PENGKLASIFIKASIAN KARYA AKHIR MAHASISWA DENGAN MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBOR (KNN) DI JURUSAN TEKNIK ELEKTRO FAKULTAS TEKNIK UNIVERSITAS NEGERI JAKARTA

Diajukan oleh :

YUNITA ANDRIANI

5235107397

PROGRAM STUDI PENDIDIKAN TEKNIK INFORMATIKA DAN KOMPUTER

JURUSAN TEKNIK ELEKTRO

FAKULTAS TEKNIK

UNIVERSITAS NEGERI JAKARTA

2016

NASKAH PUBLIKASI JURNAL

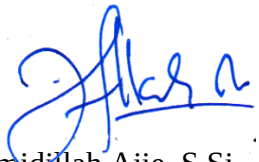
PENGKLASIFIKASIAN KARYA AKHIR MAHASISWA DENGAN MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBOR (KNN) DI JURUSAN TEKNIK ELEKTRO FAKULTAS TEKNIK UNIVERSITAS NEGERI JAKARTA

yang diajukan oleh :

Yunita Andriani
5235107397

Telah disetujui oleh :

Pembimbing I



Hamidillah Ajie, S.Si. , M.T. Tanggal 10 - 02 - 2016
NIP. 19740824 200501 1 001

Pembimbing II

Widodo, M.Kom Tanggal 10 - 02 - 2016
NIP.19720325 200501 1 002

PENGKLASIFIKASIAN KARYA AKHIR MAHASISWA DENGAN MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBOR (KNN) DI JURUSAN TEKNIK ELEKTRO FAKULTAS TEKNIK UNIVERSITAS NEGERI JAKARTA

[1]Yunita Andriani, [2]Hamidillah Ajie, S.Si., M.T, [3]Widodo, M.Kom
[1]Mahasiswi, [2]Dosen Pembimbing 1, [3]Dosen Pembimbing 2
Program Studi S1 Pendidikan Teknik Informatika dan Komputer
Universitas Negeri Jakarta
Email: [1] yunitaandriani71@gmail.com, [2]hamidillah@yahoo.com, [3]widodo03@yahoo.com

Abstrak

Karya akhir ini bertujuan untuk menguji algoritma pengklasifikasi KNN untuk pengembangan sistem klasifikasi otomatis dalam mengkategorikan karya akhir mahasiswa secara otomatis di Jurusan Teknik Elektro UNJ. Adapun proses pengklasifikasian menggunakan Algoritma K-Nearest Neighbor (KNN) dipilih karena algoritma ini cukup sederhana sehingga mudah dalam mengimplementasikannya. Penelitian ini menggunakan bagian abstrak sebagai data untuk mengklasifikasikan. Sebanyak 200 dokumen abstrak sebagai sample dan dikategorikan menjadi 4 (empat) kategori menurut bidang kajiannya, yakni : (1) Bidang Elektro, (2) Bidang Elektronika, (3) Bidang TIK, dan (4) Bidang Pendidikan.
Metode yang digunakan untuk pembobotan term adalah algoritma TF-IDF, untuk menghitung jarak antar dokumen dalam diagram n-dimensi adalah Euclidian Distance, algoritma untuk mengklasifikasikan adalah algoritma KNN, dan metode untuk validasi hasil penelitian menggunakan K-Fold Cross Validation. Hasil dari penelitian ini algoritma KNN dapat mengklasifikasikan karya akhir dengan tingkat akurasi 84,50%.

Kata kunci **:** *Algoritma KNN, Karya Akhir, K-Fold Cross Validation, Klasifikasi, TF-IDF.*

1. Pendahuluan

Karya akhir merupakan salah satu syarat kelulusan bagi mahasiswa dalam mengenyam pendidikan di jenjang Perguruan Tinggi. Karya akhir berisi hasil riset dan penelitian yang telah dilakukan oleh mahasiswa dan dapat dimanfaatkan sebagai sumber informasi yang aktual, sebagai referensi dalam melakukan penelitian lain atau penulisan sebuah jurnal, juga dapat bermanfaat sebagai materi pembelajaran.

Karya akhir ini biasanya disimpan oleh lembaga pendidikan yang bersangkutan ke dalam sebuah repositori. Dokumen dalam bentuk cetak disimpan di perpustakaan, sedangkan dokumen dalam bentuk digital disimpan ke dalam database lembaga pendidikan yang dapat berupa sistem informasi atau situs web.

Untuk memudahkan dalam melakukan proses pencarian, karya akhir diklasifikasikan ke dalam beberapa kategori sesuai dengan bidang kajiannya. Adapun saat ini, umumnya pengklasifikasian dilakukan secara manual (menggunakan manusia), sehingga prosesnya pengklasifikasian tidak efisien. Karena dalam proses klasifikasi karya akhir tersebut, dibutuhkan seseorang yang benar-benar paham mengenai materi yang dibahas di dalam karya-karya akhir tersebut, agar tidak terjadi kesalahan pengklasifikasian.

Sedangkan tiap semesternya, jumlah mahasiswa yang menghasilkan karya akhir tidak sedikit, sehingga membutuhkan sumber daya manusia yang tidak sedikit pula untuk pengklasifikasian karya akhir tersebut.

Salah satu solusi dari kelemahan klasifikasi secara manual adalah dengan membangun sistem klasifikasi dokumen secara otomatis. Dalam membangun sistem klasifikasi dibutuhkan metode atau algoritma yang dapat mengklasifikasikan karya akhir tersebut secara otomatis ke dalam kategori yang ada dengan lebih mudah dan cepat. Ada banyak metode yang dapat digunakan untuk klasifikasi dokumen secara otomatis, salah satunya adalah K-Nearest Neighbor (KNN).

Adapun untuk mengklasifikasikan karya akhir dapat dilihat dari bagian abstrak dari dokumen karya akhir tersebut. Sebab, pada umumnya, bagian abstrak sudah mencerminkan keseluruhan bahasan yang terdapat di dalam karya akhir tersebut.

Berdasarkan permasalahan di atas yang cukup luas ruang lingkupnya, penelitian akan dibatasi pada :

1. Input dokumen berupa file plain text berbahasa Indonesia.
2. Pembobotan term pada data untuk merubah menjadi data matriks menggunakan algoritma TF-IDF
3. Pre-processing data tidak mengalami proses stemming.
4. Metode klasifikasi yang akan digunakan adalah metode K-Nearest Neighbor (KNN).

5. Algoritma yang digunakan pada tahap penghitungan jarak antar titik data latih dan data uji adalah Euclidian Distance.
6. Algoritma yang digunakan untuk proses evaluasi adalah K-Fold Cross Validation.
7. Dokumen yang digunakan diambil dari bagian abstrak dari Karya Akhir Mahasiswa Jurusan Teknik Elektro, Fakultas Teknik di Universitas Negeri Jakarta.
8. Dokumen diklasifikasikan ke dalam 4 kategori, yaitu :
   a. bidang Elektro,
   b. bidang Elektronika,
   c. bidang TIK, dan
   d. bidang Pendidikan.

Berdasarkan latar belakang, identifikasi, dan pembatasan masalah, maka perumusan masalah yang dapat dibuat adalah sebagai berikut : "*Bagaimana file abstrak pada karya akhir mahasiswa Jurusan Teknik Elektro dapat diklasifikasikan dengan metode K-Nearest Neighbor dan bagaimana tingkat akurasinya?*"

1. Kajian Teori

1.1. Data Mining

Sistem komputer modern sudah dapat mengakumulasi data dalam jumlah yang sangat banyak dan memperolehnya dari berbagai sumber. Data yang dimiliki pun menjadi sangat berlimpah. Untuk mengubah data tersebut menjadi informasi yang memiliki nilai guna, dilakukanlah penggalian data atau data mining.

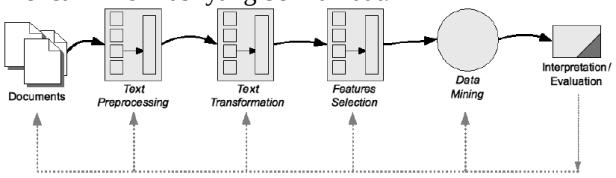
Jenis data yang dapat dilakukan proses data mining terbagi 2 (dua) jenis menurut tingkat kompleksitas data , yaitu :

1. Data dan aplikasi yang sederhana dan datanya teratur:
   a. Relational database,
   b. Data warehouse, dan
   c. Transactional database
2. Data dan aplikasi yang kompleks dan memiliki banyak dimensi :
   a. Data streams and sensor data
   b. Time-series data, temporal data, sequence data
   c. Structure data, graphs, and multi-linked data
   d. Object-relational databases
   e. Heterogeneous databases and legacy databases
   f. Spatial data and spatiotemporal data
   g. Multimedia database
   h. Text databases
   i. The World-Wide Web

Representasi dokumen dalam data mining dapat dilakukan dengan dua cara, yakni cara yang pertama dengan menggunakan tabel relasi. Misalkan tabel relasi antara variabel kata dengan variabel atribut, frekuensi dengan nilai variabel atribut, dan sebagainya. Yang kedua adalah dengan menggunakan digram.

2.2. Text Mining

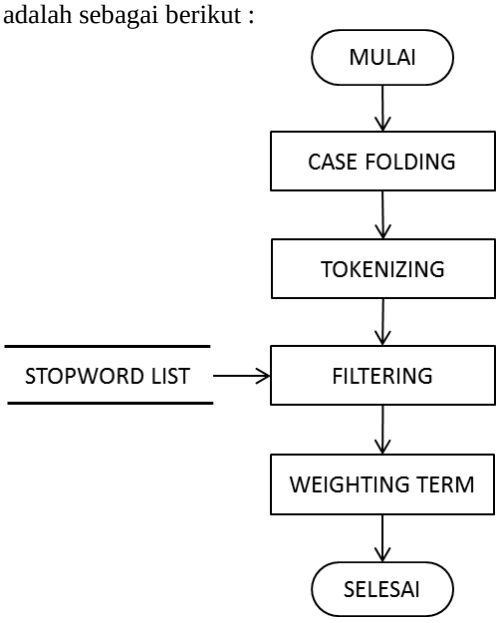
Pendekatan manual text mining secara intensif dalam laboratorium pertama muncul pada pertengahan 1980-an, namun kemajuan teknologi telah memungkinkan ranah tersebut untuk berkembang selama dekade terakhir. Text mining adalah bidang interdisipliner yang mengacu pada pencarian informasi pertambangan data, pembelajaran mesin, statistik, dan komputasi linguistik. Text mining merupakan penerapan konsep dan teknik data mining untuk mencari pola dalam teks melalui proses analisa untuk mencari informasi yang bermanfaat.



Gambar 2.1 Alur Proses Text Mining

2.3. Ekstraksi Data

Agar dapat dilakukan penggalian data, data harus diektraksi melalui tahapan pre-processing data, yang terdiri dari beberapa langkah sebagai berikut, yakni : case folding, tokenizing, filtering / stopword removal, dan stemming. Adapun diagram alur (flowchart) untuk pre-proccessing data pada umumnya adalah sebagai berikut :



Gambar 2.2 Flowchart Pre-processing Data

Tahapan case folding adalah dengan mengubah seluruh huruf dalam dokumen menjadi huruf kecil (lowercase), menghilangkan karakter numerik (angka), tanda baca, operator aritmetika, dan karakter khusus. Tokenizing adalah tahapan pemotongan string berdasarkan kata (term) yang menyusun dokumen yang diinput, menjadi token-token dalam daftar (list). Filtering adalah tahapan untuk menyaring kata (term) yang diperlukan sehingga dapat diolah lebih lanjut dalam proses data mining. Filtering dapat dilakukan dengan dua cara, yaitu dengan menghapus kata-kata

yang tidak penting (stoplist removal) atau memilih kata-kata yang penting (wordlist adding)

Untuk merubah data tersebut agar dapat ditampilkan menjadi diagram vektor, maka term-term yang terdapat dalam data tersebut harus diubah menjadi bentuk matriks. Term-term tersebut akan menentukan letak posisi data tersebut di dalam diagram dengan nilai atau bobot yang dimiliki. Pembobotan (*weighting term*) dapat dilakukan dengan algoritma TF-IDF (Term Frequency-Inverse Document Frequency). Adapun rumus weighting term menggunakan algoritma TF-IDF dapat dilihat pada persamaan 2.1 dibawah ini.

$$\boldsymbol{w_{t,d} = tf_{t,d} \times IDF_{(t)}} \text{ .....(2.1)}$$

Dimana nilai $\boldsymbol{IDF_{(t)}}$didapat dari persamaan 2.2

$$\boldsymbol{IDF_{(t)} = log \frac{|D|}{df_t}} \text{ .....(2.2)}$$

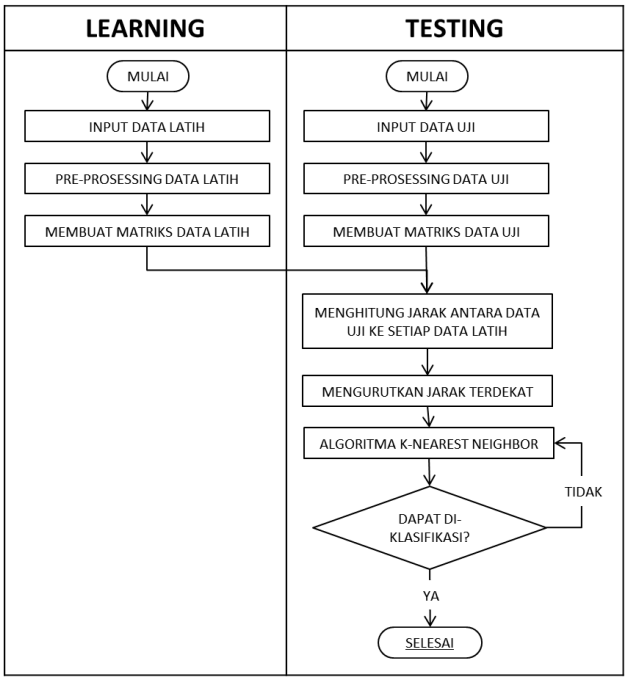
Keterangan :

$w_{t,d}$: bobot term t pada dokumen ke d
$tf_{t,d}$: frekuensi term t pada dokumen ke d
$IDF_{(t)}$: nilai inverse frekuensi dokumen untuk term t
$|D|$: jumlah dokumen yang ada dalam korpus
df_t : jumlah dokumen yang memiliki term t.

2.4. Algoritma *K-Nearest Neighbor* (KNN)

Klasifikasi adalah proses memilah-milah objek yang memiliki kemiripan satu sama lain atau yang memenuhi persyaratan sesuai dengan kategori ke dalam kelompok yang sudah didefinisikan terlebih dahulu yang disebut class. Masing-masing objek dimasukkan hanya ke dalam satu class saja, tidak pernah lebih atau tidak mempunyai class. Class dalam proses klasifikasi sudah ditentukan dan didefinisikan terlebih dahulu di dataset (labeled data).

Dalam proses klasifikasi terdapat dua tahap yang harus dilewati yaitu tahap learning dan testing. Pada tahap learning sebagian data yang telah diketahui kelas datanya (data training) digunakan untuk membentuk model perkiraan. Pada tahap testing, model perkiraan yang sudah terbentuk diuji dengan sebagian data lainnya (data testing) untuk mengetahui akurasi dari model tersebut. Bila akurasinya dapat diterima maka model ini dapat dipakai untuk prediksi kelas data yang belum diketahui. Tahapan proses klasifikasi dengan KNN seperti terlihat pada gambar 2.3



Gambar 2.3 Tahapan proses klasifikasi dengan KNN

2.5. Euclidian Distance

Dalam klasifikasi K-NN yang menggunakan diagram vektor n-dimensi, sebagai representasi data, diperlukan suatu metode untuk menghitung jarak antar titik dalam diagram. Salah satu caranya adalah dengan menggunakan algoritma Euclidian Distance. Rumus Euclidian Distance dapat dilihat pada persamaan 2.2

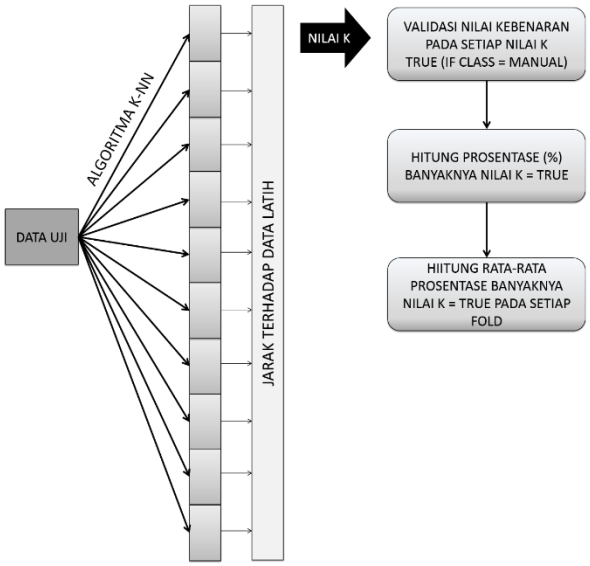
$$D_{euc}(P,Q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \text{ .....(2.2)}$$

Dengan P dan Q adalah instance yang akan diukur jaraknya, p_i dan $q_i$adalah nilai dimensi ke-i yang dimiliki oleh instance P dan Q yang diukur. dalam sebuah ruang vektor n-dimensi.

2.6. K-Fold Cross Validation

K-Fold Cross Validation adalah salah satu metode untuk mengukur tingkat akurasi sebuah model klasifikasi. Mekanisme dari metode Cross Validation cukup sederhana, yakni data dibagi menjadi k bagian sama besar.

Kemudian salah satu bagian menjadi data uji, dan sisanya menjadi data latih. Lalu lanjutkan langkah tersebut untuk semua bagian sebanyak k dengan perbandingan data yang seimbang. Umumnya nilai k yang dipakai adalah 10. Tahapan proses 10-fold cross validation dengan menggunakan model algoritma KNN. Hitung jarak dari setiap data testing terhadap data training. Input parameter nilai k tertentu. Verifikasi hasil klasifikasi setiap nilai k dengan nilai klasifikasi sebenarnya dari data testing . Pada proses akhir dilakukan perhitungan rata-rata tingkat kebenaran atau tingkat eror dari tiap fold ke-n terhadap setiap nilai k.



Gambar 2.7 Ilustrasi K-Fold Cross Validation Dengan K = 10

Tahapan proses algoritma k-fold cross validation dari proses pengolahan fold tersebut dapat dijabarkan dalam tahapan algoritma k-fold cross validation sebagai berikut[1] :

1. Baca dataset terstandarisasi
2. Masukkan nilai fold (F)
3. Masukkan nilai k
4. T = jumlah record dataset
5. S = Jumlah record data testing (T/F)
6. Tentukan L = 1
7. Tentukan M = 0
8. Partisi dataset sebanyak F, tiap partisi sebanyak S record
9. For I = 1 to F
10. F(I) = data testing
11. Not F(I) = data training
12. For N = 1 to S
13. For J = 1 to k
14. Jalankan fungsi algortima k-NN untuk setiap record (N) dalam tabel F(I) untuk nilai k = J
15. P = Hasil prediksi k-NN
16. H = Instance atribut target data testing ke-N
17. If H = P than Nilai = True; else Nilai = False
18. Replace hasil untuk K = J dengan nilai
19. J = J+1
20. Loop (step 13)
21. N = N+1
22. Loop (step 12)
23. While L < k + 1
24. $M_{(I,L)} = \frac{\sum nilai\ false\ untuk\ K=L}{T} \times 100\%$
25. Loop (step 24)
26. I = I + 1
27. Loop (step 9)
28. Selesai

[1] Bramer, Max. Op Cit hal. 83

2. Metodologi Penelitian

Dalam penelitian ini, menggunakan metode penelitian eksperimen adapun perincian metode dan algoritma yang digunakan dalam pengujian algoritma pengklasifikasi K-Nearest Neighbor (KNN) adalah sebagai berikut :

1. Penghitungan bobot kata (*weighting term*) menggunakan algoritma TF-IDF.
2. Penghitungan jarak antar data uji ke semua data latih menggunakan algoritma *Euclidian Distance*.
3. Proses validasi hasil penelitian menggunakan metode *K-Fold Cross Validation* dengan nilai K = 10 fold.

sedangkan untuk pembuatan dan pengembangan perangkat lunak menggunakan metode *waterfall*.

3. Hasil Penelitian

Pengujian algoritma menggunakan metode *K-Fold Cross Validation*. *K-Fold Cross Validation* adalah metode validasi yang membagi data menjadi beberapa fold, umumnya 10 fold. Kemudian dari masing-masing fold tersebut. Untuk pengujian algoritma, menggunakan perbandingan 9 : 1, yakni 180 abstrak menjadi data latih dan 20 abstrak menjadi data uji untuk setiap foldnya. Kemudian diekstraksi menjadi token-token per kata dan melalui tahapan *preprocessing*.

Adapun untuk validasi, hasil klasifikasi melalui aplikasi disesuaikan dengan klasifikasi secara manual dikategorikan oleh dosen ahli menjadi 4 kategori yakni :

1. Bidang Elektro,
2. Bidang Elektronika,
3. Bidang TIK dan
4. Bidang Pendidikan.

Banyaknya nilai k pada algoritma KNN yang digunakan dalam penelitian ini adalah 5. Dengan demikian, 5 data latih yang memiliki jarak yang terdekat dari data uji. Salah satu contoh hasil pengujian algoritma KNN, Fold ke- 1

FOLD KE- 1				
NO	ID_SKRIPSI	KLASIFIKASI MANUAL	KLASIFIKASI KNN	SESUAI / TIDAK SESUAI
1	1	PENDIDIKAN	PENDIDIKAN	SESUAI
2	2	PENDIDIKAN	PENDIDIKAN	SESUAI
3	3	PENDIDIKAN	PENDIDIKAN	SESUAI
4	4	PENDIDIKAN	PENDIDIKAN	SESUAI
5	5	PENDIDIKAN	PENDIDIKAN	SESUAI
6	6	ELEKTRO	ELEKTRO	SESUAI
7	7	ELEKTRO	ELEKTRO	SESUAI
8	8	PENDIDIKAN	PENDIDIKAN	SESUAI
9	9	PENDIDIKAN	PENDIDIKAN	SESUAI
10	10	ELEKTRONIKA	ELEKTRONIKA	SESUAI
11	11	PENDIDIKAN	PENDIDIKAN	SESUAI
12	12	PENDIDIKAN	PENDIDIKAN	SESUAI
13	13	ELEKTRO	ELEKTRO	SESUAI
14	14	ELEKTRO	ELEKTRO	SESUAI
15	15	ELEKTRO	ELEKTRO	SESUAI
16	17	ELEKTRONIKA	ELEKTRONIKA	SESUAI
17	18	ELEKTRONIKA	ELEKTRONIKA	SESUAI
18	20	ELEKTRONIKA	ELEKTRONIKA	SESUAI
19	25	TIK	PENDIDIKAN	TIDAK SESUAI
20	97	TIK	PENDIDIKAN	TIDAK SESUAI

Tabel 4.1 Hasil Fold-1 terdapat 18 dokumen yang sesuai dengan klasifikasi manual, sedangkan 2 dokumen tidak sesuai.

Penghitungan tingkat akurasi Fold ke-1 adalah sebagai berikut :

$$= \frac{\text{Jumlah dokumen yang sesuai}}{\text{jumlah dokumen pada fold}} \times 100\%$$

$$= \frac{18}{20} \times 100\%$$

= 90 %

Penghitungan tingkat error Fold ke-1 adalah sebagai berikut :

$$= \frac{\text{Jumlah dokumen yang tidak sesuai}}{\text{jumlah dokumen pada fold}} \times 100\%$$

$$= \frac{2}{20} \times 100\%$$

= 10 %

Adapun nilai rata-rata tingkat akurasi dan error dalam penelitian ini dijabarkan di dalam Tabel 4.2 sebagai berikut :

Tabel 4.2 Tabel nilai akurasi dan error dari pengujian algoritma

FOLD KE -	AKURASI (%)	ERROR (%)
1	90,00%	10,00%
2	90,00%	10,00%
3	80,00%	20,00%
4	85,00%	15,00%
5	85,00%	15,00%
6	80,00%	20,00%
7	70,00%	30,00%
8	90,00%	10,00%
9	85,00%	15,00%
10	90,00%	10,00%
RATA-RATA	**84,50%**	**15,50%**

4.2 Pembahasan

Data dalam sistem klasifikasi dokumen menggunakan algoritma KNN ini adalah sebagai berikut :

1. Total seluruh dokumen abstrak yang digunakan bersumber dari Jurusan Teknik Elektro Universitas Negeri Jakarta berjumlah 200 dokumen.
2. Pengujian sistem algoritma menggunakan metode K-Fold Cross Validation dimana 200 dokumen abstrak dibagi menjadi ke dalam 10 bagian (fold) dengan perbandingan data latih dan data uji sebesar 9 : 1.
3. Hasil pengujian dan besaran angka akurasi dijabarkan sebagai berikut
   a. Uji Fold ke – 1 :
      Data latih sebanyak 180 dokumen dan data uji sebanyak 20 dokumen, dengan tingkat akurasi 90% dan tingkat error 10%.
   b. Uji Fold ke – 2 :
      Data latih sebanyak 180 dokumen dan data uji sebanyak 20 dokumen, dengan tingkat akurasi 90% dan tingkat error 10%.
   c. Uji Fold ke – 3 :
      Data latih sebanyak 180 dokumen dan data uji sebanyak 20 dokumen, dengan tingkat akurasi 80% dan tingkat error 20%.
   d. Uji Fold ke – 4 :
      Data latih sebanyak 180 dokumen dan data uji sebanyak 20 dokumen, dengan tingkat akurasi 85% dan tingkat error 15%.
   e. Uji Fold ke – 5 :
      Data latih sebanyak 180 dokumen dan data uji sebanyak 20 dokumen, dengan tingkat akurasi 85% dan tingkat error 15%.
   f. Uji Fold ke – 6 :
      Data latih sebanyak 180 dokumen dan data uji sebanyak 20 dokumen, dengan tingkat akurasi 80% dan tingkat error 20%.
   g. Uji Fold ke – 7 :
      Data latih sebanyak 180 dokumen dan data uji sebanyak 20 dokumen, dengan tingkat akurasi 70% dan tingkat error 30%.
   h. Uji Fold ke – 8 :
      Data latih sebanyak 180 dokumen dan data uji sebanyak 20 dokumen, dengan tingkat akurasi 90% dan tingkat error 10%
   i. Uji Fold ke – 9 :
      Data latih sebanyak 180 dokumen dan data uji sebanyak 20 dokumen, dengan tingkat akurasi 85% dan tingkat error 15%.

j. Uji Fold ke – 10 :
Data latih sebanyak 180 dokumen dan data uji sebanyak 20 dokumen, dengan tingkat akurasi 90% dan tingkat error 10%.

Dari hasil pengujian 10 fold di atas, didapatkan nilai rata-rata akurasi sebesar 84,50% dan nilai rata-rata error 15,50%.

4. Dokumen diklasifikasikan menjadi 4 *class* yakni : Elektro, Elektronika, TIK dan Pendidikan.

Berdasarkan hasil penelitian ini, nilai akurasi terbesar adalah 90% berada di fold 1, 2, 8, dan 10. Sedangkan nilai akurasi terendah berada di fold 7 dengan nilai akurasi sebesar 70% dan nilai error 30%. Nilai akurasi pada algoritma KNN sangat bergantung pada perbandingan yang seimbang antara dokumen di masing-masing kategori dan banyaknya dokumen sebagai data latih. Semakin banyak dokumen yang menjadi data latih semakin akurat pula hasil klasifikasinya. Dan semakin seimbang jumlah dokumen per kategori di tiap fold, semakin akurat pula hasil klasifikasinya.

Namun dari data yang didapat hanya memiliki 200 dokumen dengan informasi lengkap mengenai nama penulis, nomor registrasi, judul, dan tahun lulus. Adapun kategori dari 200 dokumen tersebut adalah kategori "ELEKTRO" sebanyak 48 dokumen, kategori "ELEKTRONIKA" sebanyak 46 dokumen, "TIK" sebanyak 21 dokumen, dan kategori "PENDIDIKAN" dengan jumlah terbesar sebanyak 85 dokumen.

Fold	Bidang Elektro	Bidang Elektronika	Bidang TIK	Bidang Pendidikan
1	5	4	2	9
2	5	5	2	8
3	4	5	2	9
4	5	4	2	9
5	5	5	2	8
6	5	4	2	9
7	4	5	3	8
8	5	5	2	8
9	5	4	2	9
10	5	5	2	8
Total	48	46	21	85

Tabel 4.3 Perbandingan Kategori Manual dalam Fold

5. Kesimpulan dan Saran

5.1. Kesimpulan

Berdasarkan hasil penelitian pada Bab IV, maka dapat disimpulkan bahwa

1. Algoritma KNN mampu mengklasifikasikan dokumen karya akhir dengan cara diimplementasikan ke dalam sebuah aplikasi yang dapat mengklasifikasikan karya akhir secara otomatis dengan meng-input data abstrak dari karya akhir tersebut.
2. Proses klasifikasi karya akhir termasuk ke dalam ranah Text Mining yang datanya tidak terstruktur. Sehingga untuk membuat data teks tersebut lebih terstruktur, dokumen tersebut harus melalui proses pre-processing.
3. Tingkat akurasi dari pengklasifikasian pada pengujian algoritma KNN dapat diketahui dengan proses validasi dengan metode K Cross Validation. Adapun hasil validasi dapat dilihat pada Tabel 5.1. Nilai rata-rata akurasi sebesar 84,50% dan nilai errornya 15,50%.

5.2. Saran

Dalam penyusunan karya akhir mengenai klasifikasi dokumen berdasarkan abstrak karya akhir dengan menggunakan algoritma K-Nearest Neighbor banyak ditemui hambatan yaitu data yang didapatkan dari sumber, banyak yang tidak terstruktur baik dari segi pemberkasan yang tidak sama strukturnya, berkas softcopy yang tidak lengkap, informasi penulis, ataupun banyaknya dokumen karya akhir yang bisa dijadikan sample.

Maka diharapkan dalam penelitian selanjutnya, sistemasi karya akhir agar lebih terstruktur sehingga memudahkan dalam mengekstraksi set data latih dan set data uji serta banyaknya dokumen yang dijadikan data latih maupun data uji, harus berbanding seimbang agar dapat memiliki nilai akurasi yang lebih tinggi sehingga dapat mengklasifikasikan dokumen jauh lebih baik.

Daftar Pustaka

Berry, Michael W. & Jacob Kogan. 2010. Text Mining : Application and Theory. USA : Wiley

Brammer, Max. 2007. Principles of Data Mining. London : Springer.

Han, J. & Kamber, M. 2006. Data Mining : Concepts and Technique, 2nd Edition. San Fransisco : Morgan Kauffman Pubisher.

Tan. 2012. Data Mining – Konsep dan Aplikasi Menggunakan Matlab. Yogyakarta: Penerbit Andi

Kamber, Micheline. 2001. Data Mining Concepts and Techniques Second Edition. San Francisco: Morgan Kauffman.

Kusrini, &Emha Taufiq Luthfi. 2009. Algoritma Data Mining. Yogyakarta : Penerbit Andi.

Richard J Roger & Michael W Geatz , Data Mining Tutorial-Based Primer, United State of America : Pearson Education Inc, 2003

http://www.mohamedrabeea.com/books/book1_1165.pdf

http://ecatatan.wordpress.com/2013/05/22/k-nearest-neighbor/

http://kuliahinformatika.wordpress.com/2010/02/13/buku-ta-k-nearest-neighbor-knn/

http://yudiagusta.wordpress.com/2009/01/13/feature-selection/