JISKa, Vol. 4, No. 1, Mei, 2019, Pp. 69 – 75
ISSN : 2527 – 5836 (print) | 2528 – 0074 (online)

KLASIFIKASI DOKUMEN TUGAS AKHIR (SKRIPSI) MENGGUNAKAN *K-NEAREST NEIGHBOR*

Kitami Akromunnisa [(1)], Rahmat Hidayat [(2)]
Jurusan Teknik Informatika
Jalan Laksda Adisucipto, Yogyakarta (55281)
e-mail : kitamiakr@gmail.com [(1)], rahmat.hidayat@uin-suka.ac.id [(2)]

Abstract

Various scientific works from academicians such as theses, research reports, and practical work reports, are available in the digital version. However, this phenomenon is not accompanied by a growth of information or knowledge extracted from these electronic documents. This study aims to classify the abstract data of the informatics engineering thesis. The algorithm used in this study is the K-Nearest Neighbor. Amount of data used 50 abstract data of Indonesian language, 454 data of English abstract, and 504 title data. Each data is divided into training data and test data. Test data will be classified automatically with the classifier model that has been made. Based on the research conducted, the classification of the Indonesian essential data resulted in higher accuracy without going through a stemming process that had a 9: 1 ratio of 100.0% compared to an 8: 2 ratio of 90.0%, 7: 3 which was 80.0%, 6: 4 which is 60.0% and the data distribution using K-fold cross-validation is 80.0%.

Keywords : 3 – 5 keywords (Arial, 10 pt)

Abstrak

Berbagai karya ilmiah dari sivitas akademika seperti skripsi, laporan penelitian, laporan kerja praktek dan lain sebagainya telah tersedia dalam versi digital. Namun, pada umumnya fenomena ini tidak disertai dengan pertumbuhan jumlah informasi atau pengetahuan yang dapat disarikan dari dokumen-dokumen elektronik tersebut. Penelitian ini bertujuan untuk melakukan klasifikasi pada data abstrak skripsi teknik informatika. Algoritma yang digunakan dalam penelitian ini adalah K-Nearest Neighbor. Jumlah data yang digunakan 50 data abstrak bahasa Indonesia, 454 data abstrak bahasa inggris dan 504 data judul. Masing-masing data dibagi menjadi data latih dan data uji. Data uji akan diklasifikasikan otomatis dengan model *classifier* yang telah dibuat. Berdasarkan penelitian yang dilakukan, klasifikasi data intisari bahasa Indonesia menghasilkan akurasi lebih besar tanpa melalui proses *stemming* yang memiliki perbandingan antara data latih dan data uji sebesar 9:1 yaitu 100,0 % dibandingkan dengan perbandingan sebesar 8:2 yaitu 90,0%, 7:3 yaitu 80,0%,6:4 yaitu 60,0% serta pembagian data menggunakan *K-fold cross validation* yaitu 80,0%.

Kata Kunci : Klasifikasi, abstrak, judul skripsi, stemming, k-nearest neighbor

1. PENDAHULUAN

Semakin pesat dan murahnya perkembangan teknologi media penyimpanan digital telah mendorong terjadinya ledakan jumlah dokumen elektronik yang tersimpan dalam *repository* perpustakaan universitas. Berbagai karya ilmiah dari sivitas akademika seperti skripsi, laporan penelitian, laporan kerja praktek dan lain sebagainya telah tersedia dalam versi digital. Namun, pada umumnya fenomena ini tidak disertai dengan pertumbuhan jumlah informasi atau pengetahuan yang dapat disarikan dari dokumen-dokumen elektronik tersebut (N. Gupta, 2012) . Metode *Text Mining* merupakan pengembangan dari metode *data mining* yang dapat diterapkan untuk mengatasi masalah tersebut. Algoritma-algoritma dalam *text mining* dibuat untuk dapat mengenali data yang sifatnya semi terstruktur misalnya sinopsis, abstrak maupun isi dari dokumen-dokumen (V. Gupta, Lehal, & others, 2009)

Abstrak adalah representasi yang ringkas tetapi akurat isi suatu dokumen. Ia membedakan abstrak dari *extract*, karena sebuah *extract* adalah versi singkat dari sebuah dokumen yang dibuat dengan jalan mengambil kalimat-kalimat dari dokumen tersebut. Sedangkan abstrak, walaupun memakai berbagai kalimat yang ada dalam dokumen, merupakan sepenggal teks

yang diciptakan oleh pembuat abstrak, bukan kutipan langsung dari penulisnya (Lancaster, 1991) .

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)*, *Naïve Bayes*, *k-Nearest Neighbor* (KNN), *Decision Tree*, dan *Artificial Neural Networks (ANN*).

Beberapa aplikasi *text mining* telah diterapkan di perpustakaan terutamanya untuk pencarian bahan pustaka berbasis teks (Wiguna, 2011). Meskipun demikian belum banyak yang dikembangkan untuk tujuan analisis. Sehingga sangatlah sulit untuk dapat dengan segera mengetahui topik penelitian populer ataupun kecenderungan minat penelitian mahasiswa program studi tertentu misalnya. Efendi dan Mustakim (2017) melakukan penelitian dengan judul *Text Mining Classification* Sebagai Rekomendasi Dosen Pembimbing Tugas Akhir Program Studi Sistem Informasi, menghasilkan kesimpulan *text Mining* dengan algoritma klasifikasi yaitu *K-Nearest Neighbor(KNN)* untuk rekomendasi dosen pembimbing tugas akhir (Efendi & Mustakim, 2017).
Penelitian yang dilakukan oleh Prilianti dan Wijaya (2014) dengan judul Aplikasi *Text Mining* untuk Otomasi Penentuan Tren Topik Skripsi dengan *Metode K-Means Clustering* menghasilkan kesimpulan algoritma *k-means clustering* yang digunakan dalam proses penemuan 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 (Prilianti & Wijaya, 2014). Penelitian terkait lainnya dilakukan oleh Hidayatullah dan Ma'arif (2016), melakukan klasifikasi judul skripsi menggunakan *Naïve Bayes* dan *Support Vector Machine*, menghasilkan kesimpulan bahwa *Naïve Bayes* memiliki akurasi yang lebih tinggi dibandingkan dengan *Support Vector Machine*. Pada penelitian tersebut, algoritma SVM menggunakan *linear kernel*, sedangkan data skripsi merupakan data *multi dimensi* yang akan lebih optimal menggunakan *multiclass classification (Hidayatullah & Ma'arif, 2016)*.

Dalam penulisan makalah ini terdiri dari lima bab.bagian pertama merupakan pendahuluan yang memuat latar belakang dari penelitian ini. Penelitian-penelitian sebelumnya yang terkait dan mendukung penelitian ini dijelaskan pada bagian kedua. Bagian ketiga, dibahas metode yangdigunakan dalam penelitian ini. Bagian keempat memaparkan hasil eksperimen dan pembahaan dari hasil yang diperoleh. Bagian kelima merupakan bagian terakhir yang berisi kesimpulan dari penelitian.

Proses *pre-processing* terdapat tahap stemming, pada penelitian yang dilakukan oleh Hidayatullah (2015). Pada penelitia tersebut menghasilkan perbedaan akurasi antara *pre-processing* dengan stemming dan tanpa menggunakan stemming. Hasilnya adalah data yang tanpa melalui tahap stemming memiliki tingkat akurasi yang lebih tinggi daripada data yang melalui proses stemming (Hidayatullah, 2015).

2. METODE PENELITIAN

2.1 Data Penelitian

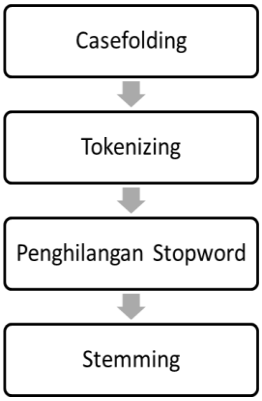
Penelitian ini menggunakan data abstrak dan judul skripsi mahasiswa Teknik Informatika tahun 2010 sampai dengan 2018 sebanyak 504 data. Data yang akan digunakan adalah data intisari yang berbahasa Indonesia, yaitu 50 data, 454 data bahasa inggris dan 504 data judul. data tersebut masing-masing dibagi dalam dua kelompok yaitu kelompok data training dan data testing. semua data telah diberi label yang terdiri dari kelas AI *(Artificial Intelligen)*, SI (Sistem Informasi), Jaringan dan RPL (Rangkaian Perangkat Lunak)

2.2 Seleksi data

Data yang terkumpul dilakukan penyeleksian terhadap data abstrak dan judul untuk menentukan data yang digunakan dalam tahap selanjutnya. Kemudian dilakukan pembagian data menjadi dua bagian yaitu data *training* dan data *testing.* Penelitian ini menggunakan dua cara yaitu *Train Test Split* dan *K-Fold.* Untuk *Train Test Split* terdapat berbagai rasio pembagian yaitu 9:1, 8:2, 7:3, dan 6:4.

2.3 *Pre-processing*

Pre-processing dalam text mining bertujuan untuk mempersiapkan data sebelum diproses pada langkah selanjutnya. Dalam penelitian ini untuk data yang menggunakan bahasa Indonesia akan membandingkan pengaruh stemming pada tahap *pre-processing,* maka dilakukan dua tahap *pre-procesing* yang berbeda. Urutan tahapan *pre-processing* yang dilakukan menggunakan stemming dapat dilihat pada Gamba 1.



Gambar 1. Tahapan pre-processing dengan stemming

2.4 Pemilihan dan Ekstraksi Fitur

1. *Term Frequency-Inverse Document Frequency (TF-IDF)*
   Term frequency bertujuan untuk menghitung kemunculan suatu *term* dalam suatu *corpus* berdasarkan bobot suatu *term* pada dokumen tertentu. Dalam suatu dokumen, apabila *term* tertentu memiliki kemunculan yang tnggi, maka akan semakin tinggi bobot dokumen untuk *term* tersebut, dan sebaliknya. *Inverse document frequency* berfungsi mengurangi bobot suatu term jika kemunculannya banyak tersebar di seluruh koleksi dokumen. Setiap kata dari data layih dan data uji akan dihiung bobotnya untuk merepresentasikan kata tersebut kedalam angka agar dapat diproses dengan algoritma KNN.

2.5 Klasifikasi

Penelitian ini menggunakan metode *Holdout* dan metode *k-fold* untuk membagi antara data *training* dengan data *testing.* untuk metode *Holdout* menggunakan rasio pembagian 9:1, 8:2, 7:3, dan 6:4. Sedangkan pembagian data menggunakan *K-fold* menggunakan k=10. Proses klasifikasi menggunakan algoritma K-NN dengan menentukan k = 3,5,7 dan 9.

3. HASIL DAN PEMBAHASAN

Berdasarkan hasil ekseprimen yang diperlihatkan oleh Tabel 1, Tabel 2, Tabel 3 diketahui bahwa hasil akurasi yang menggunakan *k-fold cross validation* menghasilkan nilai akurasi 80,0%, sedangkan akurasi yang menggunakan *Split into Train Test Sets* menghasilkan nilai akurasi pada pembagian data dengan perbandingan 9:1 menghasilkan akurasi 80,0% , pembagian data dengan perbandingan 8:2 nilai akurasi 90,0%. Pada pembagian data dengan perbandingan 7:3 menghasilkan nilai akurasi 80,0%. Sedangkan untuk pembagian data dengan perbandingan 6:4 menghasilkan nilai akurasi 60%. Untuk hasil akurasi tersebut menggunakan nilai k = 5.

Tabel 1. hasil akurasi prediksi pada data intisari bahasa indonesia

Intisari Bahasa Indonesia				
Partisi Data		k	Stemming	Tanpa stemming
K-Fold	10	5	40	80
Split Train Test Sets	6:4	1	60	60
		3	55	55
		5	60	60
		7	60	60
		9	70	70
Split Train Test Sets	7:3	1	53.3	60
		3	60	60
		5	66.6	80
		7	66.6	73.3
		9	60	66.6
Split Train Test Sets	8:2	1	70	70
		3	60	60
		5	60	90
		7	70	90
		9	80	90
Split Train Test Sets	9:1	1	80	100
		3	80	80
		5	80	80
		7	80	80
		9	80	80

Tabel 2. hasil akurasi prediksi pada data judul skripsi

Judul				
Partisi data		k	Stemming	Tanpa stemming
K-Fold	10	10	80.3	86.4
Split Train Test Sets	6:4	1	68	68.5
		3	71	71
		5	70	71
		7	71	72
		9	72.4	70
Split Train Test Sets	7:3	1	61.9	62.5
		3	65.4	65.1
		5	68.3	67.7
		7	68.3	67.7

Judul				
Partisi data		k	Stemming	Tanpa stemming
		9	69	67.7
Split Train Test Sets	8:2	1	62.5	61.5
		3	64.2	64.2
		5	63.4	63.4
		7	66.3	65.4
		9	65.3	63.4
Split Train Test Sets	9:1	1	63	63.4
		3	67.3	67.3
		5	65.3	80
		7	65.3	79
		9	69.2	78.2

Tabel 3. hasil akurasi prediksi pada data abstrak bahasa inggris

Abstrak Bahasa Inggris			
Partisi data		k	Nilai akurasi
K-Fold	10	9	76,7
Split Train Test Sets	7:3	1	55,0
		3	57,4
		5	80
		7	66,6
		9	65,1
Split Train Test Sets	8:2	1	52,3
		3	58,1
		5	60,4
		7	62,7
		9	59,3
Split Train Test Sets	9:1	1	99,4
		3	7.5
		5	77.6
		7	72
		9	73.9
Split Train Test Sets	6:4	1	55.2
		3	62.2
		5	62.2
		7	69
		9	71

KESIMPULAN

Berdasarkan penelitian yang telah dilakukan didapatkan kesimpulan bahwa klasifikasi menggunakan metode *k-nearest neighbor* bisa digunakan untuk mengklasifikasi data intisari bahasa Indonesia dan judul dengan akurasi yang lebih besar tanpa melalui proses *stemming.* Untuk partisi data yang menggunakan pembagian data *Split into train test sets* dengan rasio perbandingan 9:1 menghasilkan akurasi lebih besar dibandingkan dengan rasio perbandingan 6:4, 7:3, 8:2 dan pembagian data menggunakan *kfold cross validation*. Dari hasil yang diperoleh maka dapat disimpulkan bahwa semakin besar data latih akan semakin baik akurasinya**.**

DAFTAR PUSTAKA

Efendi, Z., & Mustakim, M. (2017). Text Mining Classification Sebagai Rekomendasi Dosen Pembimbing Tugas Akhir Program Studi Sistem Informasi. In *Seminar Nasional Teknologi Informasi Komunikasi dan Industri* (pp. 235–242).

Gupta, N. (2012). Text mining for information retrieval.

Gupta, V., Lehal, G. S., & others. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence*, *1*(1), 60–76.

Hidayatullah, A. F. (2015). The Influence of Indonesian Stemming on Indonesian Tweet Sentiment Analysis. In *Proceeding of International Conference on Electrical Engineering, Computer Science and Informatics (EECSI 2015). Palembang, Indonesia* (Vol. 2, pp. 182–187).

Hidayatullah, A. F., & Ma'arif, M. R. (2016). Penerapan Text Mining dalam Klasifikasi Judul Skripsi. *Jurnal Fakultas Hukum UII*.

Lancaster, F. W. (1991). *Indexing and abstracting in theory and practice*. Library Association London.

Prilianti, K. R., & Wijaya, H. (2014). Aplikasi text mining untuk automasi penentuan tren topik skripsi dengan metode K-Means Clustering. *Jurnal Cybermatika*, *2*(1).

Wiguna, I. (2011). *LKP: Aplikasi Katalog Online untuk Pencarian Konten Buku dengan Metode Text Mining pada Perpustakaan Stikom Surabaya*. STIKOM Surabaya.

Yuono, F. (2005). *Pembuatan Aplikasi Mining untuk Pencarian Buku Koleksi Skripsi dengan Menggunakan Association Rules Analysis, Skripsi, Universitas Kristen Petra.* Universitas Kristen Putra.