PERANCANGAN APLIKASI PENENTUAN TINGKAT KEMIRIPAN DOKUMEN TEKS MENGGUNAKAN METODE *INTELLIGENT K-MEANS*

Fika Alfiani [1)] Lely Hiryanto[2)] Tri Sutrisno[3)]

[1)]Teknik Informatika Universitas Tarumanagara
Jl. Letjen S.Parman No.1, Jakarta Barat 11440 Indonesia
email : patriciafika@gmail.com
[2)]Teknik Informatika Universitas Tarumanagara
Jl. Letjen S.Parman No.1, Jakarta Barat 11440 Indonesia
email : lelyh@fti.untar.ac.id
[3)]Teknik Informatika Universitas Tarumanagara
Jl. Letjen S.Parman No.1, Jakarta Barat 11440 Indonesia
email : tris@fti.untar.ac.id

ABSTRACT

Kesamaan antar dokumen bukan merupakan fenomena baru dalam dunia pendidikan. Sebelum hadirnya teknologi informasi, fenomena plagiat juga telah ada. Namun, hadirnya teknologi informasi secara nyata lebih mempermudah orang untuk melakukan plagiat. Plagiarisme umumnya mengacu pada penggunaan informasi, teks, gagasan yang tidak sah, tanpa referensi tang tepat terhadap sumber asli dari data asal. Intelligent K-Means merupakan salah satu metode pengelompokan data dimana jumlah klaster yang dihasilkan secara dinamis, sehingga dapat mempersempit variasi antar data yang dideteksi. Diawali dengan proses preprocessing kemudian pembobotan term atau setiap kata dengan metode tf-idf sehingga dapat mewakili dokumen yang diuji dalam perhitungan menggunakan metode Intelligent K-Means. Hasil pengujian terhadap dokumen yang sudah direkayasa dan dokumen asli skripsi mahasiswa Fakultas Teknologi Informasi Universitas Tarumanagara menunjukan bahwa metode dapat mengelompokan dokumen dengan jumlah kelompok/klaster sebanyak 3 klaster

Key words

Intelligent K-Means, K-Means, Clustering, TF-IDF

1. PENDAHULUAN

Plagiarisme atau plagiat adalah penjiplakan atau pengambilan karangan, pendapat, dan sebagainya dari orang lain dan menjadikannya seolah karangan dan pendapat sendiri[1]. Plagiat bukan merupakan fenomena baru dalam dunia pendidikan. Sebelum hadirnya teknologi informasi, fenomena plagiat juga telah ada. Namun, hadirnya teknologi informasi secara nyata lebih mempermudah orang untuk melakukan plagiat. Perilaku plagiat telah terjadi mulai dari instituisi sekolah, perguruan tinggi, sampai dengan masyarakat. Tindakan plagiat tentunya akan merusak moral civitas akademik dalam dunia Pendidikan. Perkembangan teknologi Informasi yang seharusnya dapat membantu dalam pencarian informasi ternyata disalah gunakan dalam tindak plagiarisme berupa mencuri ide/karya oranglain tanpa mencantumkan kutipan sumber terkait. Dalam mencegah plagiarisme berarti melakukan tindakan pencegahan agar plagiarisme tidak terjadi, misalnya dengan menetapkan sebuah kebijakan tentang plagiarisme atau sistem hukuman bila terbukti melakukan plagiarisme. Sebenarnya pemerintah telah memberikan perhatian khusus terkait permasalahan plagiarisme dalam Peraturan Mentri Pendidikan Nasional Republik Indonesia Nomor 17 Tentang Pencegahan dan Penanggulangan Plagiat di Perguruan Tinggi 2010 dan undang-undang Republik Indonesia Nomor 19 Tentang Hak Cipta 2002 Mendeteksi plagiarisme berarti menemukan tindakan plagiarisme yang telah terjadi. Plagiarisme umumnya mengacu pada penggunaan informasi, teks, gagasan yang tidak sah, tanpa referensi tang tepat terhadap sumber asli dari data asal. Mendeteksi plagiarisme hanya dapat mengurangi plagiarisme yang telah terjadi, tetapi mencegah plagiarisme tentu dapat menghilangkan atau paling tidak mengurangi sebagian besar plagiarisme. Namun dalam kenyataan, implementasi pencegahan plagarisme merupakan masalah moral masyarakat secara luas yang tidak dapat diselesaikan hanya dengan upaya Universitas atau Departemen[2]. Oleh karena itu, yang mungkin dilakukan adalah mendeteksi plagiarisme. Pendeteksian plagiarisme dilakukan secara manual atau dengan bantuan komputer[3]. Saat ini pendeteksian secara

manual merupakan cara yang paling akurat dalam mendeteksi plagiat. Kelemahan dari cara ini adalah menghabiskan banyak waktu dan tenaga serta tidak konsisten karena dipengaruhi juga oleh faktor emosional.

Kini tindakan plagiarisme marak terjadi dikalangan lingkup akademisi. Akademisi sering melakukan tindak penjiplakan atau plagiarisme dalam mengerjakan tugas-tugas dan yang terparah adalah tindakan penjiplakan karya ilmiah. Hal ini pun dapat terjadi di lingkungan mahasiswa Fakultas Teknologi Informasi di Universitas Tarumanagara.(Untar).

Adanya ruang karya ilmiah Fakultas Teknologi Informasi (FTI) Universitas Tarumanagara (UNTAR) yang menyimpan berbagai karya ilmiah dari sivitas akademika seperti skripsi, laporan penelitian, laporan kerja praktik dan lain sebagainya telah tersedia dalam versi digital memudahkan para mahasiswa untuk mendapatkan informasi yang dicari, namun kemudahan ini dijadikan sarana untuk melakukan tindakan plagiarisme. Tentunya tindakan seperti ini bertentangan dengan hak seseorang. Oleh karena itu, diperlukan aplikasi yang dapat meminimalisir tindak plagiat di kalangan masyarakat maupun akademisi dimana nantinya bisa diketahui tingkat kesamaannya.

2. LANDASAN TEORI

2.1 TEKS PRE-PROCESSING

Teks *pre-processing* adalah suatu tahapan awal dalam proses pengolahan teks, yang merupakan suatu tahapan yang sangat penting[4]. Tahapan dokumen teks sebelum diolah lebih lanjut, beberapa proses itu antara lain :

1. Mengubah huruf kapital menjadi huruf kecil
2. Menghapus tanda baca seperti titik, koma, petik, garis miring, garis bawah, tanda pisah, tanda seru, tanda tanya, tanda kutip.

2.2 FILTERING

Filtering merupakan proses seleksi terhadap kata-kata yang dihasilkan oleh tahap sebelumnya, dapat dilakukan dengan Algoritma *Stoplist / Stopword.* Cara kerja algoritma ini adalah dengan membuang kata-kata yang tidak deskriptif yang biasanya muncul dalam jumlah banyak dan dianggap tidak memberikan informasi penting[5].

2.3 TF-IDF

TF-IDF (Term Frequency – Inverse Document Frequency) merupakan sebuah perhitungan statistik yang bertujuan untuk menggambarkan seberapa penting sebuah kata terhadap sebuah koleksi dokumen[6]. TF-IDF sering digunakan sebagai faktor pembobotan di dalam *information retrieval* dan *text mining*. Nilai TF-IDF meningkat seiring dengan banyaknya jumlah kata tersebut muncul di dalam dokumen, tetapi diimbangi dengan jumlah dokumen yang memuat kata tersebut, nilai TF-IDF kata tersebut akan semakin kecil.

Nilai TF-IDF suatu kata pada dokumen dihitung dengan cara mengalikan *Term Frequency* suatu kata pada dokumen dan *Inverse Document Frequency* kata tersebut. Sehingga dapat ditulis sebagai berikut[6] :

$$tfidf(t,d)=tf(t,d)*idf_t$$

Keterangan :
$tf(d,f)$ = *Term Frequency* suatu kata pada dokumen
idf_t = *Inverse Document Frequency* suatu kata

2.4 INTELLIGENT K-MEANS

Algoritma *Intelligent K-Means* merupakan versi K-Means dimana nilai cluster dan centroid awalnya ditentukan menggunakan *anomalous pattern*[7]. Algoritma *Intelligent K-Means* dapat menentukan nilai *k* atau jumlah klaster secara dinamis namun untuk perhitungan dalam pengelompokan sama seperti algoritma K-Means.

Algoritma *Intelligent K-Means* menggunakan *anomalous pattern* untuk mencari dua centroid pertama[7]. Berikut adalah algoritma *anomalous pattern*:

1. Hitung *Center of Mass*(COM) atau rata-rata dari data untuk setiap variable atau atribut menggunakan persamaan :

$$x = 1, (S_{xj}) = \frac{1}{n}\sum_{i}^{n}(X_{ij})$$

Dimana :
(S) adalah *Center of Mass*(CoM), n adalah banyaknya data, i adalah indeks vector dari jumlah data, j adalah indeks vector dari jumlah atribut dan x adalah data.
2. Tentukan objek C1 dengan menentukan objek yang memiliki jarak terjauh dari CoM dengan menggunakan persamaan :

$$x = 1, Max\ d_{ix} = \sqrt{\sum_{j=1}^{m}(X_{ij} - (S_{1j}))^2}$$

Dimana :
d_{ix} adalah jarak antara pusat massa (CoM) dengan setiap objek pada data, i adalah indeks vector dari jumlah data dan j adalah indeks vector dari jumlah atribut.
3. Cari objek lainnya yaitu C2 yang mempunyai jarak terjauh dari C1 menggunakan persamaan :

$$Max\ d_{ix} = \sqrt{\sum_{j=1}^{m}(X_{ij} - C_{kj})^2}$$

4. Hitung jarak objek yang lain terhadap C1 dan C2
5. Kelompokan Objek, objek yang terdekat dengan C1 adalah klaster S_1, sedangakan yang terdekat dengan C2 adalah klaster S_2.

Kemudian dicari centroid lainnya menggunakan *anomalous pattern* dengan langkah[7] :
1. Untuk setiap klaster S_i, cari kandidat centroid baru $C_1{}'$ dengan mencari objek terjauh dari kelompok centroid C_i menggunakan persamaan (4) di atas di mana i = 1,2,n-1.
2. Optimisasi jarak antara seluruh centroid dengan menemukan nilai mean dari setiap kandidat, C_{ni}, dengan semua kelompok centroid, dimana :

C_{ni}=mean (d(C_1,$C_i{}'$), d(C_2,$C_i{}'$),…,d (C_3,$C_1{}'$))

Keterangan :
Ci adalah kelompok centroid diamana i = 1,2,..,n-1
3. Centroid yang baru :

Cn=max($Cn1$,$Cn2$,…,$Cnn-i$)

Keterangan :
Cn adalah kandidat centroid baru dimana n = 1,2,…n-1
4. Kelompokan semua objek menurut jarak terdekat dengan salah satu klaster.
5. Ulangi langkah pertama sampai tidak ada objek yang terubah klaster.

Berikut merupakan langkah – langkah dalam Algoritma *Intelligent K-Means :*

Proses klastering dimulai dengan mengidentifikasi data yang akan dikelompokan, X_{ij}(i=1,… ,n dan j=1,…,m) dimana n adalah banyaknya data yang akan dikelompokan dan m adalah banyaknya variable.
1. Pada awal iterasi tentukan titik tengah(centroid) dari klaster dengan mencari titik refrensi berupa titik pusat data, kemudian ambil titik terjauh untuk membuat klaster baru.
2. Selanjutnya kedua klaster diperbaharui sampai konvergen atau sama. Tahap ini dilakukan berulang kali dan klaster baru terbentuk hingga semua klaster menjadi konvergen.
3. Kemudian dihitung jarak setiap data dengan pusat klaster menggunakan rumus jarak Eucledian. Untuk perhitungan jarak data ke-I (Xi) pada pusat klaster ke-k(Ck) adalah sebagai berikut[8] :

$$D_{ik} = \sqrt{\sum_{j=1}^{m}(x_{ij} - C_{kj})^2}$$

4. Suatu data akan menjadi klompok klaster k apabila jarak data tersebut ke pusat klaster-k bernilai paling kecil dibandingkan dengan jarak ke pusat klaster lainnya. Jarak minimum tersebut dapat dihitung dengan menggunakan persamaan[8]:

$$Min\sum_{k=1}^{k} d_{ik} = \sqrt{\sum_{j=1}^{m}(x_{ij} - y_{kj})^2}$$

5. Nilai pusat klaster yang baru dihitung dengan mencari rata-rata dari data yang menjadi anggota pada klaster yang ditentukan. Perhitungan untuk mencari nilai pusat klaster yang baru dari rata-rata data adalah sebagai berikut[8]:

$$C_{kj} = \frac{\sum_{i=1}^{n} X_{ij}}{n}$$

2.5 Varians atau simpangan baku

Dasar penghitungan varian adalah keinginan untuk mengetahui keragaman suatu kelompok data. Salah satu cara untuk mengetahui keragaman dari suatu kelompok data adalah dengan mengurangi setiap nilai data dengan rata-rata kelompok data tersebut, selanjutnya semua hasilnya dijumlahkan

$$s = \sqrt{\frac{\sum_{x=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

3. SKENARIO PENGUJIAN

Pengujian dilakukan terhadap dokumen asli yaitu dokumen skripsi mahasiswa Teknologi Informasi Universitas Tarumanagara. Pengujian akan menampilkan hasil *cluster* yang diproses menggunakan metode *Intelligent K-Means.* Kemudian dilakukan perhitungan persentase kemunculan setiap kata dalam setiap dokumen dengan mengambil dua dokumen sebagai wakil dari anggota setiap cluster yang diuji, kemudian hasil perhitungan untuk mengetahui apakah cluster dapat meminimalkan variasi antar dokumen dicluster yang sama.

4. HASIL PENGUJIAN

Tabel 1 menunjukan hasil pengelompokan menggunakan metode *Intelligent K-Means* terhadap 4 dokumen pembanding.

Tabel 1. Hasil perhitungan cluster dan varian Bab I

Nama dokumen	Cluster	Varians
Bab I – Ayu Windy	1	1.004282
Bab I – Josselyn Sinthia Thio	1	0.323525
Bab I – Vina Tandean	1	0.290044
Bab I – Sri Whisnu A W	1	0.404096
Bab I – Fransisca Regina	1	0.400229
Bab I – Irawati Djajadi	1	2.358203
Bab I – Mishele	1	0.456026
Bab I – Jacklin Shintia Thio	1	1.482407
Bab I – Mariana	1	1.139453
Bab I – Nadia Yanitra	1	0.795019
Bab I – Victorious	1	2.217171
Bab I – Farenco	1	1.421556
Bab I – Yunita	1	0.589111
Bab I – Rosalinda	1	2.242168
Bab I – Rionaldy Trisaputra	1	2.17614
Bab I – Stevy Lie	1	1.432501
Bab I – Stephen Yan	1	1.271029
Bab I – Gabriel Fransisco	2	5.394813
Bab I – Ferry Dharmawan	3	3.059251
Bab I – Renaldo Ali	3	3.138923

Tabel 2 hasil perhitungan cluster dan varians Bab II

Nama dokumen	Cluster	Varians
Bab II – Irawati Djajadi	1	4.213532
Bab II – Rosalinda	1	3.418931
Bab II – Ayu Windy	2	2.399449
Bab II – Josselyn Sinthia Thio	2	2.263637
Bab II – Vina Tandean	2	0.833117
Bab II – Sri Whisnu A W	2	2.335465
Bab II – Fransisca Regina	2	0.727478
Bab II – Mishele	2	1.615848
Bab II – Mariana	2	1.42933
Bab II – Nadia Yanitra	2	0.863218
Bab II – Victorious	2	1.81679
Bab II – Farenco	2	2.244268
Bab II – Yunita	2	0.374132
Bab II – Rionaldy Trisaputra	2	1.358272
Bab II – Stevy Lie	2	0.588417
Bab II – Stephen Yan	2	0.904072
Bab II – Gabriel Fransisco	2	2.289932
Bab II – Ferry Dharmawan	2	2.151808
Bab II – Renaldo Ali	2	1.519385
Bab II – Jacklin Shintia Thio	3	363.7556

Tabel 3 hasil perhitungan cluster dan varians Bab III

Nama dokumen	Cluster	Varians
Bab III – Ayu Windy	1	0.455056
Bab III – Josselyn Sinthia Thio	1	1.849378
Bab III – Vina Tandean	1	0.579211
Bab III – Sri Whisnu A W	1	0.187477
Bab III – Fransisca Regina	1	1.063008
Bab III – Mishele	1	1.635292
Bab III – Mariana	1	4.665269
Bab III – Nadia Yanitra	1	1.120049
Bab III – Victorious	1	5.143833
Bab III – Farenco	1	3.181988
Bab III – Yunita	1	2.068527
Bab III – Rionaldy Trisaputra	1	0.638833
Bab III – Stevy Lie	1	1.705081
Bab III – Stephen Yan	1	0.860002
Bab III – Gabriel Fransisco	1	5.954124
Bab III – Ferry Dharmawan	1	2.365775
Bab III – Renaldo Ali	1	2.801166
Bab III – Jacklin Shintia Thio	2	134.6102
Bab III – Irawati Djajadi	3	13.88834
Bab III – Rosalinda	3	6.390453

Tabel 4 hasil perhitungan cluster dan varians Bab IV

Nama dokumen	Cluster	Varians
Bab IV – Ayu Windy	1	0.301935
Bab IV – Josselyn Sinthia Thio	1	1.276645
Bab IV – Vina Tandean	1	4.134098
Bab IV – Sri Whisnu A W	1	0.322487
Bab IV – Fransisca Regina	1	2.638963
Bab IV – Irawati Djajadi	1	1.370472
Bab IV – Mishele	1	4.718017
Bab IV – Jacklin Shintia Thio	1	0.954643
Bab IV – Mariana	1	2.552958
Bab IV – Nadia Yanitra	1	1.633421
Bab IV – Victorious	1	1.452187
Bab IV – Farenco	1	3.125674
Bab IV – Yunita	1	3.924309
Bab IV – Rosalinda	1	0.958094
Bab IV – Rionaldy Trisaputra	1	5.039227
Bab IV – Stevy Lie	1	2.029095
Bab IV – Stephen Yan	1	4.371884
Bab IV – Gabriel Fransisco	2	24.32786
Bab IV – Ferry Dharmawan	3	10.97964
Bab IV – Renaldo Ali	3	10.0141

Tabel 5 hasil perhitungan cluster dan varians Bab V

Nama dokumen	Cluster	Varians
Bab V – Gabriel Fransisco	1	1.840279
Bab V – Renaldo Ali	2	0.943783
Bab V – Ayu Windy	3	0.118452
Bab V – Josselyn Sinthia Thio	3	0.301512
Bab V – Vina Tandean	3	0.20014
Bab V – Sri Whisnu A W	3	0.280274
Bab V – Fransisca Regina	3	0.08763
Bab V – Irawati Djajadi	3	0.737712
Bab V – Mishele	3	0.225144
Bab V – Jacklin Shintia Thio	3	0.459459
Bab V – Mariana	3	0.431099
Bab V – Nadia Yanitra	3	0.202301
Bab V – Victorious	3	0.233949
Bab V – Farenco	3	0.08763
Bab V – Yunita	3	0.348185
Bab V – Rosalinda	3	0.429772
Bab V – Rionaldy Trisaputra	3	0.412245
Bab V – Stevy Lie	3	0.313288
Bab V – Stephen Yan	3	0.310559
Bab V – Ferry Dharmawan	3	0.584617

5. KESIMPULAN

- Metode Intelligent K-Means dapat mengelompokan data tanpa harus menentukan nilai kluster harus ditentukan. Sehingga data dapat lebih mengelompok berdasarkan besar perhitungan bobot dalam dokumen.

- Dari hasil pengujian bahwa metode Intelligent K-Means dapat mengecilkan variasi antar data dan setelah dihitung besar varians disetiap dokumen hasil menunjukan bahwa varians antar dokumen dapat mengelompok tergantung nilai besar variansnya.

REFERENSI

[1] Departemen Pendidikan Nasional, 1997, "Kamus Besar Bahasa Indonesia". Jakarta: Balai Pustaka

[2] Lukashenko, Graudina, and Grundspenkis, 2007, "Computer-Based Plagiarism Detection Methods and Tools : An Overview", International Conference on Computer Systems and Technologies

[3] Mahathir, Ridha Ahmad, 2011, Sistem Pendeteksi Plagiat Pada Dokumen Teks Berbahasa Indonesia Menggunakan Metode Rouge-N, Rouge-L Dan Rouge-W, (http://repository.ipb.ac.id/handle/123456789/50046)

[4] Xiaojin Zhu, 1999, "Basic Text Process", Vol. 463, ACM press

[5] Christoper D. Manning, Prabhakar Raghavan, dan Hinrich Schutze, 2008, "Introduction to Information Retrieval", (Cambrige : Cambrige University Press)

[6] Jure Leskovec, Anand Rajaraman, dan Jeffrey D. Ullman, 2011, "Mining of Massive Dataset", (Cambridge : Cambridge University Press)

[7] M.M.T. Chiang, Boris Mirkin, 2013, "Intelligent Choice of the Number of Clusters in K-Means Clustering : AnExperimental Study with Different Cluster Spread", Journal of Classification Springer, Vol. 27.

[8] Archana Singh, Avantika Yadav, Ajay Rana, 2013,"K-Means with Three different Distance Metrics", International Journal of Computer Application, Vol.67.

Fika Alfiani., mahasiswa pada program studi Teknik Informatika Universitas Tarumanagara..

Lely Hiryanto ST, M.Sc., memperoleh gelar M.Sc dari University Computer Science Curtin University Of Technology, Australia tahun 2006. Kemudian tahun 2001 memperoleh ST dari Universitas Tarumanagara, Jakarta. Saat ini sebagai Dosen Tetap program studi Teknik Informatika Universitas Tarumanagara.

Tri Sutrisno S.Si.,M.Sc., memperoleh gelar M.Sc dari Universitas Gajah Mada, Indonesia tahun 2015. Kemudian tahun 2011 memperoleh S.Si dari Universitas Diponegoro, Indonesia. Saat ini sebagai Dosen Tetap program studi Teknik Informatika Universitas Tarumanagara.