
e. Pembobotan masing-masing *term* dilakukan dengan metode TF-IDF (*Term Frequency – Inverse Document Frequency*). TF-IDF merupakan metode pembobotan *term* dengan menggunakan *termfrequency* (jumlah *term* yang terdapat pada tiap dokumen) serta *inverse document frequency* (*invers* jumlah dokumen yang memuat suatu *term*).

2.7. Proses Pengelompokan Atau Kategorisasi Dokumen

Proses pengelompokan dilakukan terhadap hasil pra-proses yang merupakan representasi data dalam bentuk model ruang vektor. Metode pertama ialah pengelompokan dokumen yang ada dengan *K-Means Clustering*. Kemudian setelah itu setiap kelompok dokumen tersebut akan diklasifikasi dengan *Multi-Class SVM*.

2.8. Penentuan Data *Training* dan *Testing*

Data *training* dan *testing* diambil dari judul tugas akhir mahasiswa program studi D3 Teknik Komputer Politeknik Harapan Bersama tegal yang diambil pada tahun 2013 dan 2014, dimana setelah dijumlahkan akan di split menjadi 70% data *training* dan 30% data *testing*.

2.9. Eksperimen dan Pengujian

Dari *k* model klasifikasi yang telah ada, maka dapat dilakukan klasifikasi dokumen baru. Pengujian dilakukan dengan mengelompokkan dokumen baru kedalam

kelompok yang ada menggunakan tetangga terdekat dari *centroid* pada masing-masing kelompok. Setelah didapatkan kelompok yang sesuai maka dilakukan proses klasifikasi dokumen baru dengan model *Multi-class SVM* pada kelompok yang bersangkutan.

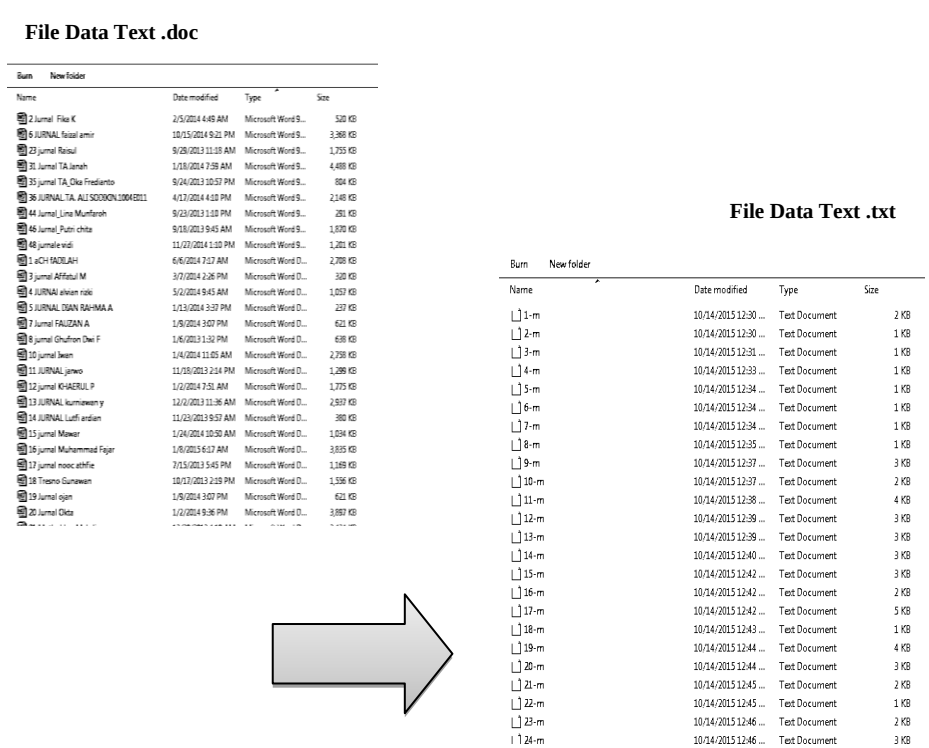
2.10. Evaluasi dan Validasi Penelitian

Sebagai evaluasi dari model yang diusulkan, yaitu dengan menggunakan metode *K-folds Cross validations* untuk mencari nilai akurasi yang kemudian hasil dari akurasi tersebut dievaluasi dengan cara membandingkan tingkat akurasi yang dihasilkan oleh model SVM dengan menggunakan *K-Means* dan dengan model SVM tanpa *K-Means*

3. HASIL DAN PEMBAHASAN

3.1 Pra-processing Data

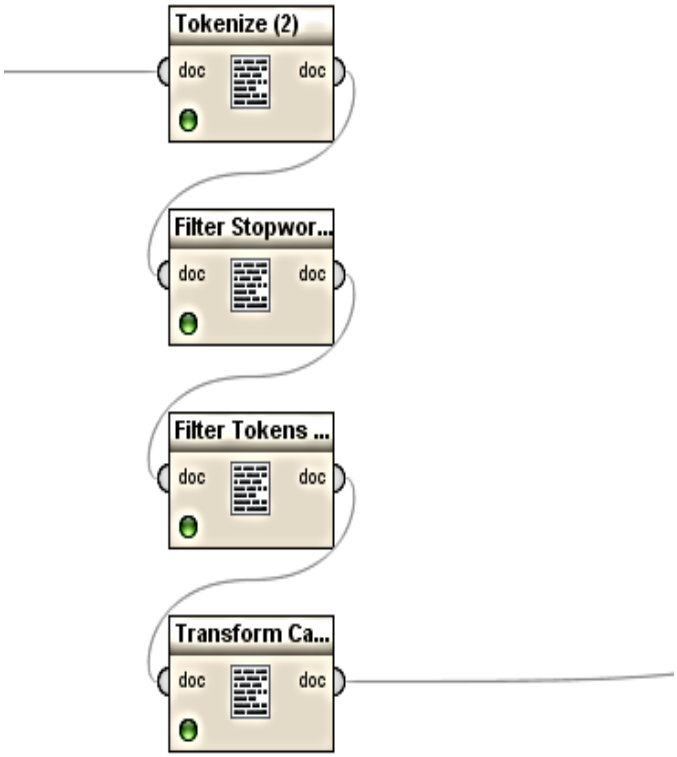
Sebelum *dataset text* diolah dengan menggunakan model yang akan digunakan, maka terlebih dahulu *dataset* yang berupa *file* bertipe .doc di *convert* menjadi *file* tipe .txt. Hal ini dilakukan agar *file text* yang akan digunakan kedalam model dengan menggunakan *tools rapid miner* dapat terbaca dengan baik, gambaran proses perubahan *file* tersebut seperti tampak pada Gambar 2 dibawah ini.



Gambar 2. Proses Perubahan *file* bertipe .doc kedalam tipe .txt

3.2 Tahapan Pengolahan Dokumen

Pada tahapan selanjutnya adalah mengolah *data text* yang akan digunakan kedalam beberapa tahapan, agar nantinya diperoleh inputan *dataset* yang sesuai dengan model yang akan digunakan, diperlihatkan pada Gambar 3 dibawah ini:



Gambar 3. Tahapan pengolahan dokumen

Seperti yang telah diperlihatkan pada Gambar 3, adapun tahapannya adalah sebagai berikut:

a. *Tokenizer*, yakni proses yang bertujuan untuk memisah teks menjadi beberapa token berdasarkan pembatas berupa spasi atau tanda baca.
b. *Filter Stopword*, yakni menghilangkan teks yang bersesuaian dengan teks yang terdapat pada daftar *stopword*, karena teks tersebut dianggap tidak dapat mewakili konten dokumen.
c. *Filter Token*: adalah memfilter teks dengan *Min* char=5 dan *Max* char=25.
d. *Transforms cases*: mentranformasikan teks kedalam *lower case.*

3.3 Pembobotan TF-IDF (*Term Frequency – Inverse Document Frequency*)

TF-IDF merupakan metode pembobotan *term* dengan menggunakan *termfrequency* yaitu jumlah *term* yang terdapat pada tiap dokumen serta *inverse document frequency* yaitu *invers* jumlah dokumen yang memuat suatu *term*.

Tabel 1. Pembobotan dokumen

administrasi	administratif	administra..	administrasi	adobe_	agama	agenda	agung	agustus	ahmad	alvana	ajaran
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0,309	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0,056	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0,211	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0,017	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0,151	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0,104	0	0	0	0	0	0,121
0	0	0	0	0	0,047	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0

3.4 Pembahasan

4.4.1 Klasifikasi Data dengan *Support Vector Machines* (SVM)

Pada tahapan pemilihan model yang sesui dengan yang diinginkan, parameter SVM yang digunakan adalah seperti pada Tabel 3.

Tabel 3. Parameter SVM

PARAMETER	VALUE
Kernel type	dot
Kernel cache	200
C	0.0
Convergence epsilon	0.001
Max iteration	100000

Setelah dilakukan ekperimen, maka didapatkan hasil dari model SVM yang digunakan adalah seperti pada Tabel 4.

Tabel 4. Hasil Akurasi model SVM

ACCURACY: 85,38%				
	TRUE.MULTIMEDIA	TRUE.DESKTOP	TRUE.WEB	CLASS PRECISION
PRED.MULTIMEDIA	24	0	0	100%
PRED.DESKTOP	2	41	5	85,42%
PRED.WEB	5	7	47	79,66%
CLASS RECALL	77,42%	85,42%	90,38%	

Dari Tabel 4 diperlihatkan bahwa hasil tingkat akurasi klasifikasi yang dihasilkan adalah sebesar 85,38%.

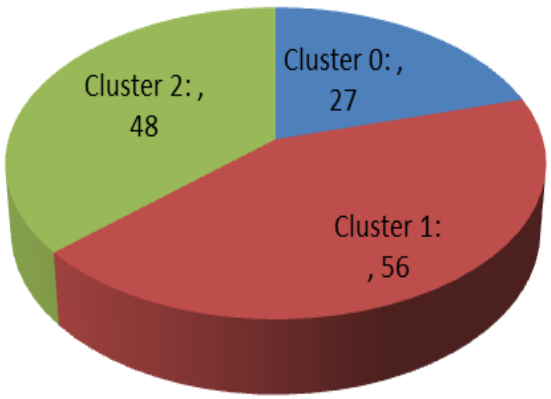
4.4.2 *Clustering Data* dengan *K-Means*

Pada tahapan yang telah dilakukan sebelumnya yaitu menggunakan SVM sebagai model yang digunakan, kemudian langkah berikutnya adalah melakukan eksperimen dengan terlebih dahulu melakukan *clustering data* yang ada dengan menggunakan *K-Means*, sehingga didapatkan hasilnya seperti pada Tabel 5.

Tabel 5. Hasil *Clustering*

CLUSTER	JUMLAH	LABEL
Cluster 0:	27	Multimedia
Cluster 1:	56	Desktop
Cluster 2:	48	Web

Apabila dibuatkan grafik maka hasilnya akan tampak pada Gambar 4 dibawah ini.



Gambar 4. Grafik *Clustering* dengan *K-Means*

4.4.3 Improvement

Berdasarkan ekperimen sebelumnya, maka untuk meningkatkan tingkat akurasi klasifikasi teks, diguanakan model hybrid yaitu menggunakan *Support Vector Machines* (SVM) dan *K-Means Clustering*. Setelah dilakukan eksperimen maka hasilnya didapatkan sebagai berikut:

Tabel 6. Hasil akurasi SVM dan *K-Means*

ACCURACY: 86,21%				
	TRUE.MULTIMEDIA	TRUE.DESKTOP	TRUE.WEB	CLASS PRECISION
PRED.MULTIMEDIA	24	0	0	96,00%
PRED.DESKTOP	1	41	4	89,13%
PRED.WEB	6	6	47	80,00%
CLASS RECALL	77,42%	85,42%	92,31%	

Dari Tabel diperlihatkan bahwa tingkat akurasi yang dihasilkan adalah sebesar 86,21%.

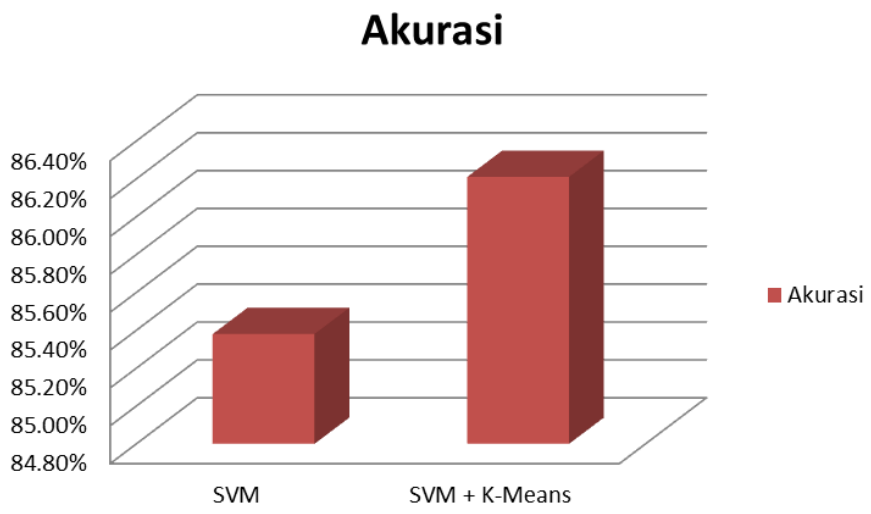
4.4.4 Evaluasi Model

Berdasarkan hasil analisis yang telah dilakukan sebelumnya, maka untuk mengevaluasi hasil dari ekperimen didapatkan hasilnya sebagai berikut:

Tabel 7. Evaluasi model

MODEL	**AKURASI**
SVM	**85,38%**
SVM + K-Means	**86,21%**

Dari hasil Tabel 7 diperlihatkan bahwa setelah dilakukan eksperimen terdapat perbedaan antara model dari SVM dibandingkan dengan model SVM + K-Means, dimana tingkat akurasi sebelumnya 85,38% menjadi 86,21%.



Gambar 5. Grafik perbandingan tingkat akurasi SVM dan SVM+*K-Means*

4. SIMPULAN

Berbagai upaya dilakukan untuk dapat meningkatkan tingkat akurasi sebuah model khususnya pada *text mining*. Untuk meningkatkan tingkat akurasi pada prediksi klasifikasi data dokumen jenis tema tugas akhir mahasiswa, maka model algoritma K-*Means* digunakan sebelum dimasukan kedalam model *Support Vector Machine* (SVM). Dari analisis yang telah dilakukan dapat diambil kesimpulan bahwa model SVM dan *K-Means* dapat digunakan oleh para pengambil kebijakan dalam menglasifikasikan kategori tugas akhir sebagai pendukung keputusan dalam penentuan tema tersebut. *K-Means* menjadi model untuk optimalisasi untuk dapat meningkatkan tingkat akurasi model SVM dalam mengklasifikasikan kategori tema tugas akhir.

5. REFERENSI

[1] Gupta V. 2009. A Survey of Text Mining Techniques and Application. *Journal of Emerging Technologies in Web Intelligence*. Vol. 1: 60-75

[2] Sebastiani F. 2002. Machine Learning in Automated Text Categorization. *ACM Computing Surveys.* Vol. 34(1): 1-47

[3] Yang Y., & Liu X. 1999. A Re-Examination of Text Categorization Methods. In *Proceedings* 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-99). Berkeley: 42-49.

[4] Wulandini F & Nugroho A.N. 2009. Text Classification Using Support Vector Machine for Web mining Based Spation Temporal Analysis of the Spread of Tropical Diseases. *International Conference on Rural Information and Communication Technology*.

[5] Trivedi S, Pardos A & Sar N. 2008. *Spectral Clustering in Educational Data Mining*. Department of Computer Science, Worcester Polytechnic Institute.

[6] Prilianti KR & Wijaya H. 2014. Aplikasi Text Mining untuk Automasi Penentuan Tren Topik Skripsi dengan Metode K-Means Clustering, *Jurnal Cybermatika,* Vol. 2(1).

[7] Jiang S., Pang G., Wu M., & Kuang L. 2012. An Improved K-Nearest-Neighbor Algorithm for Text Categorization. *In Expert Systems With Applications.*

[8] Wangsa B.K., Utomo D., & Nugroho S. 2014. Sistem Peringkas Berita Otomatis berbasis Text Mining menggunakan Generalized Vector Space Model: STudi Kasus Berita diambil dari media Massa Online. *Techne Jurnal Ilmiah Elektroteknika*. Vol. 1(2) Oktober: 231-241.

[9] Han, J., & Kamber, M. 2006. *Data Mining: Concepts and Techniques*, University of Illinois at Urbana-Champaign.

[10] Berry, M. W. 2004. Survey of text mining. *Computing Reviews.* Vol 45(9): 548.

[11] Yunliang, J., Qing, S., Jing, F., & Xiongtao, Z. 2010. The Classification for E-government Document Based on SVM. *In Web Information Systems and Mining (WISM), 2010 International Conference on*.Vol. 2: 257-260.

[12] J.Z. Liang. 2004. SVM Multi-Classifier And Web Document Classification. *Proceedings of the IEEE Third International Conference on Machine Learning and Cybernetics*.