JURNAL TIMES
TECHNOLOGY INFORMATICS & COMPUTER SYSTEM

Volume VIII No 2, Desember 2019
*p*ISSN : 2337 – 3601
*e*ISSN : 2549 – 015X

Tersedia *online* di http://ejournal.stmik-time.ac.id

PERBAIKAN PERFORMA CLUSTER K-MEANS MENGGUNAKAN SUM SQUARED ERROR (SSE) PADA ANALISIS ONLINE CUSTOMER REVIEW TERHADAP PRODUK TOKO ONLINE

Rena Nainggolan[1)] Eviyanti Purba[2)]

Komputerisasi Akuntansi, Universitas Methodis Indonesia[1,2)]
JL. Hang tuah, No. 8 Medan, Sumatera Utara-Indonesia[1,2)]
renanain99olan@gmail.com[1)]

Abstrak

Salah satu teknik dalam *Data Mining* yaitu *clustering*. C*lustering* adalah pengelompokan sejumlah data atau objek ke dalam *cluster* (*group*) sehingga setiap dalam *cluster* tersebut akan berisi data yang semirip mungkin dan berbeda dengan objek dalam *cluster* yang lainnya. *K-means* mempunyai kemampuan mengelompokkan data dalam jumlah yang cukup besar dengan waktu komputasi yang relatif cepat dan efisien. Namun, *K-means* mempunyai mempunyai kelemahan yang diakibatkan oleh penentuan pusat awal *cluster*. Hasil *cluster* yang terbentuk dari metode *K-means* ini sangatlah tergantung pada inisiasi nilai pusat awal *cluster* yang diberikan.Ini menyebabkan hasil *cluster*nya berupa solusi yang sifatnya *local optimal*. Penelitian ini melakukan pencarian pusat cluster yang paling optimum berbasis *Sum Of Squared* Error *(SSE),* diharapkan pusat cluster yang diperoleh nantinya akan menghasilkan *cluster - cluster*, dimana antar anggota *cluster* memiliki tingkat kemiripan yang tinggi. Perbaikan performa cluster K-Means akan diterapkan pada Analisis Penilaian Online Customer Review dan Rating pada Online Marketplace.

Kata Kunci : Clustering, *K-Means Clustering, Sum of Squared Error (SSE), Customer Online Review*

1. PENDAHULUAN

Pada tahun 2016 APJII (Asosiasi Penyelenggara Jasa Internet Indonesia), hasil survey menyatakan bahwa pengguna Internet Indonesia mencapai 132,7 juta dari populasi masyarakat Indonesia yaitu 256,2 juta jiwa. Berdasarkan data APJII, 63,1 juta orang menggunakan telepon genggam untuk internet. Dan 92,8 juta orang melakukan akses internet tidak tetap atau dimana saja.

Hal tersebut membuktikan bahwa setiap orang dapat terhubung ke internet dimanapun dan kapanpun. Dan situs yang paling sering dikunjungi adalah situs belanjan online shop, yaitu sebesar 82,2 juta orang. [1]

Pertumbuhan e-commerce tidak lepas dari perkembangan internet. Pada rentang tahun 2012-2015 Lembaga riset *Integrated Community Development (ICD)* Pennggunaan e-commerce meningkat sebesar 42% Angka ini lebih tinggi jika dibandingkan negara lain seperti Malaysia (14%), Thailand (22%), dan Filipina (28%) [2]. Perkembangan Online shop atau toko online melalui media internet sangat meningkat di Indonesia, bahkan di daerah terpencil. Banyaknya kemudahan dalam berbelanja online membuat para konsumen beralih untuk menggunakan fasilitas tersebut. Masyarakat hanya membutuhkan biaya berlangganan internet untuk mendapatkan fasilitas tersebut. Dalam toko online kita akan menemukan banyak sekali penjual, sehingga memudahkan kita untuk memilih barang sesuai dengan keinginan kita, hal ini membuat persaingan harga dan kualitas produk yang sangat bersaing, sehingga konsumen bisa membandingkan harga yang ditawarkan oleh para penjual.

Untuk itu dalam melakukan transaksi secara online sangat dibutuhkan kepercayaan antara si penjual dan si pembeli, salah satu faktor yang sangat mempengaruhi si konsumen untuk

membeli barang adalah dengan mengetahui riwayat si penjual dan bagaimana produk yang ditawarkan si penjual tersebut, hal in bisa di ketahui oleh si penjual dengan melihat alasan tentang produk tersebut yang bisa di baca oleh calon pembeli melalaui ulasan produk yang ada di situs toko online tersebut.

2. TINJAUAN PUSTAKA

Data Mining

Ada dua istilah dalam data mining yaitu; seperti knowledge discover ataupun pattern recognition. Masing-masing mempunyai arti yang berbeda beda dan memiliki ketepatan satu sama yang lain, istilah knowledge discovery atau penemuan pengetahuan tepat karna digunakan tujuan utama dari data mining memang untuk mendapat pengetahuan yang masih tersembunyi di dalam bongkahan data [3,4] sedangkan untuk Istilah pattern recognition atau pengenalan pola pun tetap untuk digunakan karena pengetahuan yang hendak digali dalam bongkahan data yang tengah dihadapi dan masih dalam berbentuk pola-pola yang juga masih perlu digali dari [4]

K-Means Clustering

Metode K-means clustering merupakan metode clustering yang dikenalkan oleh [5]. Metode Kmeans merupakan metode yang terkenal simpel dan cepat [6].

Metode K-Means bekerja dengan mengelompokkan data atau objek yang memiliki karakter yang sama dalam satu cluster/kelompok dan akan mengelompokkan data/objek ke cluster yang lain yang mempunyai karakteristik yang berbeda dan pada akhirnya akan menghasilkan suatu cluster atau kelompok yang memiliki tingkat kemiripan yang sangat tinggi. Langkah-langkah melakukan clustering dengan metode K-means adalah sebagai berikut [7]

1. Algoritma K-Means Algoritma k-means adalah algoritma yang mempartisi data ke dalam cluster – cluster sehingga data yang memiliki kemiripan berada pada satu cluster yang Pilih jumlah *Cluster* k.
2. Inisialisasi k pusat *Cluster* ini bisa dilakukan dengan berbagai cara. Namun yang paling sering dilakukan adalah dengan cara random. Pusat-pusat *Cluster* diberi nilai awal dengan angka-angka random.
3. Alokasikan semua data/ objek ke *Cluster* terdekat. Kedekatan dua objek ditentukan berdasarkan jarak kedua objek tersebut. Demikian juga kedekatan suatu data ke *Cluster* tertentu ditentukan jarak antara data dengan pusat *Cluster*. Dalam tahap ini perlu dihitung jarak tiap data ke tiap pusat *Cluster*. Jarak paling antara satu data dengan satu *Cluster* tertentu akan menentukan suatu data masuk dalam *Cluster* mana. Untuk menghiutng jarak semua data ke setiap tiitk pusat *Cluster* dapat menggunakan teori jarak Euclidean yang dirumuskan sebagai berikut:

   dimana:
   $D(i,j)$ = Jarak data ke i ke pusat *Cluster* j
   xki = Data ke i pada atribut data ke k
   Xkj = Titik pusat ke j pada atribut ke k

   Hitung kembali pusat *Cluster* dengan keanggotaan *Cluster* yang sekarang. Pusat *Cluster* adalah rata-rata dari semua data/ objek dalam *Cluster* tertentu. Jika dikehendaki bisa juga menggunakan median dari *Cluster* tersebut. Jadi rata-rata (mean) bukan satu-satunya ukuran yang bisa dipakai.
4. Tugaskan lagi setiap objek memakai pusat *Cluster* yang baru. Jika pusat *Cluster* tidak berubah lagi maka proses *Clustering* selesai. Atau, kembali ke langkah nomor 3 sampai pusat *Cluster* tidak berubah lagi.

Online Customer Review

Review merupakan bagian dari Electronic Word of Mouth (eWOM), [8] yaitu merupakan pendapat langsung dari seseorang dan bukan sebuah iklan. Review adalah salah satu dari beberapa faktor yang menentukan keputusan pembelian seseorang. menunjukkan bahwa orang dapat mengambil jumlah review sebagai indikator popularitas produk atau nilai dari suatu produk yang akan mempengaruhi kemauan untuk membeli suatu produk. Namun belum tentu semakin banyak review dan rating berarti produk tersebut pasti akan dibeli oleh pelanggan. Banyak faktor-faktor yang menjadi alasan keputusan pembelian suatu produk bagi pelanggan. Online review dapat menjadi alat promosi yang ampuh untuk komunikasi pemasaran. Pemasar dan vendor telah menggunakan media ini karena memberikan saluran yang murah dan berdampak untuk menjangkau pelanggan mereka. Pemasar diketahui telah mengambil keuntungan dari jaringan pengaruh antara pelanggan unt uk mempengaruhi perilaku pembelian pembeli potensial.

3. PEMBAHASAN

A. Pra Proses Data

Secara umum terdapat empat tahapan yang diselesaikan pada model penelitian ini yaitu tahap pengumpulan data ulasan produk, tahap *preprocessing* data, tahap seleksi fitur, pengujian model klasterisasi MK-Means serta pengujian performa *clustering.*

1. Pengumpulan data

Pengumpulan data (*data crawling*) bertujuan untuk menjaring data ulasan produk. Penelitian ini menggunakan data ulasan produk yang diperoleh dari situs jual beli *online*. Data dikumpulkan menggunakan aplikasi *Octoparse* yaitu sebuah aplikasi *open source* untuk *web crawler* [9]

2. Text pre-processing

Pemrosesan awal atau *text pre-processing* merupakan tahap kedua pada *text mining*.[10]. Tahap Pemrosesan awal bertujuan untuk mempersiapkan data untuk dapat dipakai pada tahap penemuan pola, misalnya mengeliminasi data yang mengandung *noise,* data yang tidak lengkap (*incomplete*) dan data yang tidak konsisten (*inconsistent*).

3. *Seleksi Fitur*

Pada tahap ini terdapat dua proses yang dilakukan, adalah sebagai berikut :

1. Case Folding

Case Folding adalah proses mengubah huruf besar menjadi huruf kecil semua.

Contoh:

```
Saya bErmain Petak Umpet
PAK MAU LAPOR
Menjadi :
saya bermain petak umpet
pak mau lapor
```

2. Non Alphanumeric Removal

Alphanumeric menghilangkan tanda baca angka.

Contoh:

```
Saya tidur... kemarin
TOLONG DIBENAHI!!! kapan beresnya??!?
Menjadi:
saya tidur kemarin
TOLONG DIBENAHI kapan beresnya
```

3. Own Stop Words Removal

Proses Own Stop Words Removal yaitu menghilangkan emotion atau ekspresi

Contoh:

```
harga cabai Rp 15.000,00
harga cabai rp 15.000,00
bbm koq naik, warga sedih #edisicurhat
telah blokir website http://www.lucu.com
Menjadi
harga cabai Rp 15.000,00
harga cabai 15.000,00
bbm naik, warga sedih
telah blokir website
```

4. *Stop Words Removal*
   Menghapus kata yang dapt mempengaruhi hasil dan tidak menghapus kata yang akan mempengaruhi hasil.
   Contoh:
   ```
   Pak kepala desa tidak tahu bahwa 3 pencuri
   di rumah itu adalah teman lamanya!
   Menjadi
   pak kepala desa tahu 3 pencuri
   rumah teman
   ```

5. *Stem*
   Proses stem yaitu menghilangkan kata awalan dan akhiran.
   Contoh:
   ```
   Mempermainkan peranan 12 domba di pementasan
   Menjadi
   main peran 12 domba di pentas
   ```

4. *Stopword Removal*
   Stop words adalah kata umum yang biasanya muncul dalam jumlah yang besar namun tidak memiliki makna. Dalam bahasa indonesia seperti "*di*", "*dengan*", "*ke*", "*yang*", "*jika*", "*akan*", dan lain sebagainya. Untuk itu perlu dilakukan penghapusan. Untuk melakukan proses penghapusan kata ini diperlukan sebuah data atau daftar kata yang diinginkan untuk dihapus.
5. *Stemming*
   Stemming pada penelitian ini didasarkan pada algoritma Nazief dan Andriani. Algoritma ini dikenal juga dengan algoritma *confix stripping* yaitu algortima khusus untuk *stemming* teks berbahasa indonesia (Mardiana, *et al*, 2016).

Setelah semua data ditransformasi ke dalam bentuk angka, maka data tersebut telah dapat dikelompokkan dengan menggunakan metode *K-Mean Clustering*. Untuk dapat melakukan pengelompokan data tersebut menjadi beberapa *cluster* perlu dilakukan beberapa langkah yaitu:

1. Tentukan telebih dahulu jumlah *cluster* yang diinginkan. Dalam penelitian ini data yang ada akan dikelompokan menjadi dua cl*uster*.
2. Tentukan titik pusat awal dari setiap *cluster*. Dalam penelitian ini titik pusat awal dibangkitkan secara random. Pusat *cluster* pada solusi awal dapat dilihat pada table 1

4. ANALISA DAN HASIL

1. Hasil Pengumpulan data
   Penelitian yang dilakukan menggunakan customer online reviews yang terdiri dari 888 data. Dari 888 data tersebut terdapat 806 komentar positif dan 82 komentar negatif.

1. kualitas bagus nyaman pakai moga awet
2. bahan bagus besar untung retur ganti ukur

3.	sesuai skripsi nyaman moga awet mantap dahh
4.	mantap barang cepat barang bagus layan bagus ☺
5.	bagus bahan sesuai gambar
6.	bagus…!
7.	terimakasih barang tidak kecewa
8.	beda gambar sama barang yang datang…!
9.	BARANG JELEK!
10.	barang besar puas terimakasih lazada

2. Pre Processing Data

Setelah data ulasan produk telah dilakukan , maka selanjutnaya dilakukan pre-processing agar data customer online reviews dapat diterapkan pada algoritma clustering. Tahapan pre-processing yang diterapkan adalah Case Folding, Non Alpha Numeric Removal, Stop words Removal, dan Stemming. Daftar stop words untuk bahasa indonesia terdiri dari 760 kata **(**Tala, 2003). Algoritma Stemming yang diterapkan adalah algoritma stemming khusus untuk bahasa Indonesia yaitu algoritma nazief-Andriani (Adriani, et al, 2007).

1.	Kemas hancur
2.	barang lama packing hancur
3.	barang datang lama
4.	kecewa tidak sesuai harap
5.	barang pesan
6.	barang tidak sesuai deskripsi
7.	barang pesan sekarang
8.	pesan
9.	tidak sesuai ukuranya
10.	lama sampai

3. String to word vector

Untuk mengubah data string menjadi vector kata di terapak algoritma TF-IDF. Hasil Penerapan TF_IDF menhasilkan matriks data dengan dimensi 70 atribut x 888 data. Terdapat 70 term pada data sebagaimana ditampilkan pada gambar berikut.

cantik, cepat,lekas, bagus, rusak, jelek, buruk, awet, mantap, bolong, cacat, sesuai, ribet, aneh, sakit, syukur sampai, ramah, profesional, tukar, telat, salah, lentur, kaku

4. Konversi Data ke Numeric

Tabel berikut merupakan contoh data akhir hasil konversi

19 0.005376,23 0.005957,27 0.005897,34 0.025563,37 0.01063,43 0.084699,72 0.504275,91 0.017341,94 0.006041,95 0.005957,99 0.010738,102 0.018866,112 0.957791,128 0.027989,131 0.034241,136 0.021461,148 1.479877,155 0.00586,169 0.121277,170 0.005802,176 0.025946,178 0.005328,188 0.026011,193 0.005868,206 0.005904,211 0.131904,1
9 0.12829,13 0.035806,18 0.005942,26 0.01387,29 0.014772,30 1.685916,

5. Attribute Selection

Data diatas masih terlalu besar dan tidak efektif, sehingga atribut yang ada harus di filter. Dengan menggunakan algoritma Cfs, sebanyak 17 atribut.

bagus, beda, besar, bohong, cacat, cepat, jelek, kecewa, lama, mantap, mengelupas, parah, rusak, suka, telat, terimakasih

6. Pengkodena K-Mean Clustering

Tabel 1. Pengkodean K-Means Clutering

Atrribute	Cluster 0	Cluster 1
Bagus	0.2675	0.0912
	0.2515	0.1941
Beda	0	0.1943
	0	0.784
Besar	0.1364	0.0315
	0.491	0.2427
Bohong	0	0.0389
	0.1579	0.4262
Cacat	0	0.0978
	0.229	0.6135
Cepat	0.1772	0.0642
	0.4984	0.3118
Jelek	0	0.3331
	0.0008	0.9205
Kecewa	0	0.8459
	0.4621	0.9738
Lama	0	0.1719
	0.2839	0.7522
Mantap	0.0138	0.6703
	0.1691	0.9731
Mengelupas	0	0.0389
	0.1579	0.4262
Parah	0	0.1239
	0.2509	0.6698
Rusak	0	0.0389
	0.1579	0.4262
Suka	0.1982	0.598
	0.4977	0.2882
Telat	0	0.0389
	0.1579	0.4262
Terimakasih	0.2336	0.0608
	0.4797	0.2653

7. Langkah selanjutnyan dilakukan penentuan cluster

Tabel 2. Hasil Cluster

Class	Cluster	
	1	2
1	733.0782	61.9128
2	35.989	0.2653
Total	769.0762	122.9238

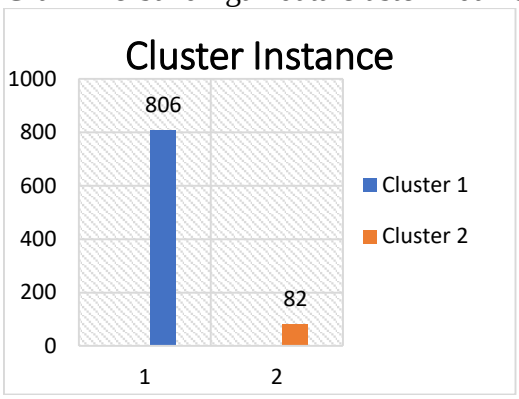
8. Clustre Instances
   Tabel 3. Cluster Instance

Cluster		Persentase
1	2	91%
806	82	9%

Dari tabel diatas dapat disimpulkan bahwa hasil pengujian terhadap 888 ulasan menghasilkan 2 cluster yaitu:

a. Cluster 1 yaitu menghasilkan sebanyak 806 (91%) ulasan yang mempunyai kemiripan yang sangat tinggi yang dikelompokkan menjadi 1 cluster
b. Cluster 2 menghasilkan 82 (9%) ulasan yang mempunyai kemiripan yang sangat tinggi yang dikelompokkan kmenjadi 1 kelompok cluster.

9. Grafik Perbandingan data cluster 1 dan cluster 2 dapat ditunjukkan pada grafik berikut.



Grafik Perbadingan data cluster 1 dan 2

5. KESIMPULAN

Berdasarkan pengujian yang dilakukan pada analisis Customer Online Reviews pada toko online dengan jumlah ulasan sebanyak 888 data, Penelitian ini menawarkan sebuah model cluster K- Means Clustering, dan pengujian menghasilakan 2 cluster, dan pengujian menghasilkan sebuah tools untuk membantu konsumen dalam mengambil sebuah keputusan untuk membeli sebuah produk mapun jasa, Karena pentingnya online customer reviews untuk meningkatkan kepercayaan terhadap perusahaan dengan cara meningkatkan kredibilitas dari penjual.

6. DAFTAR PUSTAKA

[1] Yoviriska, I. U & Wahjoedi (2018). Trend keputusan Belanja Online Mahasiswa Fakultas Ekonomi UM Angkatan 2014
[2] Sidharta, S. & Suzanto, B. (2015). *Pengaruh kepuasan transaksi online shopping dan kepercayaan konsumen terhadap sikap serta perilaku konsumen pada e-commerce*. Jurnal computech & Bisnis. 9(1) : 23-36
[3] Berry, M.W. and Browne, M., 2006. Lecture notes in data mining. World Scientific. [4] Susanto, S. and Suryadi, D., 2010. Pengantar data mining: mengagali pengetahuan dari bongkahan data.
[4] Mardiana, T., Adji, T. B., & Hidayah, I. (2016). Stemming Influence on Similarity Detection of Abstract Written in Indonesia. TELKOMNIKA (*Telecommunication Computing Electronics and Control*), 14(1), 219-227
[5] Agusta, Y., 2007. K-Means Penerapan, Permasalahan dan Metode Terkait. Jurnal Sistem dan Informatika, 3(1), pp.47-60.
[6] Lloyd, S., 1982. Least squares quantization in PCM. IEEE transactions on information theory, 28(2), pp.129-137.

[7] Lee, E.-J. & Shin, S.Y. (2014). When do consumers buy online product reviews? Effects of review quality, product type, and reviewer's photo. Computers in Human Behavior, 31, 356–366.
[8] Arthur, D. and Vassilvitskii, S., 2006, June. How slow is the k-means method?. In Proceedings of the twenty-second annual symposium on Computational geometry (pp. 144- 153). ACM.
[9] Ganjisaffar, Y. (2013). Crawler4j v. 3.5. URL http://code.google.com/p/crawler4j/.[dikutip 2017-05-13].
[10] Kumar, L., & Bhatia, P. K. (2013). Text Mining: Concepts, Process and Applications. *Journal of Global Research in Computer Science*, *4*(3), 36-39.