IMPLEMENTASI ALGORITMA K-MEANS CLUSTERING PADA ANALISIS SENTIMEN KELUHAN PENGGUNA INDOSAT

[1]Try Iryanto Saputra, [2]Rini Arianty
[1,2]Fakultas Ilmu Komputer Universitas Gunadarma
Jl. Margonda Raya No. 100, Depok 16424, Jawa Barat
[2] rinia @staff.gunadarma.ac.id,

Abstrak

Penyampaian keluhan konsumen lewat akun media sosial seperti Twitter dimaksudkan agar masalah yang dihadapi konsumen dapat diselesaikan dengan cepat. Pada penelitian ini, akan dilakukan analisis sentimen terhadap konsumen pengguna provider Indosat, menggunakan data tweet sejumlah 300 data acak yang di kumpulkan dari bulan desember 2018 hingga bulan april 2019. Data yang dianalisis adalah kalimat berbahasa Indonesia. Preprocessing pada penelitian ini terdiri dari beberapa tahapan proses antara lain tokenizing, filtering, stop word, dan stemming. Analisis dilakukan menggunakan metode K-Means Clustering. Penelitian ini berhasil menampilkan kelompok dari anggota masing-masing cluster yang berbentuk wordcloud ke dalam 3 buah wordcloud berbeda, pada wordcloud cluster 0 anggotanya berbicara tentang jaringan Indosat yang parah, pada wordcloud cluster 1 anggotanya berbicara tentang permintaan perbaikan jaringan sinyal Indosat, dan pada wordcloud cluster 2 anggotanya berbicara tentang jaringan sinyal parah Indosat pada daerah Bogor. Hasil penelitian ini diharapkan dapat menjadi masukan untuk provider dalam melihat keluhan yang masuk dari para konsumen mereka sehingga pihak provider dapat meningkatkan pelayanannya.

***Kata Kunci**: Filter, K-Means Clustering, stemming, token, Tweet.*

Abstract

Submitting a consumer complaint through this social media such as Twitter account responds so that the problems the consumer asks can be resolved quickly. In this study, sentiment analysis will be carried out by Indosat user opinions, using tweet data based on 300 random data collected from December 2018 to April 2019. The successful data is the Indonesian dialogue sentence. Preprocessing in this study consists of several processes including tokenizing, filtering, stop word, and stemming. The analysis was performed using the K-Means Clustering method. This research succeeded in displaying the groups of each group consisting of wordcloud into 3 different wordcloud, in wordcloud cluster 0 the members talked about bad Indosat networks, in wordcloud cluster 1 the members needed the help of Indosat communication networks, and in wordcloud cluster 2 its members talked about Indosat's bad signal network in Bogor area. The results of this study are expected to be input for providers in seeing complaints coming from their customers so that providers can improve their services.

***Key Words** : Filter, K-Means Clustering, stemming, token, Tweet*

PENDAHULUAN

Banyaknya pengguna media sosial aktif dan banyaknya manfaat yang diberikan oleh media sosial, membuat perusahaan ikut membuat sebuah sarana yang dapat menampung masalah yang dihadapi kliennya melalui sebuah akun media sosial. Salah

satunya adalah perusahaan *provider* di Indonesia seperti XL, Indosat, Smartfren dan lainnya, membuat akun layanan untuk menampung masalah konsumen mereka. Tujuan penyampaian keluhan konsumen lewat akun media sosial ini dimaksudkan agar masalah yang dihadapi konsumen dapat diselesaikan dengan cepat. Sayangnya dibalik manfaatnya tersebut para konsumen tersebut menggunakan kata-kata kasar untuk mengekspresikan kekecewaan atas pelayanan perusahaan tersebut. Media sosial yang digunakan perusahaan untuk akun keluhan konsumen salah satunya adalah twitter. Postingan *tweet* yang dikirim untuk perusahaan tidak hanya sebuah *tweet* yang berisi masalah, ada juga sebuah bentuk *tweet* yang mendukung dan mengapresiasi hasil dari pelayanan perusahaan tersebut. Dari masalah tersebut peneliti tertarik membuat penelitian Tentang analisis sentimen data *tweet* opini konsumen terhadap layanan *provider* Indosat. Analisis sentimen atau *opinion mining* adalah studi komputasional dari opini-opini orang, sentimen [1] dan emosi melalui entitas dan atribut yang dimiliki yang diekspresikan dalam bentuk teks [2]. Tugas dasar dalam analisis sentimen adalah mengelompokkan teks yang ada dalam sebuah kalimat atau dokumen [3] kemudian menentukan pendapat yang dikemukakan dalam kalimat atau dokumen tersebut apakah bersifat positif, negatif atau netral [4]. Analisis sentimen adalah proses yang digunakan untuk menentukan opini, emosi dan sikap yang dicerminkan melalui teks, dan biasanya diklasifikasikan menjadi opini positif dan negatif[5].

Beberapa penelitian dilakukan terkait sentimen analisis khususnya terhadap data sosial media. Penelitian dilakukan [6] menggunakan metode K-Means *Clustering*. Percobaan dilakukan pada 1000 data yang dikelompokan menjadi lima *cluster* yaitu cinta, marah, sedih,senang dan takut .Implementasi metode K-Means pada penelitian tersebut menghasilkan akurasi sebesar 76,3%. Selanjutnya penelitian dilakukan [7] menggunakan metode K-Means untuk optimasi klasifikasi tema tugas akhir mahasiswa menggunaan *Support Vector Machine* (SVM). Hasil yang diperoleh memiliki tingkat akurasi yang lebih baik yaitu 86,21%. Penelitian yang dilakukan [8] menggunakan dataset positif dan dataset negatif, untuk analisis sentimen *review* film menggunakan algoritma K-Means. Dari setiap uji coba menggunakan data dengan jumlah berbeda menunjukan bahwa akurasi mencapai 57.83% didapatkan menggunakan 300 dataset positif dan 300 dataset negatif, sedangkan 700 dataset positif dan 700 dataset negatif menunjukan akurasi mencapai 56.71%, jika menggunakan 1000 dataset positif dan 1000 dataset negatif akurasinya mencapai 50,40%. Penelitian [9] melakukan penelitian analisis sentimen kurikulum 2013 pada twitter menggunakan *Ensemble Feature* dan metode K-Nearest Neighbor. Penelitian ini digunakan untuk mengetahui opini yang berkembang

mengenai kurikulum 2013 yang dibagi kedalam opini positif atau opini negatif. Berdasarkan hasil pengujian dengan menggabungkan kedua fitur tersebut mendapatkan hasil akhir akurasi mencapai 96%. Penelitian analisis sentimen [10] data komentar sosial media facebook dengan K-Nearest Neighbor dengan studi kasus pada akun jasa ekspedisi barang J&T. dalam tahapan penelitiannya dilalui melalui beberapa tahap yaitu *preprocessing* yang terdiri dari *case folding, tokenizing, stopword removal*, dan *stemming*. Hasil yang didapatkan dari implementasi metode KNN dengan akurasi tertinggi adalah 79,21% sedangkan akurasi terendahnya adalah 70,3%.

METODE PENELITIAN

Analisis sentimen terhadap data keluhan pengguna Indosat lewat media sosial twitter terdiri atas tahapan preprocessing kalimat berupa case folding, tokenization, filterisasi, stopword, stemming dan labelisasi menggunakan TF-IDF yang kemudian akan dilakukan clustering menggunakan K-means *clustering* untuk mendapatkan *wordcloud*.

Data Tweet

Pengumpulan data dilakukan secara manual dengan mengumpulkan komentar pada website twitter dengan alamat https:// twitter.com/IndosatCare. Data komentar yang dikumpulkan dari @IndosatCare diambil dari bulan desember 2018 hingga bulan April 2019. Komentar yang diambil sebagai data berdasarkan komentar opini dari konsumen @IndosatCare terhadap layanan yang diberikan oleh @IndosatCare terhadap konsumen dengan jumlah data yang dikumpulkan ialah 300 data komentar yang terdiri dari komentar positif dan negatif seperti dapat dilihat pada Gambar 1. Data kemudian simpan data dengan ekstensi CSV(*Comma Separated Values File)*.

Gambar 1. Pengambilan Data KomentarPP

Preprocessing

Preprocessing pada penelitian ini terdiri atas tahapan sebagai berikut :

1. *Case Folding*. Data komentar akan di proses untuk merubah semua bentuk huruf besar yang ada pada data komentar dirubah menjadi huruf kecil.
2. *Tokenizing*. Data komentar yang telah masuk ke proses *tokenizing*, akan berubah menjadi potongan kata yang masing-masing katanya memiliki nilai yang berbeda.
3. *Filtering* dan *Stopword*. Pada tahap *filterin*g merupakan tahapan ketiga dari proses *preprocessing*, pada tahap ini dilakukan penghapusan tanda baca dan hashtag yang tidak memiliki arti. Pada proses ini kata-kata yang memiliki nilai kecil pada data komentar akan dihi-langkan dan akan disesuaikan dengan daftar *stopword*.
4. *Stemming*. Pada tahap *stemming* meru-pakan lanjutan dari tahap *filtering* dan *Stopword*, data komentar yang telah masuk tahap *filtering* dan *Stopword* yang berupa penggalan kata dasar dari suatu komentar akan dirubah menjadi bentuk kata baku bahasa Indonesia yang baik dan benar.

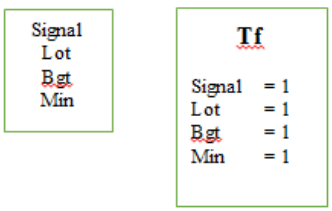
Pembobotan *Term Weighting* TF-IDF

Setelah melakukan proses *pre-processing*, selanjutnya masuk ke proses pembobotan kata. Pembobotan kata dilakukan untuk memberikan sebuah nilai terhadap suatu kata berdasarkan tingkat kepentingan kata pada kumpulan dokumen yang telah dimasukan ke dalam sistem. Sebagai contoh digunakan lima kalimat komentar sebagai salah satu contoh perhitungan, dengan uraian satu kalimat digunakan sebagai data uji dan empat kalimat digunakan sebagai data latih.

Tahap pembobotan kata terdapat empat langkah yang harus dilakukan, diantaranya sebagai berikut:

1. Menghitung jumlah *term frequency* (*tf*) tiap kata. Setelah melakukan tahap *pre-processing*, langkah selanjutnya ialah menghitung jumlah *term frequency*(tf), penghitungan dilakukan dengan memecah kalimat menjadi kata perkata seperti dapat dilihat pada Gambar 2.
2. Menghitung jumlah *document frequency* (*df*) tiap kata.
3. Menghitung *inverse document frequency* (*idf*) menggunakan persamaan 1.

$$idf = log \frac{N}{df} \quad (1)$$



Gambar 2. Contoh Perhitungan *Term Frequency* pada Kalimat

Contoh:

Pada kata "jam" hanya terdapat didalam satu dokumen saja, maka perhitungannya adalah sebagai berikut :

$$idf = \log\frac{5}{1} = \log 5 = 0.6989700043$$

Pada kata "gue" ditemukan pada dua dokumen yang sama, maka perhitungannya adalah sebagai berikut :

$$idf = \log\frac{5}{2} = \log 2.5 = 0.3979400087$$

Kata "min" ditemukan pada tiga dokumen yang berbeda, maka perhitungannya adalah sebagai berikut:

$$idf = \log\frac{5}{3} = \log 1.66666667 = 0.2218487505$$

4. Menghitung bobot atau *weight* (*w*) dari hasil perkalian *term frequency* (*tf*) dengan *inverse document frequency* (*idf*). Langkah pertama dilakukan dengan mengkalikan nilai tf dengan idf, hasil dari perkalian seluruh kata dijumlahkan untuk mendapatkan bobot atau *weight* (w) dari setiap kalimat dapat dilihat pada Persamaan 2.

$$tfidf = tf(t,d) * idf \quad (2)$$

Persamaan 3 digunakan untuk mendapatkan nilai bobot atau *weight* dari keseluruhan kalimat menggunakan rumus sebagai berikut :

$$w = (tfidf_1 + tfidf_2 + \cdots + tfidf_n) \quad (3)$$

Contoh hasil perhitungan tf-idf kalimat negatif pada masing-masing kata dan nilai bobot dalam setiap kalimat disajikan pada Tabel 1.

Analisis Metode K-Means *Clustering*

Langkah-langkah yang dilakukan untuk menganalis menggunakan metode K-Means yaitu:

1. Menentukan jumlah *k cluster*. Jumlah *cluster* yang digunakan ialah 2 *cluster*.
2. Menentukan *centroid* awal (C_i). Setelah melakukan *pre-processing* dan pembobotan (w), maka setiap bobot kalimat akan digunakan untuk perhitungan K-Means *clustering*. Berikut data yang didapat dari perhitungan bobot (w):
   1) Kalimat 1 (k_0) = 9.08661006
   2) Kalimat 2 (k_1) = 2.79588002
   3) Kalimat 3 (k_2) = 8.38764005
   4) Kalimat 4 (k_3) = 3.49485002
   5) Kalimat 5 (k_4) = 2.09691001

Tabel 1. Contoh Perhitungan Bobot Kalimat

No	Kata	Tf	Idf	Tf-idf
1	signal	1	0.6989700043	0.6989700043
2	lot	1	0.6989700043	0.6989700043
3	Bgt	1	0.6989700043	0.6989700043
4	Min	1	0.6989700043	0.6989700043
Bobot atau *Weight (w)*				**2.79588002**

Berdasarkan data hasil pembobotan kalimat di atas, diambil nilai terkecil (kalimat 5 (k_4) = 2.09691001) dan nilai terbesar (kalimat 1 (k_0) = 9.08661006) untuk menentukan *centroid*.

3. Mengelompokkan setiap titik data ke *cluster* terdekat untuk menemukan *centorid* baru (C_i). Setelah menentukan titik pusat atau *centroid*, selanjutnya adalah mengelompokan data ke dalam masing-masing *cluster*. Setiap *cluster* akan men-dapatkan anggota yang ditentukan dari nilai bobot (w) kalimat, hasil penentuan akan menentukan anggota tersebut masuk ke dalam *cluster* C1 atau *cluster* C2. Masing-masing *cluster* akan memiliki anggota seperti berikut ini :
    1) C1 =k_1, k_3 , k_4
    2) C2 =k_0, k_2
4. Menghitung jarak masing-masing anggota *cluster*, untuk mendapatkan *centroid* baru dengan menggunakan persamaan 4.

$$C_{(i)} = \frac{(x_1 + x_2 + \cdots + x_n)}{\sum x} \quad (4)$$

Keterangan

$C_{(i)}$ = *centroid* baru

x_1, x_2, x_n= merupakan nilai pada masing-masing anggota *cluster*

$\sum x$ = jumlah anggota *cluster*

Mengulangi kembali langkah 3 dan 4 sampai pusat *cluster* dan anggota *cluster* tidak berubah atau tetap sama.

HASIL DAN PEMBAHASAN

Hasil perhitungan jarak masing-masing anggota *cluster*, untuk mendapatkan *centroid* baru dengan menggunakan persamaan 4 didapatkan nilai *cluster* C1 sebesar 2.79588002 dan *cluster* C2 sebesar 8.73712505.

Gambar 3 merupakan hasil Tabel *content* yang mengambil data dari tahap *stemming* untuk diproses kedalam perhi-tungan K-Means *Clustering* untuk menen-tukan kumpulan kalimat dari data yang diolah masuk ke *cluster* mana dengan menampilkan kumpulan data *testing* dengan menampilkan masing-masing kalimat tersebut memasuki *cluster* yang sesuai dengan hasil dari algoritma K-Means *Clustering.*

Implementasi Gambar 3 kemudian dibuat sebuah *wordcloud* untuk menampilkan sebaran kata dari data komentar sesuai masing-masing *clusternya* dengan menam-pilkan *wordcloud* dengan judul *cluster 0, cluster 1 dan cluster 2* seperti dapat dilihat pada Gambar 4.

Gambar 3. Implementasi K-Means Clustering

Gambar 4. Implementasi Tampilan Cluster K-Means

KESIMPULAN DAN SARAN

Implementasi metode K-Means *Clustering* menggunakan data komentar tweet @Indosatcare telah berhasil dibuat dan menghasilkan 3 buah *cluster* dengan anggota berbeda ditiap *clusternya.* Penelitian ini berhasil menampilkan kelompok dari anggota masing-masing *cluster* yang berbentuk *word-cloud* ke dalam 3 buah *wordcloud* berbeda, pada *wordcloud cluster* 0 anggotanya berbicara tentang jaringan Indosat yang parah, pada *wordcloud cluster* 1 anggotanya berbicara tentang permintaan perbaikan jaringan sinyal Indosat, dan pada *wordcloud cluster* 2 anggotanya berbicara tentang jaringan sinyal parah Indosat pada daerah Bogor.

Hasil penelitian ini diharapkan dapat menjadi masukan untuk *provider* dalam melihat keluhan yang masuk dari para konsumen mereka sehingga pihak *provider* dapat meningkatkan pelayanannya.

Pengembangan lebih lanjut dapat dilakukan untuk menambahkan metode klasifikasi lainnya kedalam metode K-Means *Clustering*, dengan menggabungkan 2 metode yang berbeda untuk mendapatkan hasil yang berbeda.

DAFTAR PUSTAKA

[1] B. Pang dan L. Lee, "Opinion Mining and Sentiment Analysis", *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, hal. 1–135, 2008.

[2] A. Chaudhuri dan J. Chant, "Protein-interaction mapping in search of effective drug target", *BioEssays* 27(9): 958—969, 2005

[3] B. Liu, *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publisher, 2012

[4] C. D. Manning, P. Raghavan, dan H. Schutze, *An introduction to information retrieval*. Cambridge: Cambridge University Press, 2008.

[5] Y. Y. Luhulima, "Sentimen Analysis Pada Review Barang Berbahasa Indonesia dengan Metode K-Nearest Neighbor", Skripsi Sarjana, Universitas Brawijaya, Malang, Indonesia, 2013.

[6] B. Nugroho, *Membuat Website Sendiri Dengan PHP & MySql*. Jakarta: *MediaKita*, 2009.

[7] O. Somantri, S. Wiyono dan Dairoh, "Metode *K-Means* untuk optimasi klasifikasi tema tugas akhir mahasiswa menggunakan *Support Vector Machine* (SVM)", *Scientific Journal of Informatics*, vol. 3, no. 1, 2016.

[8] S. Budi, "*Text mining* untuk analisis sentimen review film menggunakan algoritma *K-Means*", TehcnoCOM, vol. 16, no. 1, 2017

[9] N. D. Mentari, M. A. Fauzi, dan L. Muflikhah, L, "Analisis sentimen kurikulum 2013 pada sosial media Twitter menggunakan metode *K-Nearest Neighbor* dan *Feature Selection Query Expansion Ranking*", *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer (J-PTIIK)*, vol. 2, no. 8, hal. 2739–2743, 2018.

[10] A. Salam, J. Zeniarja, dan R. S. U. Khasanah, "Analisis sentimen data komentar sosial media Facebook dengan *K-Nearest Neighbor* (Studi kasus pada akun jasa barang J&T Ekspress Indonesia", dalam Prosiding SINTAK, 2018, hal.480–486.