IMPLEMENTASI *MODIFIED K-MEANS CLUSTERING* PADA PENGELOMPOKAN DATA KOMENTAR SENTIPOL

Skripsi

oleh

RUDDY CAHYANTO

71130097

PROGRAM STUDI INFORMATIKA FAKULTAS TEKNOLOGI INFORMASI

UNIVERSITAS KRISTEN DUTA WACANA

2020

IMPLEMENTASI *MODIFIED K-MEANS CLUSTERING* PADA PENGELOMPOKAN DATA KOMENTAR SENTIPOL

Skripsi

Diajukan kepada Program Studi Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana
Sebagai Salah Satu Syarat dalam Memperoleh Gelar
Sarjana Komputer

Disusun oleh

RUDDY CAHYANTO
71130097

PROGRAM STUDI INFORMATIKA FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS KRISTEN DUTA WACANA
2020

PERNYATAAN KEASLIAN SKRIPSI

Saya menyatakan dengan sesungguhnya bahwa skripsi dengan judul:

IMPLEMENTASI MODIFIED K-MEANS CLUSTERING PADA PENGELOMPOKAN DATA KOMENTAR SENTIPOL

yang saya kerjakan untuk melengkapi sebagian persyaratan menjadi Sarjana Komputer pada pendidikan Sarjana Program Studi Informatika Fakultas Teknologi Informasi Universitas Kristen Duta Wacana, bukan merupakan tiruan atau duplikasi dari skripsi kesarjanaan di lingkungan Universitas Kristen Duta Wacana maupun di Perguruan Tinggi atau instansi manapun, kecuali bagian yang sumber informasinya dicantumkan sebagaimana mestinya.

Jika dikemudian hari didapati bahwa hasil skripsi ini adalah hasil plagiasi atau tiruan dari skripsi lain, saya bersedia dikenai sanksi yakni pencabutan gelar kesarjanaan saya.

Yogyakarta, 8 Januari 2020

RUDDY CAHYANTO

71130097

HALAMAN PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR UNTUK KEPENTINGAN AKADEMIS SECARA ONLINE UNIVERSITAS KRISTEN DUTA WACANA YOGYAKARTA

Saya yang bertanda tangan di bawah ini:

NIM	: 71130097
Nama	: Ruddy Cahyanto
Prodi / Fakultas	: Informatika / Fakultas Teknologi Informasi
Judul Tugas Akhir	: Implementasi Modified K-Means Clustering pada Pengelompokan Data Komentar Sentipol

bersedia menyerahkan Tugas Akhir kepada Universitas melalui Perpustakaan untuk keperluan akademis dan memberikan **Hak Bebas Royalti Non Ekslusif (*Non-exclusive Royalty-free Right*)** serta bersedia Tugas Akhirnya dipublikasikan secara online dan dapat diakses secara lengkap (full access).

Dengan Hak Bebas Royalti Nonekslusif ini Perpustakaan Universitas Kristen Duta Wacana berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk database, merawat, dan mempublikasikan Tugas Akhir saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta. Demikian pernyataan ini saya buat dengan sebenar-benarnya.

Yogyakarta, 22 Januari 2020

(71130097 – Ruddy Cahyanto)

HALAMAN PERSETUJUAN

Judul Skripsi	:	IMPLEMENTASI MODIFIED K-MEANS CLUSTERING PADA PENGELOMPOKAN DATA KOMENTAR SENTIPOL
Nama Mahasiswa	:	RUDDY CAHYANTO
N I M	:	71130097
Matakuliah	:	Skripsi (Tugas Akhir)
Kode	:	TIW276
Semester	:	Gasal
Tahun Akademik	:	2019/2020

Telah diperiksa dan disetujui di
Yogyakarta,
Pada tanggal 8 Januari 2020

Dosen Pembimbing I	Dosen Pembimbing II
Antonius Rachmat C., S.Kom.,M.Cs.	Danny Sebastian, S.Kom., M.M., M.T.

HALAMAN PENGESAHAN

IMPLEMENTASI MODIFIED K-MEANS CLUSTERING PADA PENGELOMPOKAN DATA KOMENTAR SENTIPOL

Oleh: RUDDY CAHYANTO / 71130097

Dipertahankan di depan Dewan Penguji Skripsi
Program Studi Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana - Yogyakarta
Dan dinyatakan diterima untuk memenuhi salah satu syarat memperoleh gelar
Sarjana Komputer
pada tanggal 17 Desember 2019

Yogyakarta, 8 Januari 2020
Mengesahkan,

Dewan Penguji:

1. Antonius Rachmat C., S.Kom.,M.Cs.
2. Danny Sebastian, S.Kom., M.M., M.T.
3. Joko Purwadi, M.Kom
4. Willy Sudiarto Raharjo, S.Kom.,M.Cs.

Dekan	Ketua Program Studi
(Restyandito,S.Kom.,MSIS.,Ph.D.)	(Gloria Virginia,Ph.D.)

UCAPAN TERIMA KASIH

Dalam menyelesaikan penelitian ini, penulis mengucapkan terima kasih kepada Bapak Antonius Rachmat dan Bapak Danny Sebastian yang telah membimbing, mendampingi serta banyak membantu dalam proses penelitian. Penulis juga mengucapkan terima kasih kepada rekan-rekan dan keluarga yang selalu mendoakan dan memberi dukungan dalam menyelesaikan penelitian ini.

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I
PENDAHULUAN

1.1. Latar Belakang Masalah

Pada saat ini, media sosial tidak hanya menjadi sarana untuk menjalin pertemanan dengan banyak orang tanpa dibatasi ruang dan waktu. Media sosial juga banyak digunakan untuk hal-hal lain, misalnya sebagai sarana untuk mempromosikan produk hingga menjadi sarana untuk kampanye politik, seperti yang terjadi pada pemilihan presiden Republik Indonesia periode 2014-2019 yang mempertemukan Jokowi dan Prabowo sebagai lawan politik.

Rachmat dan Lukito (2016) pada jurnal penelitiannya yang berjudul "Sentipol: *Dataset* Sentimen Komentar pada Kampanye Pemilu Presiden 2014 dari Facebook Page" melakukan pembangunan *dataset* dengan mengumpulkan status dan komentar terhadap calon presiden Indonesia pada masa kampanye pemilu tahun 2014 dari facebook *page*. Penelitian tersebut menghasilkan sebuah *dataset* yang berisi 3400 komentar dari 68 status. Masing-masing komentar diberi label positif, negatif dan netral.

Pada jurnal penelitian tersebut, penulis juga juga telah mencoba melakukan klasifikasi pada *dataset* yang telah dihasilkan. Klasifikasi dilakukan dengan menggunakan metode *Naive Bayes* dan *Support Vector Machine* dimana keakuratan hasil klasifikasi tersebut masing-masing 83,32% dan 84,82%. Namun pada penelitian tersebut, Rachmat dan Lukito (2016) belum mencoba melakukan *clustering* pada *dataset* sentipol yang dibangun. Oleh karena itu, pada penelitian ini penulis akan mencoba menerapkan teknik *clustering* pada *dataset* sentipol.

Clustering merupakan salah satu teknik dalam data *mining* yang mengelompokkan suatu himpunan data ke dalam kelompok-kelompok (*clusters*) data yang serupa (Han & Kamber, 2006). Salah satu algoritma yang umum digunakan dalam *clustering* adalah algoritma *K-means*. Algoritma *K-means* sering digunakan dalam *clustering* karena kemudahan penggunaannya. Namun, kualitas

hasil akhir *clustering* dengan menggunakan algoritma *K-means* sangat bergantung pada pemilihan *centroid* (titik pusat *cluster*) awal (Alrabea, Senthilkumar, Al-Shalabi, & Bader, 2013). Pada algoritma *K-means*, pada umumnya pemilihan *centroid* awal dilakukan secara *random*, sehingga memungkinkan hasil akhir *clustering* berbeda meskipun menggunakan data uji yang sama (Fabregas, Gerardo, & Tanguilig III, 2017). Selain itu, jika *centroid* awal yang dipilih buruk, maka *cluster* yang dihasilkan bisa sangat tidak optimal.

Oleh karena itu, berdasarkan pada salah satu kelemahan algoritma *K-means* tersebut, Sujatha dan Sona (2013) melakukan penelitian mengenai algoritma *K-means* yang dimodifikasi, dengan menambahkan algoritma pada penentuan *centroid* awal. Hasilnya pun cukup baik, dalam proses *clustering*, algoritma yang digunakan membutuhkan jumlah iterasi dan waktu yang lebih sedikit dari algoritma *K-means* biasa. Berdasarkan pada penelitian yang dilakukan oleh Sujatha dan Sona (2013), penulis akan menggunakan algoritma *Modified K-means* pada sistem *clustering* yang akan dibangun, sehingga diharapkan mampu menghasilkan hasil *clustering* yang sama pada data uji yang sama pula. Selain itu, sistem yang dibangun juga diharapkan bisa menghasilkan *cluster* yang memiliki akurasi yang lebih baik dibandingkan algoritma *K-means* konvensional.

1.2. Perumusan Masalah

Rumusan masalah pada penelitian ini adalah sebagai berikut:

1. Bagaimana penerapan *Modified K-means Clustering* dalam mengklasterisasi data komentar pada *dataset* sentipol?
2. Bagaimana hasil analisis *clustering* data komentar pada *dataset* sentipol berdasarkan nilai *purity* yang dihasilkan?

1.3. Batasan Masalah

Sistem *clustering* sentipol yang akan dibangun memiliki beberapa batasan sebagai berikut:

1. *Dataset* sentipol diambil dari penelitian yang telah dilakukan oleh Rachmat dan Lukito (2016) melalui http://ti.ukdw.ac.id/~crowd/*dataset*.php.
2. *Dataset* yang akan diteliti menggunakan bahasa Indonesia.

3. Fokus penelitian hanya pada data komentar.
4. Daftar *stopword* diambil dari http://static.hikaruyuuki.com/wp-content/uploads/*stopword*_list_tala.txt.
5. *Stemming* dilakukan menggunakan library JSastrawi untuk bahasa pemrograman java.

1.4. Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah sebagai berikut:

1. Menerapkan metode *Modified K-Means Clustering* dalam mengklasterisasi data komentar pada *dataset* sentipol.
2. Menganalisis hasil *clustering* data komentar pada *dataset* sentipol berdasarkan nilai *purity* yang dihasilkan.

1.5. Manfaat Penelitian

Manfaat penelitian yang akan dilakukan adalah sebagai berikut:

1. Mengetahui seberapa baik *dataset* sentipol yang telah dibangun untuk proses pembelajaran *clustering* berdasarkan analisis *cluster*.
2. Dengan menggunakan algoritma *Modified K-Means*, diharapkan hasil *clustering* bisa memiliki nilai *purity* yang lebih besar, serta menghasilkan *cluster* yang lebih konsisten dibanding algoritma *K-Means* konvensional.

1.6 Metodologi Penelitian

Adapun metodologi penelitian yang akan dilakukan adalah sebagai berikut:

1. Studi Pustaka dan Pengumpulan Data

Tahap pertama pada penelitian ini adalah studi pustaka. Pada tahap ini, penulis mencari dasar-dasar teori terkait topik penelitian dari berbagai sumber referensi, seperti buku, jurnal, internet dan lain-lain. Selain itu, pada tahap ini juga dilakukan pengambilan *dataset* sentipol yang akan digunakan sebagai *input* dari sistem yang akan dibangun melalui link berikut http://ti.ukdw.ac.id/~crowd/*dataset*.php.

2. Perancangan Sistem

Pada tahap ini, dilakukan perancangan sistem *clustering* yang meliputi perancangan alur kerja sistem, perancangan *database*, dan perancangan antarmuka sistem yang akan dibangun.

3. Implementasi

Tahap ini merupakan tahap implementasi atau pembuatan sistem dengan mengacu pada rancangan yang telah dibuat pada tahap sebelumnya.

4. Pengujian Sistem

Pada tahap ini dilakukan pengujian sistem, apakah sistem yang dibangun sudah berjalan dengan baik dan sesuai dengan rancangan awal. Selain itu juga akan dilakukan evaluasi hasil *clustering* dengan menggunakan metode *purity*.

5. Penulisan Laporan

Proses terakhir dalam penelitian ini adalah penulisan laporan Tugas Akhir. Laporan berisi Bab 1 sampai dengan Bab 5 sesuai dengan format yang telah ditetapkan oleh pihak Universitas Kristen Duta Wacana.

1.7 Sistematika Penulisan

Sistematika penulisan laporan tugas akhir ini terdiri dari 5 bab. Bab 1 merupakan pendahuluan, yang berisi penjabaran mengenai latar belakang, rumusan masalah, batasan masalah, dan tujuan dilakukannya penelitian. Selain itu, dalam bab ini juga menjabarkan mengenai metode penelitian yang akan dilakukan serta sistematika penulisan laporan tugas akhir.

Bab 2 berisi tinjauan pustaka dan landasan teori. Tinjauan pustaka berisi pemaparan mengenai penelitian-penelitan yang pernah dilakukan sebelumnya yang terkait dengan penelitian yang akan dilakukan oleh penulis. Sedangkan landasan teori berisi penjelasan mengenai teori-teori yang terkait dengan penelitian yang akan dilakukan.

Bab 3 berisi tentang metodologi penelitian yang akan dilakukan. Bab ini memaparkan mengenai kebutuhan sistem, alat-alat yang dibutuhkan dalam pembuatan sistem, dan perancangan sistem yang meliputi perancangan alur kerja

sistem, perancangan *database* serta perancangan antarmuka sistem. Selain itu, pada bab ini juga akan dijelaskan mengenai rancangan pengujian yang akan dilakukan.

Bab 4 berisi hasil dan pembahasan. Bab ini memaparkan hasil dan pembahasan sistem yang telah dibangun sesuai dengan rancangan sistem pada bab 3. Selain itu, pada bab ini juga memaparkan hasil pengujian yang telah dilakukan. Sedangkan pada bab 5 berisi kesimpulan terkait penelitian yang telah dilakukan dan juga saran untuk pengembangan penelitian selanjutnya.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil pengujian dan analisis *cluster* yang telah dilakukan, maka penulis dapat mengambil beberapa kesimpulan dari penelitian mengenai Implementasi *Modified K-means Clustering* pada Pengelompokan Data Komentar Sentipol sebagai berikut:

1. Penggunaan *stemming* mampu meningkatkan kualitas *cluster* yang dihasilkan. Pada metode *K-means*, penggunaan *stemming* meningkatkan kualitas *cluster* sebesar 1,7%. Sedangkan pada metode *Modified K-means*, penggunaan *stemming* meningkatkan kualitas *cluster* sebesar 1,9% berdasarkan pada rata-rata nilai *purity* yang dihasilkan.
2. Metode *Modified K-means* menghasilkan kualitas *cluster* atau nilai *purity* yang lebih baik dibandingkan metode *K-means* biasa. Pada *clustering* yang menggunakan *stemming*, *Modified K-Means* mampu menghasilkan persentase rata-rata nilai *purity* sebesar 42% sedangkan *K-Means* menghasilkan persentase rata-rata nilai *purity* sebesar sebesar 39,1%. Sementara pada *clustering* yang tidak menggunakan *stemming*, *Modified K-Means* mampu menghasilkan persentase rata-rata nilai *purity* sebesar 40,1% sedangkan persentase rata-rata nilai *purity* yang dihasilkan pada *K-Means* sebesar 37,4%.
3. Faktor *random* pada penentuan *centroid* awal juga mempengaruhi kualitas dan konsistensi hasil *cluster* yang ada. Dari 5 kali pengujian dengan data yang sama, metode *Modified K-Means* yang dalam penerapannya menggunakan algoritma *modified centroid selection* dalam menentukan *centroid* awal tidak mengalami perubahan sama sekali, baik dari segi *purity* maupun *cluster* yang dihasilkan karena tidak adanya faktor *random*. Sementara pada metode *K-Means* yang mana terdapat faktor *random* pada pemilihan *cenroid* awal, selalu terjadi perubahan *cluster* pada setiap pengujian yang dilakukan sehingga mempengaruhi nilai *purity* yang dihasilkan.
4. Rata-rata persentase nilai *purity* yang dihasilkan dibawah 50% karena masalah pelabelan dan kurangnya jumlah *cluster* yang ditetapkan. Dari hasil pengujian

dengan label yang diperbaiki secara manual, terjadi peningkatan nilai *purity* sebesar 0,5%. Sedangkan pada pengujian penambahan jumlah *cluster*, dari jumlah *cluster* sebesar 3 sampai dengan 7, selalu terjadi peningkatan nilai *purity* setiap kali jumlah *cluster* ditambah.

5.2 Saran

Adapun saran untuk pengembangan yang lebih lanjut mengenai penelitian ini adalah sebagai berikut:

1. Perlu adanya penelitian lebih lanjut menggunakan *dataset* lain dengan topik yang berbeda sebagai perbandingan.
2. Melakukan pendekatan pengujian yang berbeda, misalnya dengan melakukan penghitungan *purity* secara manual atau tidak menggunakan label yang ada. Selain itu juga bisa dilakukan penelitian untuk mencari jumlah *cluster* yang terbaik.

DAFTAR PUSTAKA

Alrabea, A., Senthilkumar, A. V., Al-Shalabi, H., & Bader, A. (2013, June). Enhancing K-Means Algorithm with Initial Cluster Centers Derived from Data Partitioning along the Data Axis with PCA. *Journal of Advances in Computer Networks, 1*(2), 137-142. doi:10.7763/JACN.2013.V1.28

Fabregas, A. C., Gerardo, B. D., & Tanguilig III, B. T. (2017). Enhanced Initial Centroids for K-Means Algorithm. *International Journal of Information Technology and Computer Science, 9*(1), 26-33. doi:10.5815/ijitcs.2017.01.04

Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* United States of America: Cambridge University Press.

Han, J., & Kamber, M. (2006). *Data Mining Concepts and Techniques* (2 ed.). San Francisco: Elsevier Inc.

Kadhim, A. I., Cheah, Y. N., & Ahamed, N. H. (2014). Text Document Preprocessing and Dimension Reduction Techniques for Text Document Clustering. *International Conference on Artificial Intelligence with Applications in Engineering and Technology*, (hal. 69-73). Kinabalu.

Maimon, O., & Rokach, L. (2010). *Data Mining and Knowledge Discovery Handbook* (2 ed.). New York, USA: Springer Science+Business Media. doi:10.1007/978-0-387-09823-4

Manning, C. D., Raghavan, P., & Schutze, H. (2009). *An Introduction to Information Retrieval.* Cambridge, England: Cambridge University Press.

Rachmat, A., & Lukito, Y. (2016). Sentipol: Dataset Sentimen Komentar pada Kampanye Pemilu Presiden Indonesia 2014 dari Facebook Page. *Konferensi Nasional Teknologi Informasi dan Komunikasi*, 218-228. Retrieved from https://knastik.ukdw.ac.id/2016/makalah/artikel/e7-j1.pdf

Rudy. (2009). *Perbandingan Metode K-Means dan Hierarchical Agglomerative Clustering untuk Pengelompokan Dokumen Teks.* Duta Wacana Christian University. Special Region of Yogyakarta: Duta Wacana Christian University. Retrieved from http://sinta.ukdw.ac.id

Sujatha, S., & Sona, A. S. (2013, February). New Fast K-Means Clustering Algorithm using Modified Centroid Selection Method. *International Journal of Engineering Research & Technology*, *2*(2), 1-9. Retrieved from https://www.ijert.org/download/2353/new-fast-k-means-clustering-algorithm-using-modified-centroid-selection-method

Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2013). *Introduction to Data Mining* (2 ed.). New York: Pearson.