e-ISSN 2338-5197

Penerapan *Text Mining* Pengelompokkan Judul Kerja Praktek Menggunakan Metode *K-Means Clustering* dengan *Cosine Similarity*

Ika Kurnia Saputri[a,1], Tedy Setiadi[a,2] Lisna zahrotun [a,3]

[a] Program Studi Teknik Informatika, Universitas Ahmad Dahlan,
Prof. Dr. Soepomo, S.H., Janturan, Umbulharjo, Yogyakarta 55164
[1]ika.putri2801@gmail.com; [2] tedy.setiadi@tif.uad.ac.id; [3]lisna.zahrotun@tif.uad.ac.id

Abstrak

Kerja Praktek adalah kegiatan mahasiswa yang dilakukan di masyarakat maupun di perusahaan atau instansi untuk mengaplikasikan ilmu yang diperoleh dan melihat relevansinya di masyarakat maupun melalui jalur pengembangan diri dengan mendalami bidang ilmu tertentu dan aplikasinya. Dalam pelaksanaanya, tidak sedikit mahasiswa bingung menentukan sebuah instansi, perusahaan ataupun tempat lain untuk dijadikan tempat Kerja Praktek. Oleh karena itu, perlu adanya pengelompokkan Judul Kerja Praktek sehingga dapat menjadi salah satu referensi pengetahuan bagi mahasiswa untuk mengetahui pola kelompok judul kerja praktek yang ada. Dalam pengelompokkan tersebut, dapat menggunakan metode Text Mining K-Means Clustering dengan Cosine Similarity untuk dapat mengelompokkan judul kerja praktek. Penelitian ini dilakukan untuk mengkaji tentang algoritma K-Means Clustering dengan Cosine Similarity dan mengimplementasikan algoritma K-Means Clustering dengan Cosine Similarity dengan melakukan tahapan tokenizing, filtering dan stemming sehingga pada akhirnya akan didapatkan cluster-cluster judul kerja praktek. Data yang digunakan adalah data judul kerja praktek Teknik Informatika Universitas Ahmad Dahlan sebanyak 355 data. Hasil pengujian dilakukan uji purity sebanyak 5 kali percobaan, dengan mengkombinasikan parameter M yang berbeda-beda sebagai titik pusat cluster diperoleh nilai terbaik sebesar 0,85 dengan kombinasi M=6 yang artinya semakin mendekati 1 mengindikasikan bahwa semakin banyak dokumen yang berhasil dikelompokkan dengan benar.

Kata Kunci: *Text Mining, K-Means Clustering, Cosine Similarity.*

1. Pendahuluan

Text mining adalah salah satu teknik yang dapat digunakan untuk melakukan klasifikasi, dimana *text mining* merupakan variasi dari data *mining* yang berusaha menemukan pola yang menarik dari sekumpulan data tekstual yang berjumlah besar. Selain klasifikasi, *text mining* juga digunakan untuk menangani masalah *clustering*, *information extraction*, dan *information retrival* [1] .

Algoritma *K-Means* merupakan metode yang digunakan pada teknik *clustering*. *K-Means* adalah salah satu algoritma *unsupervised learning* yang paling sederhana yang dikenal dapat menyelesaikan permasalahan *clustering* dengan baik. Ide utamanya adalah mendefinisikan *centroid* sejumlah K untuk masing-masing klaster [2].

Data yang digunakan dalam penelitian ini berupa data teks judul kerja praktek. Kerja praktek adalah kegiatan mahasiswa yang dilakukan di masyarakat maupun di perusahaan atau instansi untuk mengaplikasikan ilmu yang diperoleh dan melihat relevansinya di masyarakat maupun melalui jalur pengembangan diri dengan mendalami bidang ilmu tertentu dan aplikasinya. Kerja praktek umumnya mempunyai bobot 2 (dua) SKS dan dilaksanakan dalam kurun waktu 1-3 bulan, disesuaikan dengan kebijakan fakultas. Setelah melaksanakan kerja praktek, mahasiswa akan menyerahkan judul kerja praktek tersebut dibagian Tata Usaha. Namun, dalam pelaksanaanya dibagian Tata Usaha judul kerja

e-ISSN 2338-5197

praktek tersebut hanya didata dan belum pernah dilakukan pengolahan, dilakukan evaluasi terhadap judul kerja praktek tersebut.

Dengan melihat data Judul Kerja Praktek yang ada di Teknik Informatika dan kemampuan dari metode *K-Means Clustering*, maka diharapkan ada suatu aplikasi yang dapat mengelompokkan Judul Kerja Praktek yang ada di Teknik Informatika.

2. Kajian Pustaka

Penelitian ini mengacu pada penelitian terdahulu [2]. Pada penelitian ini membahas problem pengelompokkan dokumen dengan metode *cluster*. Pada penelitian ini dibahas juga analisa perbandingan *cluster* data dengan mengkolaborasikan metode *k-means* dengan metode hierarki untuk penentuan pusat awal *cluster*.

Penelitian ini juga mengacu pada penelitian [3]. Pada penelitian ini telah berhasil mengembangkan sebuah aplikasi yang mengimplementasikan *cosine similarity* dan algoritma *smith-waterman* untuk mendeteksi kemiripan teks. Langkah-langkah pendekteksian kemiripan teks dilakukan dengan melalui tahap pertama *preprocessing* yaitu *case folding* dan *tokenizing, filtering, stemming*. penelitian ini dapat mendeteksi tingkat kemiripan teks dari sangat mirip hingga sagat tidak mirip berdasarkan kemunculan kata di dalamnya dengan menggunakan *cosine similarity*.

Pada penelitian ini akan dikembangkan sebuah program yang menjelaskan tahapan *text mining*, dapat menghasilkan pola sebagai pengetahuan, dan dapat menampilkan hasil pengelompokkan dari penerapan metode *k-means clustering* dengan *cosine similarity*.

2.1. *Text Mining*

Text mining adalah salah 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 [4].

2.2. *Clustering*

Cluster adalah kumpulan dari objek yang mirip dengan objek lainnya dan berada pada kelompok yang sama. Sedangkan proses untuk mengelompokkan data baik itu bersifat fisik atau abstrak kedalam suatu kelompok atau kelas yang memiliki kesamaan sifat disebut *clustering*[5].

2.3. *K-Means*

Algoritma *K-Means* merupakan salah satu metode pengelompokkan data *nonhierarki* (sekatan) yang berusaha mempartisi data yang ada ke dalam bentuk dua atau lebih kelompok. Metode ini mempartisi data ke dalam kelompok sehingga data berkarakteristik sama dimasukkan ke dalam satu kelompok yang sama dan data yang berkarakteristik berbeda dikelompokkan ke dalam kelompok yang lain [6].

Algoritma *K-Means* dimulai dengan penentuan jumlah kelompok (*cluster*). Tahap selanjutnya mengalokasikan data ke dalam kelompok (*cluster*) secara acak. Kemudian tentukan pusat kelompok terdekatnya dengan *cosine similarity*. Perbaharui pusat kelompok berdasarkan kelompok yang di dapat sebelumnya. Proses berulang terus sampai tidak ada satupun dokumen berpindah kelompok.

Untuk memperbaharui pusat kelompok pada iterasi selanjutnya menggunakan persamaan (1) dengan memperhatikan distribusi kemunculan term yang ada.

Rumus memperbaharui pusat baru adalah sebagai berikut :

$$Pusat = \big((nt - nr) \times (R(1,2) - R.ND)\big) \qquad (1)$$

nt = Frekuensi tertinggi kemunculan term

nr = Frekuensi terendah kemunculan term

 e-ISSN 2338-5197

$R(1,2)$ = Bilangan acak antara 1 sampai dengan 2

$R.ND$ = Sembarangan bilangan acak

2.4. *Cosine Similarity*

Cosine similarity adalah ukuran kesamaan yang lebih umum digunakan dalam information retreval dan merupakan ukuran sudut antara vektor dokumen d1 dan d2. d1 merepresentasikan kemunculan term pada dokumen 1 dan d2 merepresentasikan kemunculan term pada dokumen 2 [4].

Adapun rumus kesamaan kosinusnya adalah sebagai berikut:

$$\text{cosine}(d_1, d_2) = \frac{(d_1 . d_2)}{\|d_1\| \cdot \|d_2\|} \quad (2)$$

d_1 = Dokumen 1

d_2 = Dokumen 2

Ketika dua dokumen identik, sudutnya adalah nol derajat dan kesamaanya adalah satu (1), dan ketika dua dokumen tidak identik sama sekali, sudutnya adalah 90 derajat dan kesamaanya adalah nol (0) [3].

2.5. *Outlier*

Outlier atau data pencilan adalah kumpulan obyek-obyek yang dipandang sangat berbeda dibandingkan keseluruhan data [5].

2.6. *Purity*

Untuk melakukan evaluasi terhadap hasil pengelompokkan, maka dilakukan pengukuran *purity*. *Purity* menghitung rasio antara jumlah dokumen yang pengelompokkannya benar dengan total dokumen yang dianalisis. Nilainya berkisar dari 0 sampai 1. Semakin mendekati 1 mengindikasikan bahwa semakin banyak dokumen yang telah sesuai atau benar pengelompokkannya. Sebaliknya, semakin mendekati 0 menunjukkan semakin sedikit dokumen yang benar pengelompokkannya [4].

Adapun rumus rasio sebagai berikut :

$$purity(\Omega, C) = \frac{1}{N} \sum_k \max_j \left|\omega_k \cap C_j\right| \quad (3)$$

Ω = $\{\omega_1, \omega_2 \dots \omega_k\}$ adalah himpunan cluster

ω_k = himpunan dokumen dalam ω_k

C = $\{C_1, C_2 \dots C_j\}$ adalah himpunan class

C_j = himpunan dokumen dalam C_j

N jumlah data

$\max_j$ banyak elemen yang sama dalam

e-ISSN 2338-5197

3. Metode Penelitian

3.1. Tahapan *Text Mining*

Pada bagian ini menjelaskan tentang tahapan-tahapan pengelompokkan judul kerja praktek. Berikut proses yang menjelaskan tentang tahapan-tahapan dalam pengelompokkan judul kerja praktek.

Load data → *Tokenizing* → *Filtering* → *Stemming* → Menentukkan titik pusat → *Clustering* → Tes Akurasi.

A. Load Data

Sebelum dilakukan load data, data terlebih dahulu dilakukan *preprocessing* data yaitu dengan menghilangkan tanda . , “” () & - /dan lain sebagainya.

B. Tokenizing

Setelah judul kerja praktek berhasil di *upload* maka tahapan selanjutnya adalah tahapan *tokenizing* yaitu tahap pemotongan *string* berdasarkan tiap kata penyusunnya. Judul kerja praktek akan dipecah menjadi per kata.

C. Filtering

Filtering adalah proses mengecek satu per satu kata-kata yang ada pada setiap judul kerja praktek kemudian membuang kata-kata yang dianggap tidak penting. Kata yang dihapus adalah kata hubung atau konjungsi yaitu “yang”, “dan”, “dengan”, “di”, “kepada”, “untuk” dan lain sebagainya.

D. Stemming

Stemming adalah tahapan setelah proses filtering, pada tahap ini kata-kata yang memiliki imbuhan akan dikembalikan ke kata dasar bentuk awal.

E. Menentukan titik pusat

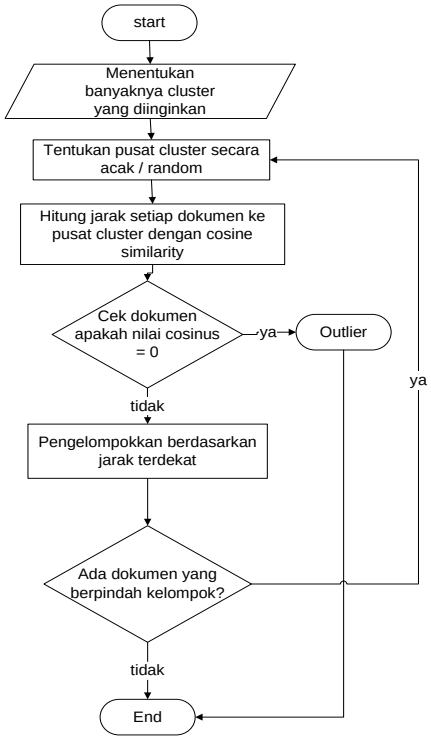
Proses ini sangat penting dilakukan karena untuk menentukan berapa banyak kluster yang akan dibentuk nantinya. Proses ini dapat dilakukan setelah proses *stemming* selesai.

F. Clustering

Proses dilakukan dengan menentukan banyaknya *cluster* terlebih dahulu. Kemudian sistem akan menentukan titik pusat secara *random* atau acak. Selanjutnya sistem mengelompokkan judul kerja praktek berdasarkan jarak kedekatan dengan rumus *cosine similarity* sesuai pada persamaan (2). Sistem akan mencari hasil *similarity* dengan memilih titik pusat terbesar pada masing-masing judul kerja praktek. Dan mengecek satu per satu judul kerja praktek apakah termasuk *outlier* atau tidak. Proses berhenti jika dokumen sudah tidak berpindah kelompok. Hasil akhirnya akan terbentuk pola-pola *cluster*.

Flowchart langkah algoritma *k-means clustering* dengan *cosine similarity* dapat dilihat dalam Gambar 3.1 berikut ini :

e-ISSN 2338-5197



Gambar 3.1 Flowchart langkah algoritma K-Means Clustering dengan Cosine Similarity

G. Tes Akurasi

Tes akurasi akan dilakukan untuk mengetahui sebaik apakah algoritma k-means clustering dalam hal akurasi. Pada tes akurasi dilakukan dengan perhitungan manual menggunakan persamaan rumus *purity* (3). Semakin mendekati 1 mengindikasikan bahwa semakin banyak dokumen yang telah sesuai atau benar pengelompokkannya begitupun sebaliknya.

3.2. Evaluasi Pola dan Representasi Pengetahuan

Pada tahapan ini juga dilakukan uji model pola untuk mengetahui nilai kesesuaian dan kesalahan atau error pada pola yang ditemukan. Pola yang sudah ditemukan kemudian direpresentasikan kepada pengguna agar mudah dipahami. Pada representasi pengetahuan algoritma *k-means clustering* dilakukan ekplorasi terhadap kecenderungan kemiripan dari tiap judul kerja praktek terhadap judul kerja praktek lainnya. Dengan melakukan pemetaan kedekatan dari sebuah judul kerja praktek dengan judul kerja praktek lainnya sehingga pada akhirnya akan didapatkan *cluster-cluster* judul kerja praktek.

4. Pembahasan

4.1. Tahapan Text Mining

A. Load Data

Setelah dilakukan *preprocessing* data yaitu dengan menghilangkan tanda . , “” () & - /dan lain sebagainya. Jumlah data judul kerja praktek sebanyak 355 data akan di Load atau dimasukkan kedalam sistem dalam bentuk .xls.

B. Tokenizing

Setelah judul kerja praktek berhasil di *upload* maka tahapan selanjutnya adalah tahapan *tokenizing* yaitu tahap pemotongan *string* berdasarkan tiap kata penyusunnya. Judul kerja praktek akan dipecah menjadi per kata.

Judul :

 e-ISSN 2338-5197

Pelatihan internet dan pembuatan blog di SDN 1 jetis bantul

Hasil Tokenizing :

Pelatihan-internet-dan-pembuatan-blog-di-SDN-1-jetis-bantul

C. Filtering

Filtering adalah proses mengecek satu per satu kata-kata yang ada pada setiap judul kerja praktek kemudian membuang kata-kata yang dianggap tidak penting. Kata yang dihapus adalah kata hubung atau konjungsi yaitu “yang”, “dan”, “dengan”, “di”, “kepada”, “untuk” dan lain sebagainya.

Judul :

Pelatihan internet **dan** pembuatan blog **di** SDN 1 jetis bantul

Hasil Filtering :

Pelatihan-internet-pembuatan-blog-SDN-1-jetis-bantul

D. Stemming

Stemming adalah tahapan setelah proses filtering, pada tahap ini kata-kata yang memiliki imbuhan akan dikembalikan ke kata dasar bentuk awal.

Judul :

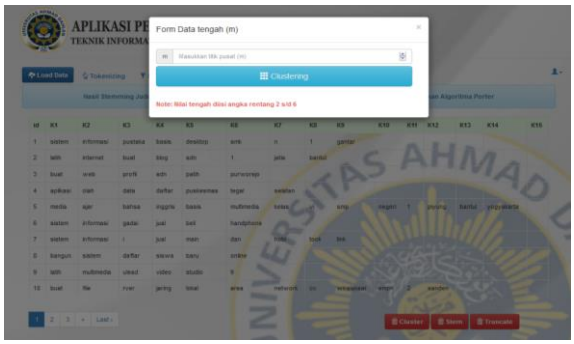
Pelatihan internet **pembuatan** blog SDN 1 jetis bantul

Hasil Stemming :

latih-internet- **buat**-blog- SDN-1-jetis-bantul

E. Menentukan titik pusat

Proses ini sangat penting dilakukan karena untuk menentukan berapa banyak kluster yang akan dibentuk nantinya. Proses ini dapat dilakukan setelah proses *stemming* selesai. Penginputan nilai titik pusat diisi antara 2 sampai dengan 6.



Gambar 1. Input Nilai Titik Pusat

F. Clustering

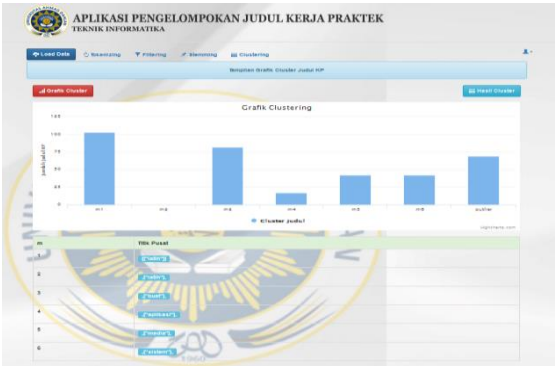
Proses dilakukan dengan menentukan banyaknya *cluster* terlebih dahulu. Kemudian sistem akan menentukan titik pusat secara *random* atau acak. Selanjutnya sistem mengelompokkan judul kerja praktek berdasarkan jarak kedekatan dengan rumus *cosine similarity* sesuai pada persamaan (2). Sistem akan mencari hasil *similarity* dengan memilih titik pusat terbesar pada masing-masing judul kerja praktek. Dan mengecek satu per satu judul kerja praktek apakah termasuk *outlier* atau tidak. Proses berhenti jika dokumen sudah tidak berpindah kelompok.

e-ISSN 2338-5197

Gambar 2. Hasil *Clustering*

Setelah proses *clustering* selesai maka hasil akhirnya akan terbentuk pola-pola *cluster*. Pola-pola *cluster* akan ditampilkan dalam bentuk detail laporan yang berisi judul kerja praktek yang telah berhasil dikelompokkan selain itu ditampilkan juga dalam data *statistic* dari proses *k-means clustering* seperti jumlah judul kerja praktek pada masing-masing cluster. Seperti pada Gambar 3 dan Gambar 4.

Gambar 3. Pola Cluster



Gambar 4. Grafik Cluster

G. Tes Akurasi

Tes Akurasi atau pengujian sistem pada penelitan ini menggunakan metode *Purity test*. Hasil dari pengujian sistem pada penelitian ini *purity test* dilakukan untuk mengetahui baik buruknya *cluster* yang dihasilkan pada proses *clustering*. *Purity test* dilakukan dengan data judul kerja praktek sebanyak 355 judul kerja praktek dengan mengkombinasikan parameter M menggunakan persamaan rumus (3).

Hasil dari *purity test* yang sudah dilakukan dapat dilihat pada Tabel 1 berikut ini :

 e-ISSN 2338-5197

Tabel 1. Hasil Pengujian Purity

No	Jumlah Data	M	Purity
1	355	2	0,30
2	355	3	0,46
3	355	4	0,51
4	355	5	0,64
5	355	6	0,85

Uji *purity* yang sudah dilakukan menggunakan data set sebanyak 355 judul kerja praktek dengan mengkombinasi parameter M sebagai titik pusat *cluster* dihasilkan nilai *purity* terbaik sebesar 0,85 dengan kombinasi M sebesar 6. Karena hasil purity mendekati 1 mengindikasikan bahwa semakin banyak dokumen yang telah sesuai atau benar pengelompokkannya.

4.2. Evaluasi Pola dan Representasi Pengetahuan

Dilihat pada tabel 1 Hasil Pengujian *Purity* bisa disimpulkan bahwa pengelompokkan terbaik menggunakan data set sebanyak 355 dengan M = 6 sebesar 0,85. Adapun data judul yang termasuk dalam 6 *cluster* adalah sebagai berikut :

1. Pada *cluster* 1 memiliki titk pusat *cluster* kata latih dan jumlah judul dalam *cluster* ini sebanyak 103 data
2. Pada *cluster* 2 dengan nama *cluster* latih dihasilkan pengelompokkan judul sebanyak 0 data dikarenakan nama *cluster* sama dengan nama *cluster* pada cluster 1.
3. Pada *cluster* 3 memiliki titk pusat *cluster* kata buat dan jumlah judul dalam *cluster* ini sebanyak 68 data.
4. Pada *cluster* 4 memiliki titk pusat *cluster* kata aplikasi dan jumlah judul dalam *cluster* ini sebanyak 17 data.
5. Pada *cluster* 5 memiliki titk pusat *cluster* kata media, ajar dan jumlah judul dalam *cluster* ini sebanyak 56 data.
6. Pada *cluster* 6 memiliki titk pusat *cluster* kata media, ajar dan jumlah judul dalam *cluster* ini sebanyak 42 data.

5. Kesimpulan dan Saran

5.1. Kesimpulan

Uji *purity* yang sudah dilakukan pada Sistem aplikasi web dengan *text mining* untuk pengelompokkan judul kerja praktek dengan metode *k-means clustering* yaitu menggunakan data set sebanyak 355 judul kerja praktek dengan mengkombinasikan parameter M sebagai titik pusat cluster didapatkan hasil bahwa nilai *purity* terbaik adalah sebesar 0,85 dengan kombinasi M = 6 mengindikasikan bahwa semakin banyak dokumen yang telah sesuai atau benar pengelompokkannya.

5.2. Saran

Penelitian aplikasi web dengan *text mining* untuk pengelompokkan judul jurnal publikasi dengan metode *k-means clustering* masih memiliki beberapa kekurangan, maka diharapkan adanya pengembangan lebih lanjut lagi. Pada proses *upload* data masih dilakukan secara manual yaitu dengan menggunakan file berekstensi .xls. Diharapkan untuk dapat dikembangkan lagi sehingga proses *upload* data menjadi otomatis dengan menggunakan teknik *crawler* atau semacamnya. Pada

e-ISSN 2338-5197

proses *clustering* perlu adanya pengombinasian algoritma sehingga menjadi lebih efisien untuk memproses data dalam jumlah besar.

Daftar Pustaka

[1] Kurniawan, B., Effendi, S., sitompul, O.S., 2012, Klasifikasi Konten Berita Dengan Metode Text Mining, *Jurnal Dunia Teknologi Informasi*, Vol.1, No. 1, Hal.14-19.

[2] Alfina, T., Santosa, B., Barakbah, A.R., 2012, Analisa Perbandingan Metode Hierarchical Clustering, K-Means Dan Gabungan Keduanya Dalam Cluter Data (Studi Kasus : Problem Kerja Praktek Jurusan Teknik Industri Its), *Jurnal Teknik ITS*, vol.1, ISSN: 2301-9271, September 2012.

[3] Imbar, R.V., Adelia., Ayub, M., Rehatta, A., 2014, Implementasi Cosine Similarity dan Algoritma Smith-Waterman untuk Mendeteksi Kemiripan Teks, *Jurnal Informatika*, Vol. 10, nomor 1, Juni 2014:31-42.

[4] Prilianti, K.R., Wijaya, H., 2014, Aplikasi Text Mining Untuk Automasi Penentuan Tren Topik skripsi dengan Metode K-Means Clustering, *Jurnal Cybermatika*, Vol. 2, No. 1, juni 2014.

[5] Han, J., Kamber, M., Pei, J., 2006, Data Mining: Concepts and Techniques. Morgan Kaufmann.

[6] Prasetyo, E., 2014, *Data Mining-Mengolah Data Menjadi Informasi Menggunakan Matlab*. Yogyakarta.Andi.