SNESTIK

Seminar Nasional Teknik Elektro, Sistem Informasi, dan Teknik Informatika

https://ejurnal.itats.ac.id/snestik dan https://snestik.itats.ac.id

Informasi Pelaksanaan :
SNESTIK IV - Surabaya, 27 April 2024
Ruang Seminar Gedung A, Kampus Institut Teknologi Adhi Tama Surabaya

Informasi Artikel:
DOI : 10.31284/p.snestik.2024.5846

Prosiding ISSN 2775-5126

Fakultas Teknik Elektro dan Teknologi Informasi-Institut Teknologi Adhi Tama Surabaya
Gedung A-ITATS, Jl. Arief Rachman Hakim 100 Surabaya 60117 Telp. (031) 5945043
Email : snestik@itats.ac.id

Implementasi Algoritma Agglomerative pada Pengelompokkan Data Tweet

Eka Yoga Kartika Aji, S. Nurmuslimah, Rani Rotul Muhima

Institut Teknologi Adhi Tama Surabaya

e-mail: eka.yoga.kartika@gmail.com

ABSTRACT

Twitter is a microblogging social media platform that is still widely used today. It is still in demand for promotional media because it uses the concept of microblogging, which is quite effective as a means of writing. In recent years, many Twitter users have used it as a means of promotion to market their products. This study aims to group tweets based on the results of the data pre-processing stage and use word-weighted values to measure the words in each tweet. It also investigated the application of the agglomerative algorithm to form data groups based on the similarity of the word weight values using the TF-IDF method. The dataset was derived from 3 accounts, namely @TokopediaCare, @ShopeeCare, and @BlibliCare, totaling 1050 records. The agglomerative algorithm produced groups of tweets that were like a hierarchy. Each document could be connected to other documents that spread from the data distribution and then were grouped in a hierarchical shape, producing one final cluster group. After grouping the data, they were then validated using the Silhouette Coefficient method to measure whether the results of clustering were good or not. The agglomerative algorithm also allowed deeper identification of sub-topics that might exist in the Twitter dataset. The research results provide insight into how Twitter users interact in a conversation and the obstacles they experienced with the social media administrator of e-commerce. Social media, especially Twitter, must develop a communication strategy for customers.

Keywords*: Twitter, TF-IDF, Clustering, Agglomerative, Silhouette Coefficient*

ABSTRAK

Twitter merupakan platform media sosial *microblogging* yang masih banyak digunakan hingga saat ini. Twitter masih diminati untuk media promosi karena twitter menggunakan konsep micro blogging dimana cukup efektif sebagai sarana menulis. Beberapa Tahun terakhir pengguna Twitter banyak menggunakan sebagai sarana promosi untuk memasarkan produk yang dijualnya. Penelitian ini bertujuan untuk mengelompokkan tweet berdasarkan hasil dari tahap preprocessing data yang dilakukan dan nilai pembobotan kata untuk mengukur kata-kata yang ada di dalam setiap tweet, dan penerapan algoritma Agglomerative untuk membentuk kelompok data berdasarkan kemiripan nilai dari pembobotan kata menggunakan metode TF-IDF. Dataset yang digunakan didapatkan dari 3 akun yaitu @TokopediaCare, @ShopeeCare, @BlibliCare berjumlah 1050 data. Algoritma Agglomerative menghasilkan kelompok tweet berbentuk seperti hirarki di mana setiap dokumen dapat terhubung dengan dokumen lain yang menyebar. Dari sebaran data tersebut dikelompokkan dalam bentuk hirarki dan menghasilkan satu kelompok cluster akhir. Data yang telah dikelompokkan selanjutnya divalidasi menggunakan metode Silhouette Coefficient untuk mengukur hasil dari *clustering* apakah sudah baik atau belum. Algoritma Agglomerative juga memungkinkan identifikasi lebih mendalam terhadap sub-topik yang mungkin ada di dalam dataset Twitter. Hasil dari penelitian ini dapat memberikan gambaran tentang bagaimana percakapan pengguna Twitter dengan admin e-commerce di platform Twitter mengenai kendala yang dialami dan proses pengembangan strategi komunikasi terhadap pelanggan dalam lingkup media sosial khususnya Twitter.

Kata Kunci: Twitter, TF-IDF, Clustering, Agglomerative, Silhouette Coefficient

PENDAHULUAN

Media sosial menjadi sumber informasi yang sangat berguna untuk pegiat usaha dalam memasarkan kegiatan bisnis yang dilakoni dengan menampilkan produk, layanan, peringkat produk, ulasan produk, dan opini. Oleh karena itu, analisis pasar dari berbagai topik yang terkait dengan metode penambangan teks menggunakan data dari media sosial diharapkan dapat memantik inovasi, mengenali peluang baru, dan meningkatkan reputasi pegiat usaha. Seiring berkembangnya perilaku konsumen dalam membeli suatu produk, maka cara pemasaran yang digunakan juga harus semakin bervariasi [1]. Hal ini dimaksudkan untuk menjaring lebih banyak konsumen dari berbagai kalangan. Berangkat dari sinilah peran digital marketing bisa masuk menggantikan pemasaran. Saluran digital berusaha untuk mempengaruhi keputusan pembelian konsumen, sedangkan dari sisi konsumen ingin mencari produk yang sesuai dan memberikan penilaian berupa suka, kesaksian, pendapat, saran, peringkat, umpan balik, dan rekomendasi yang akan mempengaruhi performa penjualan dan menarik konsumen lain [2].

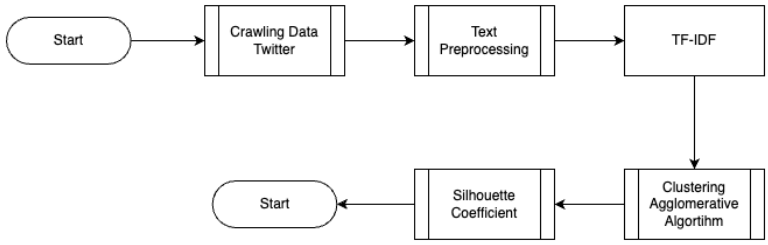
Data dari media sosial khususnya Twitter, dapat dihimpun dan dianalisis untuk mengetahui hal-hal apa saja yang berkaitan dengan ketertarikan konsumen.[3] Analisis klaster adalah salah satu tugas penambangan data paling dasar dalam analisis eksplorasi data dan pengaplikasiannya meluas ke pengambilan informasi. Sebagai alat dasar, clustering adalah teknik yang bertujuan untuk mengelompokkan kumpulan data (objek) ke dalam cluster, di mana objek yang serupa harus digabungkan ke dalam sebuah cluster dan cluster yang berbeda.

Para peneliti telah mengembangkan algoritma yang berbeda, termasuk berbagai kelompok data, grup data yang berbeda, algoritma yang sama, atau memilih dampak besar pada partisi pengelompokan akhir [4]. Agglomerative Clustering teknik clustering ini memiliki alur melakukan pemeriksaan ganda untuk meningkatkan tingkat akurasi secara keseluruhan, metode ini juga cocok untuk ekstraksi data seperti data media sosial karena mencari tingkat kesamaan antar objek [5]

Dari paparan di atas, tujuan dilakukannya penelitian ini tidak lain adalah untuk menerapkan algoritma Agglomerative pada dataset yang berisi teks sesuai dengan parameter yang telah ditentukan yang dapat menghasilkan analisa di mana hasil analisa tersebut dapat bermanfaat untuk penelitian selanjutnya.

METODE

Penelitian ini menggunakan beberapa alur metode guna mendapatkan hasil yang maksimal dan sesuai dengan langkah awal, yakni studi literatur berupa jurnal dan sumber ilmiah yang mendukung penelitian ini; melakukan observasi permasalahan dan mencari dataset yang digunakan untuk penelitian berupa data tweet dari media sosial Twitter; masuk ke tahap penelitian, yaitu tahap pre-processing data dan pengujian algoritma menggunakan dataset yang telah didapatkan. Dari data yang dihasilkan dapat ditarik kesimpulan secara keseluruhan.



Gambar 1. Flowchart system.

HASIL DAN PEMBAHASAN

Pembahasan Data I

Hal pertama yang dilakukan dalam penelitian ini adalah mengambil data dari twitter menggunakan API resmi yang disediakan oleh twitter, di mana API tersebut berisi *consumer key*, *consumer secret*, *access token*, dan *access token secret* yang digunakan untuk validasi agar penulis dapat mengakses data Twitter yang dibutuhkan. Setelah proses validasi berhasil, maka langkah selanjutnya adalah memasukkan kata kunci yang ingin didapatkan. Jika data tweet melalui kata kunci yang dimasukkan telah didapat, maka proses berikutnya adalah *library tweepy* untuk crawling data tweet di platform Twitter dan library csv untuk membuat file .csv. Data tweet yang berhasil terkumpul berjumlah 1050 dimana data ini masih harus melalui beberapa proses pengolahan data, seperti parsing data, case folding, tokenizing, filtering, normalisasi, stemming, dan pengolahan lainnya yang akan dijabarkan dalam pembahasan berikut. Sebelum masuk pada tahap case folding, terlebih dahulu diadakan parsing data. Parsing data merupakan proses pembagian kalimat menjadi satu dokumen di mana dalam proses ini dapat diketahui dokumen mana saja yang dibutuhkan untuk diolah dalam proses selanjutnya. Apabila data yang digunakan adalah suatu artikel, maka setiap paragraf akan dibagi menjadi satu dokumen, sedangkan jika data yang kita gunakan adalah ulasan, tweet, postingan di media sosial, maka setiap ulasan/tweet tersebut akan dianggap sebagai satu dokumen.

Pengolahan data pertama berupa tahap case folding, yakni proses perataan data tweet yang awal mulanya huruf kapital diubah menjadi huruf kecil semuanya. Hal ini dimaksudkan

untuk mempermudah proses pengolahan data. Jika proses case folding telah dilakukan, berikutnya masuk ke tahap tokenizing. Tokenizing dilakukan untuk menghilangkan karakter spesial, angka, dan tanda baca yang tidak relevan dan memecah setiap kata yang ada di dalam satu kalimat. Proses ini menggunakan library re, library string dan library nltk. Fungsi nltk ini digunakan untuk memanggil function word_tokenizer yang digunakan untuk membagi setiap kata di dalam kalimat dan digunakan juga untuk menghilangkan karakter spesial.

Tahap berikutnya adalah filtering. Filtering digunakan untuk menghilangkan kata-kata yang dianggap tidak relevan, misalnya kata hubung dan kata tidak relevan lainnya. Kata-kata tersebut dikumpulkan menjadi satu file yang disebut stopwords, di mana stopwords ini dapat membuat satu file berisi list yang berisi stopwords dalam Bahasa Indonesia. Setelah proses filtering selesai, maka selanjutnya masuk ke tahap stemming. Proses stemming sendiri bertujuan untuk mengembalikan kata yang berulang atau mencari poin dari setiap kata. Tahap stemming ini menggunakan library sastrawi dan library swifter, di mana keduanya memiliki fungsi tersendiri. Library sastrawi digunakan untuk proses stemming Bahasa Indonesia sedangkan library swifter digunakan untuk mempercepat proses stemming.

Tahap setelah proses stemming selesai adalah TF IDF. TF IDF merupakan tahap untuk mendapatkan nilai ekstraksi data agar didapatkan nilai dari masing-masing teks. Proses TF IDF ini menggunakan library scikit learn di mana di dalam library scikit learn terdapat module tfidfvectorizer yang digunakan untuk menghitung nilai ekstraksi TF-IDF. Proses TF IDF sendiri memiliki beberapa tahapan, yakni mulai proses load data set, proses deklarasi library dan membuat variable corpus, proses untuk melihat urutan kata, proses untuk melihat hasil dari TF IDF sendiri, proses melihat hasil dari TF IDF pada setiap kata, dan proses perhitungan TF IDF yang kemudian disimpan dalam bentuk file excel. Berikut gambar dan penjelasan singkat sebagai gambaran proses dalam penelitian ini:

Pembahasan Data II

Setelah proses di atas dan mendapatkan nilai ekstraksi pada setiap kata, maka selanjutnya data akan diolah menggunakan algoritma agglomerative. Tahap ini menggunakan bahasa pemrograman python 3 dan menggunakan jupyter notebook yang dibantu oleh library scikit-learn untuk proses clustering, sedangkan untuk visualisasi hasil clustering menggunakan library matplotlib. Setelah proses clustering selesai selanjutnya menghitung nilai rata rata dari setiap cluster proses ini dilakukan agar mengetahui hasil cluster yang dihasilkan baik atau buruk. Untuk proses ini menggunakan metode Silhouette Coefficient metode ini menghasilkan output -1 sampai 1 dimana semakin tinggi nilai yang didapatkan maka semakin bagus tetapi nilai ini dipengaruhi juga berapa banyak cluster yang dihitung.

KESIMPULAN

Dari penelitian yang telah dilakukan, berhasil didapatkan beberapa kesimpulan, yakni saat melakukan *Tokenizing* ada beberapa kata yang terlewat karena bahasa slang berkembang setiap hari dan berbeda di setiap daerah; dengan menggunakan 1050 dataset mendapatkan kesimpulan bahwa Algoritma Agglomerative tidak selalu dapat menentukan jumlah cluster di awal untuk perhitungan cluster, tetapi tidak menutup kemungkinan bisa juga menentukan jumlah cluster di awal; mendapatkan nilai rata-rata Silhouette Coefficient 0.564 dimana nilai tersebut dapat dikatakan baik karena masih dalam range 0.5.

DAFTAR PUSTAKA

[1] N. A. Morgan, K. A. Whitler, H. Feng, and S. Chari, "Research in marketing strategy," *Journal of the Academy of Marketing Science*, vol. 47, no. 1. Springer New York LLC, pp. 4–29, Jan. 15, 2019. doi: 10.1007/s11747-018-0598-1.

[2] J. Choi, J. Yoon, J. Chung, B. Y. Coh, and J. M. Lee, "Social media analytics and business intelligence research: A systematic review," *Inf Process Manag*, vol. 57, no. 6, Nov. 2020, doi: 10.1016/j.ipm.2020.102279.

[3] P. Harrigan, T. M. Daly, K. Coussement, J. A. Lee, G. N. Soutar, and U. Evers, "Identifying influencers on social media," *Int J Inf Manage*, vol. 56, Feb. 2021, doi: 10.1016/j.ijinfomgt.2020.102246.

[4] A. C. Benabdellah, A. Benghabrit, and I. Bouhaddou, "A survey of clustering algorithms for an industrial context," in *Procedia Computer Science*, Elsevier B.V., 2019, pp. 291–302. doi: 10.1016/j.procs.2019.01.022.

[5] M. Divyapushpalakshmi and V. Ramachandran, "Improved Overlapping Community Detection in Weighted Complex Social Network Using Hybrid Agglomerative Hierarchical Clustering for optical networks," 2022, doi: 10.21203/rs.3.rs-1537014/v1.