Text Mining Untuk Analisis Sentimen Review Film Menggunakan Algoritma K-Means

Text Mining For Movie Review Sentiment Analysis Using K-Means Algorithm

Setyo Budi
Jl. Imam Bonjol No. 207, Semarang, 50131, Telp. (024) 3517261
Fakultas Ilmu Komputer, Universitas Dian Nuswantoro Semarang
e-mail: setyoinator@gmail.com

Abstrak

Kemudahan manusia didalam menggunakan website mengakibatkan bertambahnya dokumen teks yang berupa pendapat dan informasi. Dalam waktu yang lama dokumen teks akan bertambah besar. Text mining merupakan salah satu teknik yang digunakan untuk menggali kumpulan dokumen text sehingga dapat diambil intisarinya. Ada beberapa algoritma yang di gunakan untuk penggalian dokumen untuk analisis sentimen, salah satunya adalah K-Means. Didalam penelitian ini algoritma yang digunakan adalah K-Means. Hasil penelitian menunjukkan bahwa akurasi K-Means dengan dataset digunakan 300 positif dan 300 negatif akurasinya 57.83%, 700 dokumen positif dan 700 negatif akurasinya 56.71%%, 1000 dokumen positif dan 1000 negatif akurasinya 50.40%%. Dari hasil pengujian disimpulkan bahwa semakin besar dataset yang digunakan semakin rendah akurasi K-Means.

Kata Kunci : Text Mining, Analisis Sentimen, K-Means, Review Film

Abstract

Human convenience in using the website resulting in increased text document in the form of opinions and information. For a long time text documents will increase. Text mining is one of the techniques used to explore a collection of text documents that can be taken essence. There are several algorithms used for extracting documents for sentiment analysis, one of which is a K-Means. In this study the algorithm used is K-Means. The results showed that the accuracy of the K-Means with the dataset used 300 positive and 300 negative 57.83% accuracy, 700 positive and 700 negative document accuracy of 56.71 %%, 1000 positive and 1000 negative documents accuracy of 50.40 %%. From the test results concluded that the larger the dataset used the lower the accuracy of the K-Means.

Keywords: *Text Mining, Sentiment Analysis, K-Means, Movie Review*

1. PENDAHULUAN

Kemudahan manusia didalam menggunakan website mengakibatkan bertambahnya dokumen teks yang berupa pendapat dan informasi. Dalam waktu yang lama dokumen teks akan bertambah besar. Banyaknya dokumen teks berasal dari berbagai sumber seperti *review*, opini, berita, paper, buku, perpustakaan digital, pesan e-mail dan halaman web. Teknik yang berkembang untuk penggalian dokumen teks saat ini adalah *text mining*. Text Mining merupakan suatu proses pengambilan intisari dari dokumen teks sehingga didapatkan hasil yang berguna untuk tujuan tertentu [1]. Kemudian Sesuai dengan buku *The Text Mining Handbook* [2], *text mining* dapat didefinisikan sebagai suatu proses menggali informasi dimana seorang *user* berinteraksi dengan sekumpulan dokumen menggunakan *tools* analisis yang merupakan komponen-komponen dalam *data mining* yang salah satunya adalah kategorisasi. Penambangan

dokumen teks dari website yang berisi komentar, pendapat, *feedback*, kritik dan *review* merupakan hal penting, karena apabila dikelola dengan baik maka dapat memberikan keuntungan berupa informasi yang bermanfaat untuk membantu individu atau organisasi didalam pengambilan sebuah keputusan [3]. Ada beberapa kategori yang termasuk didalam teknik *text mining* salah satunya adalah analisis sentimen, yaitu suatu proses memahami, mengekstrak, dan mengolah data tekstual secara otomatis, atau merupakan studi komputasi pendapat, perasaan dan emosi yang dinyakan dalam bentuk teks. Ada beberapa algoritma atau metode yang di gunakan untuk analisis sentimen, antara lain Naïve Bayes (NB) [4,5] Support Vector Machine (SVM) [4,5,6] dan *clustering* K-Means[3]. Didalam penelitian ini algoritma yang digunakan adalah K-Means dengan seleksi fitur Information Gain. Dengan tujuan untuk mengetahui kinerja algoritma K-Means tanpa seleksi fitur dan menggunakan seleksi fitur.

2. METODE PENELITIAN

Model penelitian yang digunakan didalam penelitian ini adalah model eksperimen dan evaluasi. *Tools* yang digunakan untuk eksperimen penelitian ini adalah RapidMiner versi 5. Tahapan penelitian ini adalah a) Pengumpulan Dataset, b) *Preprocessing*, c) Model yang diusulkan d) Eksperimen dan Pengujian Model, e) Evaluasi hasil.

2.1 Pengumpulan Dataset

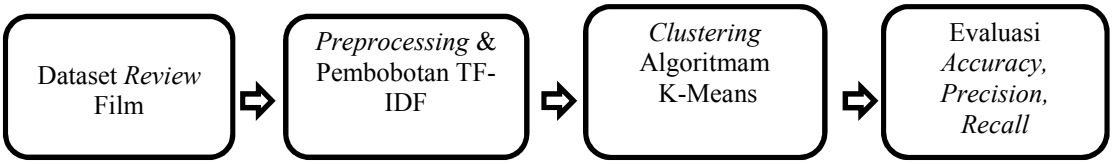
Dataset yang digunakan didalam penelitian ini adalah dataset *review* film yang diunduh dari http://www.cs.cornell.edu/People/pabo/movie-review-data/. [7].

2.2 Preprocessing

Preprosesing merupakan tahapan untuk mengubah struktur isi dari suatu dokumen kedalam format yang sesuai, berupa kumpulan *term* atau kata, agar dapat diproses oleh algoritma *clustering* [8].

2.3 Model yang diusulkan

Model yang diusulkan pada penelitian ini adalah seperti pada gambar 1.



Gambar 1. Model yang diusulkan

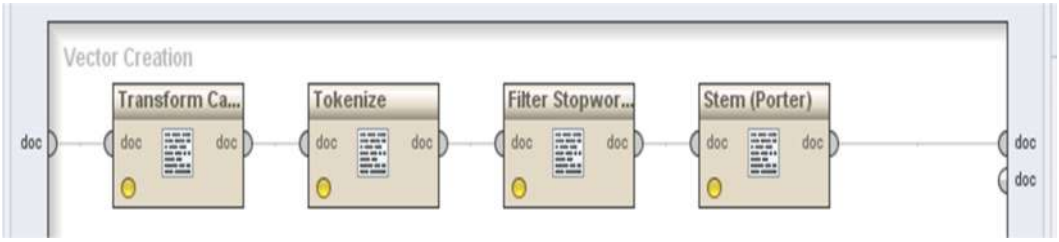
1. Dataset *Review* Film

   Hasil unduh dataset dari http://www.cs.cornell.edu/People/pabo/movie-review-data/ dikelompokkan menjadi 3 bagian yaitu : 1) 300 dokumen *review* positif dan 300 dokumen *review* negatif, 2) 700 dokumen *review* positif dan 700 dokumen *review* negatif, 3) 1000 dokumen *review* positif dan 1000 *review* dokumen negatif .

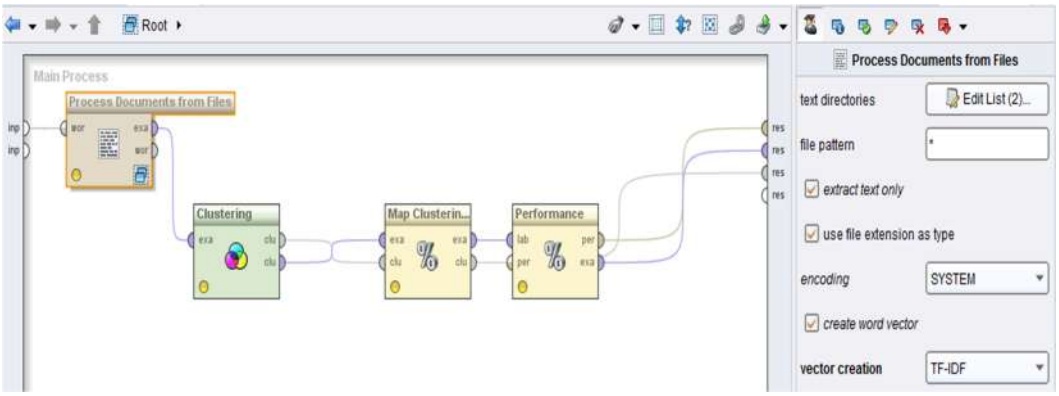
2. *Preprocessing* dan pembobotan TF-IDF

   Tahapan *preprocessing* yang digunakan dalam penelitian ini adalah Transform Cases/Case folding, Tokenizing, Stopword, Stemming (porter), dilanjutkan dan proses

pembobotan TF-IDF. Pembobotan TF-IDF (*Term Frequency - Inverse Document Frequency*) adalah salah satu algoritma yang digunakan untuk menghitung skor atau pembobotan *term* atau kata didalam suatu dokumen [9]. Proses *Preprocessing* disajikan pada gambar 2 dan untuk pembobotan TF-IDF disajikan pada gambar 3.



Gambar 2. Proses *View Preprocessing*



Gambar 3. Proses *View Preprocessing* dan TF-IDF

3. *Clustering* Algoritma K-Means

Metode K-Means diperkenalkan oleh James B Mac Queen pada tahun 1967 dalam Proceding of the 5th Berkeley Symposium on Mathematical Statistics and Probability [10]. *K-Means* adalah salah satu teknik *unsupervised learning* yang paling sederhana dan baik untuk memecahkan masalah *clustering* [11]. Berikut adalah *pseudocode* dari algoritma K-Means.

Algoritma K-Means *Clustering*

Input : Koleksi Dokumen *D*={*d1, d2, d3, ... dn*};

Jumlah *cluster* (*k*) yang akan dibentuk*;*

Output : *k cluster;*

Proses :

a. Memilih *k* dokumen untuk dijadikan *centroid* (titik pusat *cluster*) awal secara random*;*
b. Hitung jarak setiap dokumen ke masing-masing *centroid* menggunakan persamaan *cosines similarity* kemudian jadikan satu *cluster* untuk tiap-tiap dokumen yang memiliki jarak terdekat dengan *centroid*;
c. Tentukan *centroid* baru dengan cara menghitung nilai rata-rata dari data-data yang ada pada *centroid* yang sama;
d. Kembali ke langkah 2 jika posisi *centroid* baru dan *centroid* lama tidak sama;

Proses *clustering* dengan menggunakan algoritma K-Means sesuai model yang diusulkan apabila di formulasikan di dalam *tools* rapid miner adalah sesuai dengan gambar 3, dimana urutannya adalah Process Document From Files, Clustering, Map Clustering kemudian Performance. Dari model tersebut nantinya diketahui *accuracy*, *precision* dan *recall* dari algoritma K-Means.

4. Evaluasi *Accuracy*, *Precision* dan *Recall*
   Evaluasi dilakukan dengan mengamati kinerja algoritma K-Means meliputi *accuracy*, *precision* dan *recall*. Dalam penelitian ini digunakan *Confusion Matrix* untuk mengukur kinerja *clustering*.

Tabel 1: Tabel *Confusion Matrix*

		Predicted	
		Negative	Positive
Actual	*Negative*	A	b
	Positive	C	d

1. a adalah jumlah prediksi yang tepat bahwa instance bersifat negatif
2. b adalah jumlah prediksi yang salah bahwa instance bersifat positif
3. c adalah jumlah prediksi yang salah bahwa instance bersifat negatif
4. d adalah jumlah prediksi yang tepat bahwa *instance* bersifat positif

a. *Accuracy* (AC) adalah proporsi jumlah prediksi dataset yang benar. Hal ini ditentukan dengan menggunakan persamaan :

$$AC = \frac{a + d}{a + d + b + c} \quad (1)$$

b. *Recall* adalah rasio proporsi jumlah dokumen *review* positif yang diidentifikasikan dengan benar, yang dihitung dengan menggunakan persamaan :

$$R = \frac{d}{c+d} \quad (2)$$

c. *Precision* (P) adalah proporsi prediksi jumlah dokumen *review* positif yang benar, dihitung dengan menggunakan persamaan :

$$P = \frac{b}{b+d} \quad (3)$$

2.4 Eksperimen dan Pengujian Model
Setelah model yang diusulkan dibuat, selanjutnya dilakukan eksperimen dengan menguji model yang ada dengan dataset yang sudah dikelompokkan yang menjadi 3 kelompok seperti pada tahapan penelitian.

2.5 Evaluasi Hasil
Setelah pengujian model dilakukan maka langkah selanjutnya mengevaluasi hasil eksperimen, sehingga diketahui *accuracy*, *precision* dan *recall* dari algoritma K-Means.

3. HASIL DAN PEMBAHASAN

Didalam penelitian ini akan dilakukan pengujian model dengan menggunakan dataset yang sudah dikelompokkan menjadi 3 kelompok. Untuk pengujian model dibagi menjadi 3 skenario yaitu, skenario pertama : pengujian model dengan dataset 300 dokumen *review* positif dan 300 *review* dokumen negatif, skenario kedua: Pengujian model dengan dataset 700 dokumen *review* positif dan 700 *review* dokumen negatif, dan skenario ketiga : pengujian model dengan dataset 1000 dokumen *review* positif dan 1000 dokumen *review* negatif.

3.1 Skenario Pengujian

3.1.1 Skenario Pertama : Pengujian model dengan dataset 300 *review* positif dan 300 *review* negatif.

Skenario ini bertujuan untuk mengetahui kinerja algoritma K-Means dengan dataset yang digunakan sebanyak 300 dokumen *review* positif dan 300 dokumen *review* negatif. Hasil pengujian disajikan pada tabel 2.

Tabel 2**:** Hasil pengujian dengan 300 dokumen positif dan 300 negatif.

Dokumen	*Accuracy*	*Precision*	*Recall*
300 (+) & 300(-)	57.83%	60.83%	45.67%

3.1.2 Skenario Kedua : Pengujian model dengan dataset 700 *review* positif dan 700 *review* negatif.

Skenario ini bertujuan untuk mengetahui kinerja algoritma K-Means dengan dataset yang digunakan sebanyak 700 dokumen *review* positif dan 700 dokumen *review* negatif. Hasil pengujian disajikan pada tabel 3.

Tabel 3**:** Hasil pengujian dengan 700 dokumen positif dan 700 negatif.

Dokumen	*Accuracy*	*Precision*	*Recall*
700 (+) & 700(-)	56.71%	62.05%	34.57%

3.1.3 Skenario Ketiga : Pengujian model dengan dataset 1000 *review* positif dan 1000 *review* negatif.

Skenario ini bertujuan untuk mengetahui kinerja algoritma K-Means dengan dataset yang digunakan sebanyak 1000 dokumen *review* positif dan 1000 dokumen *review* negatif. Hasil pengujian disajikan pada tabel 4.

Tabel 4. Hasil pengujian dengan 1000 dokumen positif dan 1000 negatif.

Dokumen	*Accuracy*	*Precision*	*Recall*
1000(+) & 1000(-)	50.40%	50.30%	67.20%

3.2 Analisis *Accuracy*

Setelah ketiga skenario pengujian dilakukan, selanjutnya dilakukan analisis *accuracy* berdasarkan jumlah dataset. Analisis *accuracy* disajikan di tabel 5.

Tabel 5**:** A*ccuracy* K-Means dengan jumlah dataset berbeda

Dokumen	300 (+) &300 (-)	700 (+) & 700(-)	1000 (-) & 1000 (+)
Accuracy	57.83%	56.71%	50.40%

Tabel 5 menunjukkan bahwa *accuracy* K-Means dengan masing-masing jumlah dataset 300 dokumen *accuracy* =57.83%, 700 dokumen = 56.71% dan 1000 dokumen *accuracy* K-Means 50.40%, sehingga menurut hasil penelitian ini dapat simpulkan bahwa semakin besar dataset yang digunakan semakin kecil *accuracy* K-Means.

3.3 Analisis *Precision*

Untuk mengetahui *precision* dengan jumlah dataset *review* film yang berbeda, dapat dilihat dari hasil pengujian pada tabel 6 dibawah ini.

Tabel 6**:** *Precision* K-Means dengan jumlah dataset yang berbeda

Dokumen	300 (+) &300 (-)	700 (+) & 700(-)	1000 (-) & 1000 (+)
Precision	60.83%	62.05%	50.30%

Tabel 6 menunjukkan bahwa *precision* dengan jumlah dataset yang berbeda, *precision* dengan dataset masing-masing 300 dokumen adalah 60.83%, 700 dokumen 62.05% dan 1000 dokumen adalah 50.30%. Dari hasil pengujian tidak bisa disimpulkan bahwa semakin besar dataset yang digunakan semakin kecil *precision* K-Means. Terbukti *precision* dengan dataset 700 dokumen lebih besar daripada menggunakan dataset 300 dokumen.

3.4 Analisis *Recall*

Untuk mengetahui *recall* masing-masing pengujian dapat di lihat pada tabel 7.

Tabel 7. Perbandingan *Precision* K-Means

Dokumen	300 (+) &300 (-)	700 (+) & 700(-)	1000 (-) & 1000 (+)
Recall	45.67%	34.57%	67.20%

Tabel 7 Menunjukkan bahwa *recall* pada masing-masing pengujian dengan jumlah dataset seperti di tabel 7, bahwa *recall* untuk dataset 300 dokumen positif dan 300 dokumen negatif sebesar 45.67%, 700 dokumen positif dan 700 dokumen negatif sebesar 34.57% dan dengan dataset 1000 dokumen positif dan 1000 dokumen negatif sebesar 67.20%. Hasil ini tidak bisa di pastikan seperti *accuracy*, semakin besar dataset yang digunakan semakin kecil *recall* K-Means.

4. KESIMPULAN

1. Hasil pengujian menunjukkan bahwa *accuracy* algoritm K-Means dengan menggunakan dataset 300 dokumen *review* positif dan 300 dokumen *review* negatif adalah 57.83%, dataset 700 dokumen positif dan 700 dokumen negatif *accuracy* K-Means adalah 56.71% kemudian menggunakan dataset 1000 dokumen positif dan 1000 dokumen negatif *accuracy* K-Means adalah 50.40%.
2. Sesuai hasil pengujian sesuai dengan tabel 4 dapat disimpulkan bahwa semakin besar dataset yang digunakan maka semakin rendah *accuracy* K-Means.

5. SARAN

Dalam penelitian ini dataset yang digunakan adalah menggunakan dataset *review* film berjumlah 1000 dokumen *review* positif dan 1000 dokumen *review* negatif. Agar di ketahui lebih lanjut kinerja algoritma K-Means untuk *text mining* maka untuk penelitian selanjutnya akan menambah jumlah dataset yang digunakan. Selain itu karena dataset yang digunakan didalam dipenelitian ini berbahasa inggris, maka untuk penelitian selanjutnya dataset yang digunakan adalah *review* film berbahasa Indonesia.

DAFTAR PUSTAKA

[1] Witten, I. H., Frank E., Hall, M. A., 2011, Data Mining, Practical Machine Learning Tools and Techniques, Ed. 3, Burlington: Morgan Kaufmann, USA.

[2] Francis, L, Flynn, M., 2010, Text Mining Handbook, *Casualty Actuarial Society E-Forum*, p.1, Spring.

[3] G. Li and F. Liu, 2010, A Clustering-based Approach on Sentiment Analysis, in *Intelligent Systems and Knowledge Engineering (ISKE),* 2010 International Conference on 2010 Nov 15, pp.331-337, Australia, IEEE

[4] Stylios, G., Christodoulakis, D., Besharat, J., Vonitsanou, M., Kotrotsos, I., Koumpouri, A. and Stamou, S., 2010. Public opinion mining for Governmental Decisions. *Electronic Journal of e-Government*, Vol. *8*(2), pp.203-214.

[5] Keefe, T. O., Koprinska, I., 2009, Feature Selection and Weighting Methods in Sentiment Analysis, *Proceedings of the 14th Australasian Document Computing Symposium,* pp. 1-8, Dec 4.

[6] Abbasi, A., Chen, H., Salem, A., 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. *ACM Transactions on Information Systems (TOIS)*, *26*(3), p.12.

[7] Pang, B., Lee, L., 2004, A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts, *ACL '04 Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics,* Barcelona, Spain, July 21.

[8] L. Suanmali, L., N. Salim, N. dan Binwahlan, M.S., 2008, Automatic Text Summarization Using Feature Based Fuzzy Extraction, *Jurnal Teknologi Maklumat,* Jilid 20, Bil.2,

Desember.

[9] Baeza-Yates, R. and Ribeiro-Neto, B., 1999. *Modern information retrieval* (Vol. 463). New York: ACM press.

[10] MacQueen, J., 1967, June. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 14, pp. 281-297).

[11] Velmurugan, T. and Santhanam, T., 2010. Computational complexity between K-means and K-medoids clustering algorithms for normal and uniform distributions of data points. *Journal of computer science*, *6*(3), p.363.