IT FOR SOCIETY, Vol. 04, No. 02
ISSN 2503-2224

Analisa Judul Skripsi untuk Menentukan Peminatan Mahasiswa Menggunakan Vector Space Model dan Metode K-Nearest Neighbor

Dewi Marini Umi Atmaja
President University
Bekasi, Indonesia
marini.umiatmaja@gmail.com
dewi.atmaja@student.president.ac.id

Rila Mandala
President University
Bekasi, Indonesia
rilamandala@president.ac.id

Abstrak - Sulitnya menentukan klasifikasi judul skrpsi berdasarkan peminatan yang diambil oleh mahasiswa informatika unjani merupakan salah satu permasalahan penting yang dihadapi oleh pihak Jurusan. Tujuan dari penelitian ini yaitu memberikan sebuah penunjang keputusan bagi pihak Jurusan agar setiap judul skripsi yang diajukan oleh mahasiswa sesuai dengan peminatan. Berdasarkan hasil penelitian yang telah dilakukan, model yang dibangun menggunakan algoritma KNN menghasilkan tingkat akurasi yang lebih tinggi jika dibandingkan dengan model yang dibangun menggunakan algoritma VSM. Nilai akurasi tertinggi berdasarkan hasil pengujian pada penelitian ini adalah sebasar 96,85%.

Keywords: Text Mining, Vector Space Model, K-Nearest Neighbor

I. PENDAHULUAN

Tugas akhir atau yang biasa disebut dengan skripsi merupakan salah satu matakuliah yang wajib diambil oleh mahasiswa tingkat akhir. Skripsi merupakan sebuah karya ilmiah yang ditulis oleh mahasiswa tingkat akhir yang membahas tentang bidang atau topik tertentu berdasarkan dari suatu hasil kajian pustaka yang ditulis oleh para ahli, hasil pengembangan, atau hasil penelitian dilapangan.

Universitas Jenderal Achmad Yani atau yang biasa disebut UNJANI merupakan suatu organisasi pendidikan yang memiliki banyak data Skripsi hasil karya mahasiswa dari beberapa Fakultas dan Jurusan. Salah satu Jurusan yang dimiliki oleh UNJANI yaitu Informatika. Dalam Jurusan Informatika terbagi atas dua kepeminatan atau konsentrasi, yaitu Sistem Cerdas Data mining (SCDM) dan Sistem Informasi Enterprise (SIE). Peminatan tersebut, dipilih oleh Mahasiswa pada semester 4 (empat) perkuliahan. Menentukan sebuah tema tugas akhir atau skripsi untuk mencari masalah penelitian menjadi salah satu kesulitan utama bagi mahasiswa, hal ini tentunya berpengaruh terhadap tepat atau tidaknya mahasiswa tersebut menyelesaikan perkuliahan. Judul skripsi yang harus sesuai dengan topik dan peminatan merupakan suatu ketetapan dari UNJANI yang harus dipatuhi oleh mahasiswa dan dosen pembimbing.

Kesulitan yang banyak dialami oleh para pengambil kebijakan di UNJANI dalam hal ini pada jurusan Informatika adalah menentukan klasifikasi topik dan peminatan berdasarkan judul tugas akhir yang diajukan oleh mahasiswa. Selama ini klasifikasi topik skripsi hanya berdasarkan pada perkiraan terhadap isi konten yang akan diteliti oleh mahasiswa, sehingga kesesuaian antara judul, topik dan peminatan yang dipilih oleh mahasiswa seringkali tidak sesuai. Hal ini dapat diselesaikan dengan mengaplikasikan *text mining*. Adapun proses-proses yang dilakukan dalam *text mining* menghasilkan pola-pola data, tren, dan ekstraksi serta pengetahuan yang potensial dari data teks. Diantara beberapa proses yang dapat dilakukan dalam *text mining* yaitu proses klasifikasi teks. Klasifikasi teks ini dapat didefinisikan sebagai proses untuk menentukan suatu dokumen teks ke dalam suatu kelas tertentu. Dalam melakukan proses klasifikasi teks terdapat beberapa algoritma yang dapat digunakan, diantaranya yaitu menggunakan algoritma *K-Nearest Neighboard* (KNN) dan *Vector Space Model* (VSM). Adapun data yang digunakan dalam penelitian ini yaitu data alumni mahasiswa Informatika Unjani dari tahun 2010 hingga saat ini.

II. LANDASAN TEORI

A. Text Mining

Teks 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 (2006), 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. Menurut Berry, M. W. (2004), tahapan dalam text mining secara umum adalah tokenizing, filtering, stemming, tagging, dan analyzing.

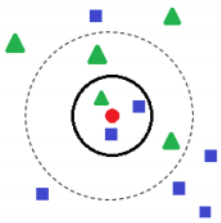
B. Vector Space Model

Vector Space Model (VSM) adalah suatu cara yang digunakan untuk mengukur tingkat kemiripan (*similarity*) antar satu dokumen dengan suatu *query*. Pada model ini, *query* dan dokumen dianggap sebagai vector-vektor pada ruang n dimensi. Dimana n adalah jumlah dari seluruh *concept* yang ada didalam indeks. Dalam proses *clustering* ataupun klasifikasi membutuhkan nilai *similarity* dari satu dokumen dengan dokumen lain, umumnya diukur dengan fungsi *similarity* tertentu.

Fungsi *similarity* tersebut umumnya dinamakan sebagai fungsi *cosine similarity. Cosine similarity* adalah perhitungan kesamaan antara dua vektor n dimensi dengan mencari kosinus dari sudut diantara keduanya dan sering digunakan untuk membandingkan dokumen dalam text mining.

C. K-Nearest Neighbor

Algoritma *Nearest Neighbor* merupakan salah satu metode klasifikasi yang digunakan untuk pemecahan masalah pada bidang *Data Mining*. Sama halnya dengan beberapa metode klasifikasi lainnya, algoritma ini memiliki ciri yaitu dengan pendekatan untuk mencari kasus dengan menghitung kedekatan kasus yang baru dengan kasus yang lama. Adapun teknik yang digunakan yaitu berdasarkan bobot dari sejumlah objek kasus yang ada. Metode KNN dikenal juga dengan *lazy learner* (pembelajar malas) karena tidak ada proses belajar (dari data) melainkan belajar dari data (tetangga) terdekat secara langsung pada saat klasifikasi.



Gambar 1 Klasifikasi Berdasarkan Tetangga Terdekat

Terdapat beberapa cara untuk mengukur kedekatan jarak antara data baru (*testing data*) dan data lama (*training data*), diantaranya yaitu euclidean distance dan manhattan distance (city block distance), yang paling sering digunakan adalah euclidean distance (Bramer,2007), dengan rumus:

$$\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_n - b_n)^2}$$

Dimana a = a_1,a_2, …, a_n, dan b = b_1, b_2, …, b_n mewakili n nilai atribut dari dua record.

Untuk mengukur jarak dari atribut yang mempunyai nilai besar, seperti atribut pendapatan, maka perlu dilakukan proses normalisasi. Normalisasi bisa dilakukan dengan rumus min-max normalization atau Z-score standardization (Larose, 2006). Jika data training terdiri dari atribut campuran antara data numerik dan kategori, maka lebih baik menggunakan min-max normalization (Larose, 2006). Untuk menghitung kemiripan kasus, digunakan rumus sebagai berikut (Kusrini, 2009):

$$Similarity\ (p,q) = \frac{\sum_{i=1}^{n} f(pi, qi) Xwi}{wi}$$

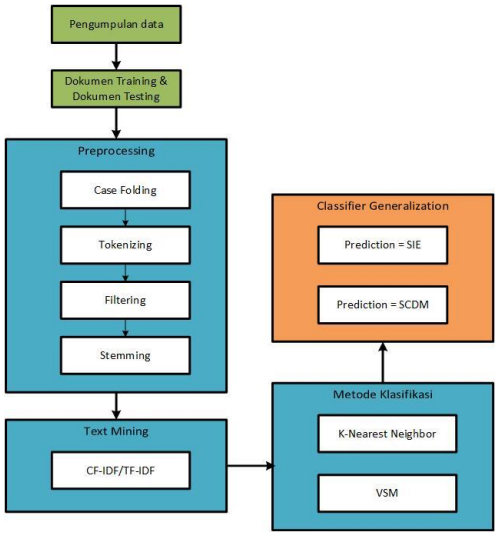
III. METODE PENELITIAN

A. Metode yang Diusulkan

Desain penilitian merupakan suatu rencana kerja yang terstruktur dan komprehensif mengenai hubungan-hubungan antar variable yang disusun sedemikian rupa agar hasil penelitian dapat memberikan jawaban pertanyaan pertanyaan penelitian (Umar, 2008). Adapun metode dan model yang diusulkan dalam penelitin ini yaitu metode K-Nearest Neighbor (KNN) dan Vector Space Model (VSM). Metode KNN digunakan untuk melakukan klasifikasi judul skripsi berdasarkan topik dan peminatan mahasiswa. Model VSM digunakan untuk mengukur kemiripan antara suatu dokumen dengan suatu query. Proses ekstraksi pola serta penemuan informasi dan pengetahuan dari data menggunakan proses *text mining*. Dengan metode dan model yang diusulkan dalam penelitian ini diharapkan dapat membantu pihak Jurusan Informatika Unjani dalam melakukan klasifikasi judul skripsi agar tidak terjadi perbedaan atau tidak adanya kesesuaian antara judul skripsi yang diajukan dengan topik dan peminatan mahasiswa.

B. Kerangka Pemikiran

Kerangka pemikiran dalam penelitian ini dibuat dengan tujuan agar penelitian dapat dilakukan secara bertahap serta konsisten dan merupakan garis besar dari langkah-langkah penelitian yang dilakukan dalam penelitian ini.



Gambar 2 Kerangka Pemikiran

C. Tahap Praprocessing Data

Tahap awal sebelum melakukan proses pengelompokkan dokumen adalah mempersiapkan teks yang ada di dalam dokumen. Pada tahapan preprocessing ini dilakukan beberapa subproses diantaranya yaitu :

1. Tokenizer, yakni proses yang bertujuan untuk memisahkan teks menjadi beberapa token berdasarkan pembatas berupa spasi atau tanda baca. Proses tokenizer akan ditunjukkan pada Tabel 1

Tabel 1 Proses Tokenizer

Teks Input	Teks Output
penjadwalan kuliah fakultas mipa unjani menggunakan algoritma genetika	penjadwalan kuliah fakultas mipa unjani menggunakan algoritma genetika

2. Proses selanjutnya adalah menghilangkan teks yang tidak sesuai dengan teks yang terdapat pada daftar *stopword*, karena teks tersebut dianggap tidak dapat mewakili konten dokumen. Proses tersebut akan ditunjukkan pada Tabel 2

Tabel 2 Proses Stopward

Teks Input	Teks Output
penjadwalan kuliah fakultas mipa Unjani menggunakan algoritma genetika	penjadwalan kuliah genetika unjani algoritma

3. Kemudia pada teks yang masih tersisa dilakukan proses *stemming*, yaitu proses pengubahan teks menjadi bentuk dasarnya. Proses tersebut akan ditunjukkan pada Tabel 3

Tabel 3 Proses Stemming

Teks Input	Teks Output
penjadwalan kuliah genetika Unjani algoritma	jadwal kuliah genetika unjani algoritma

4. Selanjutnya, setiap kata tersebut disebut sebagai *term*, yang nantinya setiap *term* akan didaftarkan dan diberi bobot.s

D. Pembobotan Term Frequency

Setelah tahap praprocessing selesai dilakukan, tahap selanjutnya yaitu menghitung frekuensi kemunculan kata (*term*) pada dokumen uji dan dokumen latih, serta mengitung frekuensi jumlah dokumen yang mengandung kemunculan kata (*Document Frequency*). Tahap selanjutnya, setelah mendapatkan nilai *term frequency* dari setiap konsep per dokumen yaitu mencari nilai bobot *concept frequency*. Setelah mendapatkan nilai dari CF-IDF tahapan selanjutnya yaitu mencari nilai kemiripan antar dokumen dengan menggunakan metode *cosine similarity* dengan menghitung sudut koordinat antar dokumen. Semakin besar nilai yang didapatkan maka dokumen tersebut semakin mirip dengan dokumen uji. Untuk mendapatkan nilai matriks *weighted document* maka nilai CF-IDF dari setiap *term* uji akan dikalikan dengan nilai CF-IDF dokumen pembanding. Untuk setiap kolom dari dokumen akan dihitung nilai kemiripannya nilai matriks WDq x WDi kemudian dihitung Panjang setiap dokumen termasuk dokumen uji (Q) dengan cara mengkuadratkan bobot pada setiap term dalam setiap dokumen, lalu jumlahkan nilai kuadrat dan terakhir diakarkan.

E. *K-Fold Cross Validation*

Penelitian ini menggunakan metode *5-fold cross validation* dengan menggunakan 400 data judul skripsi sebagai data uji dan data latih. Data tersebut dibagi menjadi 5 bagian eksperimen data secara acak. Setiap bagian eksperimen dilakukan 5 percobaan dengan 5 parameter *k* yang berbeda. Parameter tersebut yaitu 3,5,7,9, dan 11. Pada penelitian ini dilakukan *5-fold cross validation* dikarenakan sangat direkomendasikan untuk menentukan parameter *k* yang optimal. Hasil dari percobaan ini akan didapatkan nilai *k* terbaik yang nantinya akan digunakan pada proses klasifikasi *K-Nearest Neighbor*. Pada Tabel 4 merupakan ilustrasi pembagian data pada proses *k-fold cross validation.*

Tabel 4 Ilustrasi Pembagian Data K-Fold Cross Validation

400 Data					
	A (80 data)	B (80 data)	C (80 data)	D (80 data)	E (80 data)
Ekperimen1	A (Data Uji)	BCDE (Data Latih)			
Eksperimen2	A (Data Latih)	B (Data Uji)	CDE (Data Latih)		
Eksperimen3	AB (Data Latih)		C (Data Uji)	DE (Data Latih)	
Eksperimen4	ABC (Data Latih)			D (Data Uji)	E (Data Latih)
Eksperimen5	ABCD (Data Latih)				E (Data Uji)

F. Perhitungan K-Nearest Neighbor

Nilai kedekatan dokumen uji dengan dokumen latih menggunakan metode K-Nearest Neighbor dapat diperoleh dari rumus cosine similarity. Setelah mendapatkan hasil dari perhitungan *cosine similarity*, tahap selanjutnya yaitu mengurutkan dokumen berdasarkan nilai yang paling besar. Nilai yang paling besar memiliki arti bahwa nilai kemiripannya semakin dekat. Setelah tahap pengurutan nilai kedekatan selesai dilakukan, tahap selanjutnya yaitu mencari kelas dari dokumen uji dengan nilai K optimal hasil dari proses *K-Fold Cross Validation*. Diasumsikan nilai K optimal untuk pengujian adalah 3, maka yang diambil adalah 3 nilai kedekatan terbesar hasil perhitungan cosine similarity. 3 nilai kedekatan terbesar tersebut dapat dilihat pada Tabel 5.

Tabel 5 Hasil Perankingan Berdasarkan Nilai K Optimal

Ranking	Dokumen	Cosine Similarity	Kelas (Topik Penelitian)	Peminatan
1.	D2	0.013986	Sistem Keamanan Informasi	SCDM
2.	D3	0.01135	Algoritma Genetika	SCDM
3.	D1	0.003450	Algoritma Genetika	SCDM

IV. HASIL DAN PEMBAHASAN

Objek penelitian ini adalah pengujian terhadap model *Vector Space Model* (VSM) dan metode *K-Nearest Neighbor* (KNN) untuk mengetahui nilai akurasi serta prediksi yang dapat digunakan dalam proses klasifikasi penentuan peminatan berdasarkan judul skripsi yang diajukan oleh mahasiswa. Sumber data yang digunakan sebagai obyek dalam penelitian ini adalah data judul skripsi mahasiswa Informatika UNJANI dari tahun 2010 hingga saat ini.

A. Perbandinga Metode VSM dan KNN

Hasil pengujian metode VSM dan KNN akan dibandingkan untuk mengetahui nilai akurasi tertinggi dari kedua metode tersebut. Perbandingan kedua metode akan dijabarkan pada beberapa Tabel berikut ini:

1. Pengujian dengan Atribut (Judul, keyword, Topik) dan Label (Peminatan).

Tabel 6 Perbandingan 1

Nilai	VSM	KNN
Accuracy	95,29 %	96,08%
Recall	95,28 %	96,07
Precision	95,36 %	96,18

2. Pengujian dengan Atribut (Judul, Keyword, Peminatan) dan Label (Topik).

Tabel 7 Perbandingan 2

Nilai	VSM	KNN
Accuracy	73,83 %	73,79 %
Recall	66,40 %	64,60 %
Precision	68,80 %	70,97 %

3. Pengujian dengan Atribut (Judul, Keyword) dan Label (Peminatan).

Tabel 7 Perbandingan 3

Nilai	VSM	KNN
Accuracy	74,11 %	71,45 %
Recall	74,22 %	71,27 %
Precision	78,14 %	77,90 %

4. Pengujian dengan Atribut (Judul, Keyword) dan Label (Topik).

Tabel 8 Perbandingan 4

Nilai	VSM	KNN
Accuracy	63,62 %	64,37 %
Recall	60,80 %	60,13 %
Precision	67,69 %	71,38 %

5. Pengujian dengan Atribut (Judul, Topik) dan Label (Peminatan).

Tabel 9 Perbandingan 5

Nilai	VSM	KNN
Accuracy	96,85 %	96,85 %
Recall	96,84 %	96,84 %
Precision	96,94 %	96,94 %

6. Pengujian dengan Atribut (Judul, Peminatan) dan Label (Topik).

Tabel 10 Perbandingan 6

Nilai	VSM	KNN
Accuracy	36,65 %	39,25 %
Recall	23,16 %	23,67 %
Precision	9,16 %	9,81 %

7. dengan Atribut (Keyword, Topik) dan Label (Peminatan).

Tabel 11 Perbandingan 7

Nilai	VSM	KNN
Accuracy	95,03 %	96,33 %
Recall	95,03 %	96,32 %
Precision	95,17 %	96,42 %

8. Pengujian dengan Atribut (Keyword, Peminatan) dan Label (Topik).

Tabel 12 Perbandingan 8

Nilai	VSM	KNN
Accuracy	74,09 %	74,84 %
Recall	66,55 %	65,32 %
Precision	68,86 %	71,65 %

9. Pengujian dengan Atribut (Topik) dan Label (Peminatan).

Tabel 13 Perbandingan 9

Nilai	VSM	KNN
Accuracy	96,85 %	96,85 %
Recall	96,84 %	96,84 %
Precision	96,94 %	96,94 %

10. Pengujian dengan Atribut (Peminatan) dan Label (Topik).

Tabel 14 Perbandingan 10

Nilai	VSM	KNN
Accuracy	36,65 %	38,99 %
Recall	23,16 %	23,52 %
Precision	9,16 %	9,77 %

B. Analisa Hasil Pengujian

Pada tahap ini akan dilakukan proses analisa hasil pengujian yang bertujuan untuk mengetahui karakter dari setiap metode. Proses pengujian dilakukan pada setiap kombinasi atribut untuk mengetahui nilai akurasi yang diperoleh dari masing-masing atribut dengan menggunakan metode VSM dan metode KNN.

C. Hasil Prediksi Algoritma VSM

prediction(P...	confidence(...	confidence(...
SCDM	1	0
SCDM	1	0
SCDM	1	0
SIE	0	1

Gambar 3 Hasil Prediksi VSM

Gambar 3 merupakan hasil prediksi menggunakan metode VSM, nilai confidence SCDM dan nilai confidence SIE hanya terdiri dari 2 nilai yaitu 1 dan 0. Jika nilai confidence SCDM = 1 maka data training dikategorikan kedalam kelas SCDM, jika nilai confidence SIE = 1 maka data training dikategorikan kedalan kelas SIE. Nilai confidence yang hanya terdiri dari 1 dan 0 ini disebabkan karena metode VSM dalam menentukan kelas dari suatu data uji berdasarkan nilai jarak terdekat antara dokumen uji dengan dokumen latih. Data uji yang memiliki nilai kedekatan paling besar di beri nilai 1, selain itu bernilai 0.

Dalam penelitian ini juga dilakukan perbandingan pengujian dengan memprediksi label (peminatan) yang memiliki jumlah kelas = 2 dan label (topik) yang memiliki jumlah kelas = 8 seperti yang terlihat pada Tabel 15.

Tabel 15 Perbandingan Label

Label	Atribut data latih	KNN				VSM			
		akurasi	precision	recall	relative error	akurasi	precision	recall	relative error
Peminatan	Topik, Judul, Kata kunci	96.08	96.18	96.07	8.31	95.29	95.36	95.28	4.71
Topik	Peminatan, Judul, Kata Kunci	73.79	70.97	64.60	31.06	73.83	68.80	66.40	26.17

Hasil pengujian tersebut menunjukkan bahwa metode pengklasifikasian KNN menghasilkan nilai akurasi yang lebih besar apabila diterapkan pada label (peminatan) yang memiliki jumlah kelas lebih sedikit dibandingkan dengan label (topik) yang memiliki kelas lebih banyak

D. Hasil Prediksi Algoritma KNN

prediction(PEMINATAN)	confidence(SCDM)	confidence(SIE)
SCDM	1	0
SIE	0.200	0.800
SCDM	1	0
SIE	0.200	0.800

Gambar 4 Hasil Prediksi KNN

Gambar 4 merupakan hasil prediksi menggunakan metode KNN, nilai confidence SCDM dan nilai confidence SIE sangat bervariasi yaitu berada diantara range 0 hingga 1, tidak seperti nilai confidence algoritma VSM yang hanya memiliki 2 nilai yaitu 0 dan 1. Nilai confidence ini diperoleh karena cara kerja dari metode KNN yang mengklasifikasikan data uji berdasarkan jumlah tetangga terdekat, dan tidak menjadikan jarak kedekatan antara dokumen uji dengan dokumen latih sebagai dasar pengklasifikasian.

E. Faktor Selisih Nilai Algoritma

Setelah dilakukan analisa, maka dapat dikatakan bahwa faktor yang mempengaruhi nilai selisih antara algoritma VSM dan algoritma KNN yaitu:

1. Jumlah kelas pada label

Semakin banyak jumlah kelas dalam suatu label, maka nilai akurasi yang dihasilkan dari kedua metode akan semakin menurun.

2. Daftar kata dalam atribut keyword

Penggunaan seluruh kata pada judul skripsi dapat mengakibatkan masuknya dokumen-dokumen yang berbeda kelas ke dalam satu kelas, hal ini dikarenakan metode KNN dan VSM juga memproses kata-kata yang bersifat umum atau tidak signifikan pada saat proses praprocesing data. Semakin banyak daftar kata yan g bersifat umum dalam daftar keyword, maka nilai selisih antara kedua metode akan semakin besar.

3. Atribut Topik

Algoritma KNN dalam mengklasifikasikan suatu data uji menggunakan atribut topik sebagai acuan, sehingga apabila atribut topik dihilangkan, maka nilai akurasi dari metode KNN akan menurun dan mempengaruhi nilai selisih antara algoritma KNN dan VSM.

V. KESIMPULAN DAN SARAN

A. Kesimpulan

Penelitian ini menghasilkan model yang dapat melakukan klasifikasi dan prediksi judul skripsi berdasarkan peminatan mahasiswa pada Jurusan Informatika Unjani. Secara garis besar, terdapat dua jenis model yang dibangun

dengan dua pendekatan yang berbeda, yaitu Vector Space Model dan K-Nearest Neighbor.

Berdasarkan hasil eksperimen, model VSM memiliki akurasi yang lebih rendah dengan perbedaan yang cukup sigifikan jika dibandingkan dengan model yang dihasilkan dari algoritma KNN. Algoritma KNN cukup efektif dalam melakukan klasifikasi dan prediksi pada data yang terdiri dari dua kelas, dan tidak efektif apabila diterapkan pada data yang memiliki banyak kelas. Adapun atribut yang sangat berpengaruh dalam penentuan klasifikasi judul skripsi dipenelitian ini yaitu atribut 'topik' dengan tingkat akurasi mencapai 96,85 %.

B. Saran

Penelitian mengenai Analisa judul skripsi ini masih dapat dikembangkan untuk mendapatkan hasil yang lebih baik. Berikut saran penulis untuk pengembangan lebih lanjut:

1. Menambahkan atribut baru yang berpegaruh terhadap klasifikasi dan prediksi judul skripsi.
2. Melakukan pengujian menggunakan algoritma lain yang belum pernah dilakukan sebelumnya.
3. Pemilihan keyword sebaiknya terbatas pada kata-kata kunci yang signifikan pada penentuan peminatan dalam judul skripsi.

REFERENCES

[1] A. Fitria, M. dan H. Azis, "Analisis Kinerja Sistem Klasifikasi Skripsi Menggunakan Metode Naive Bayes Classifier," *Jurnal Ilmu Komputer dan Teknologi Informasi*, pp. 102 - 106, 2018.

[2] A. F. Hidayatullah dan M. R. Ma'arif, "Penerapan Text Mining dalam Klasifikasi Judul Skripsi," *Seminar Nasional Aplikasi Teknologi Informasi*, pp. A-33 - A-36, 2016.

[3] K. R. Prilianti dan H. Wijaya, "Aplikasi Text Mining untuk Automasi Penentuan Tren Topik Skripsi dengan Metode K-Means Clustering," *Jurnal Cybermatika*, pp. 1 - 6, 2014.

[4] O. Somantri, S. Wiyono dan D. , "Optimalisasi Support Vector Machine (SVM) untuk Klasifikasi Tema Tugas Akhir Berbasis K-Means," *Jurnal Telematika*, pp. 60 - 68, 2016.

[5] R. T. Wahyuni, D. Prastiyanto dan E. Supraptono, "Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF pada Sistem Klasifikasi Dokumen Skripsi," *Jurnal Teknik Elektro*, pp. 18 - 23, 2017.

[6] J. Han, M. Kamber dan J. Pei, Data Mining Concept and Techniques, Elsevier, Morgan Kaufman, 2012.

[7] D. Suyanto, Data Mining untuk Klasifikasi dan Klaterisassi Data, Bandung: Informatika, 2017.

[8] D. Nofriansyah dan G. W. Nurcahyo, Algoritma Data Mining dan Pengujian, Yogyakarta: Deepublish, 2015.

[9] R. T. Wulandari, Data Mining Teori dan Aplikasi Rapidminer, Yogyakarta: Gava Media, 2017.