Laporan Tugas Besar Data Mining

MEMPREDIKSI ULASAN KUALITAS MAKANAN DENGAN MENGGUNAKAN K-MEANS CLUSTERING

M.Ilham Arief Himawan[1)], **Khoirunnisa**[2)], **Nadhea Allya**[3)],

Sandy Wijaya[4)], **Danar Zahra**[5)].

Program Studi Sains Data, Jurusan Sains, Institut Teknologi Sumatera

Email : milham.120450057@student.itera.ac.id [1)] khoirunnisa.120450029@student.itera.ac.id [2)]
nadhea.120450007@student.itera.ac.id [3)] sandy.120450047@student.itera.ac.id [4)]
danar.120450093@student.itera.ac.id [5)]

Abstrak

Media sosial memberikan kesempatan kepada konsumen untuk memberikan review ulasan atas produk berbentuk barang maupun makanan yang mereka beli dalam suatu platform belanja online.Tujuan pemberian ulasan ialah untuk memberikan nilai untuk kualitas produk yang diposting pada platform Amazon. Ulasan atau *review* sangat mempengaruhi penjual untuk bagaimana memproduksi barang jualan mereka. Penilaian pengguna sangat besar dan berpengaruh untuk rating platform Amazon.Ulasan pada data ini terdiri 9 kolom ulasan dengan total 110.000 dataset. Ulasan pelanggan tentang produk dan peringkat keseluruhan diberikan oleh mereka kemudian diterbitkan di halaman produk. Berdasarkan hasil pemodelan ulasan pengguna permasalahan ini diangkat untuk memprediksi ulasan kualitas makanan pada Amazon Fine Food menggunakan k-means. Ini akan meningkatkan pilihan ulasan bermanfaat Amazon di bagian atas bagian ulasan dan meningkatkan keputusan pembelian pelanggan. Ini juga dapat membantu pengulas lain sebagai panduan untuk menulis ulasan yang bermanfaat.

Kata Kunci :Amazon, K-Means,Pengguna

I. PENDAHULUAN

1.1 Latar Belakang

Dalam kehidupan bermasyarakat produk biasanya dibeli oleh pelanggan berdasarkan ulasan sebelumnya dari produk yang diberikan oleh pengguna melalui jejaring sosial seperti Facebook, Twitter, Instagram dan forum lainnya. Salah satunya Amazon, Amazon merupakan merupakan salah satu toko online terbesar di dunia yang terbentuk tahun 1994 dan sudah berhasil menguasai pasar online di dunia. Jaringannya sudah tersebar sangat luas di berbagai negara hingga ke seluruh dunia. Amazon didirikan oleh Jeff Bezos pada tahun 1994. Amazon memberikan rekomendasi tentang produk mereka dan sekitar 20% dari penjualan Amazon dipicu oleh ulasan pada sistem

rekomendasi. Namun, pengguna mungkin memerlukan sistem rekomendasi independen. Tantangan nyata dalam mengkategorikan ulasan berdasarkan sudut pandang pengguna adalah menganalisis sentimen yang disampaikan dalam ulasan.

Media sosial telah memberikan banyak kesempatan kepada konsumen dalam hal mengukur kualitas produk dengan membaca dan memeriksa ulasan yang diposting oleh pengguna platform belanja online. Selain itu, platform online seperti Amazon.com memberikan opsi kepada pengguna untuk memberi label ulasan sebagai 'Bermanfaat' jika mereka menganggap konten ulasan itu berharga. Ini membantu konsumen dan produsen untuk mengevaluasi preferensi umum secara efisien dengan berfokus terutama pada ulasan bermanfaat yang dipilih. Namun, ulasan yang baru-baru ini diposting mendapatkan suara yang relatif lebih sedikit dan ulasan dengan suara lebih tinggi masuk ke radar pengguna terlebih dahulu.

Studi ini menangani masalah ini dengan membangun sistem klasifikasi teks otomatis untuk memprediksi kegunaan ulasan online terlepas dari waktu postingnya. Studi ini dilakukan pada data yang dikumpulkan dari Amazon.com yang terdiri dari review makanan enak. Fokus penelitian sebelumnya sebagian besar tetap menemukan korelasi antara ukuran manfaat ulasan dan fitur berbasis konten ulasan. Selain menemukan fitur berbasis konten yang signifikan, penelitian ini menggunakan tiga pendekatan berbeda untuk memprediksi manfaat ulasan yang mencakup fitur vektor, fitur sentris ulasan dan ringkasan, dan fitur berbasis penyisipan kata. Apalagi konvensional pengklasifikasi yang digunakan untuk klasifikasi teks seperti mesin vektor dukungan.

Karena jumlah ulasan yang diberikan oleh pengguna sangat besar dan mereka hanya memiliki beberapa kalimat yang berisi pendapat tentang produk menjadi sulit bagi pelanggan potensial untuk membaca mereka untuk membuat keputusan yang tepat tentang pembelian hasil. Studi lain menyarankan untuk menambah fitur produk yang dikomentari oleh pelanggan dalam ulasan mereka dan kemudian mengidentifikasi pendapat setiap ulasan.

1.2 Permasalahan

Permasalahan yang diangkat pada tugas besar ini adalah memprediksi review kualitas makanan oleh pengguna platform Amazon. Sehingga bisa menyediakan kerangka kerja[1] untuk meningkatkan klasifikasi ulasan pelanggan pada produk.

1.3 Data Set

Amazon adalah salah satu iklan paling populer situs jaringan di seluruh dunia. Ia menjual buku, musik, elektronik, barang-barang rumah tangga dan barang-barang kebutuhan manusia lainnya. Data amazon ini bersumber dari situs kaggle,didalam ulasan ini terdiri 9 kolom ulasan dengan total 110.000 dataset. Ulasan pelanggan tentang produk dan peringkat keseluruhan diberikan oleh mereka diterbitkan di

1 Kerangka kerja yang digunakan untuk mengembangkan website

halaman produk. Kami melihat peringkat yang diberikan oleh pengulas dan

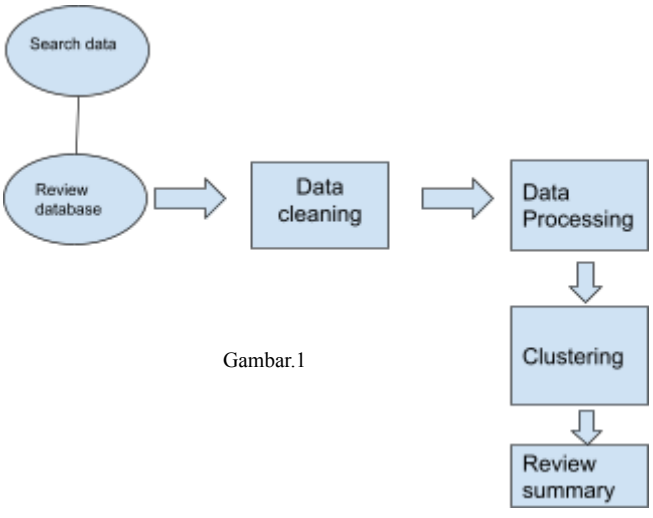
menghubungkan kata-kata dari ulasan kemudian meninjau dan mengelompokkannya menurut 5 kategori berbeda. Pengguna dalam himpunan data ini memiliki hubungan sosial berbasis pada produk yang mereka beli di situs. Kemudian mengumpulkan ulasan pengguna aktif di Amazon yang memberi lebih dari lima ulasan untuk buku yang berbeda dan juga memeriksa asal usul pengguna. Kami kemudian meneliti lebih lanjut pengguna ini untuk membangun sub-jaringan Amazon dari segi wilayah.

II. METODE

Pada penelitian ini algoritma yang digunakan dalam menyelesaikan permasalahan adalah K-Means. Teknik ini menciptakan kelompok objek target berdasarkan informasi yang ditemukan dalam sebuah data yg digunakan dimana di dalam data tersebut dilakukan sebuah pengkodingan yang membedakan objek dan hubungan di antara mereka. Kondisi dalam cluster tertentu mirip satu sama lain dan berbeda dari kondisi di cluster yang lainnya termasuk pada k-means. K-Means merupakan sebuah algoritma yang ada di dalam data mining yang dapat dipergunakan untuk melakukan pengelompokan atau clustering dari suatu data.K-Means juga merupakan algoritma clustering dengan metode partisi untuk mengelompokkan data berdasarkan variabel atau feature.

K Means adalah algoritma pembelajaran mesin yang paling umum dan paling sederhana dan mengikuti pendekatan berulang yang mencoba mempartisi himpunan data menjadi jumlah "K" yang berbeda subkelompok yang telah ditentukan sebelumnya dan tidak tumpang tindih di mana setiap titik data hanya dimiliki oleh satu sub kelompok menurut kualitas.Metode K-Means itu sendiri merupakan metode yang termasuk dalam algoritma clustering berbasis jarak yang membagi data ke dalam sejumlah cluster dan algoritma ini hanya bekerja pada atribut numerik.

Arsitektur sistem yang dibuat ditunjukkan dalam gambar 1. Mencari dan menyimpan database berupa csv, melakukan pembersihan dengan mengubah review Numerik menjadi ulasan kategoris tentang kondisi di atas 3 adalah positive dan di bawah 3 negatif karena peringkat ulasan dengan 3 tidak banyak berguna. Selanjutnya text processing dengan menemukan kalimat yang mengandung html. Selanjutnya Clustering dan visualisasi agar memudahkan menarik kesimpulan dari hasil yang sudah didapatkan.

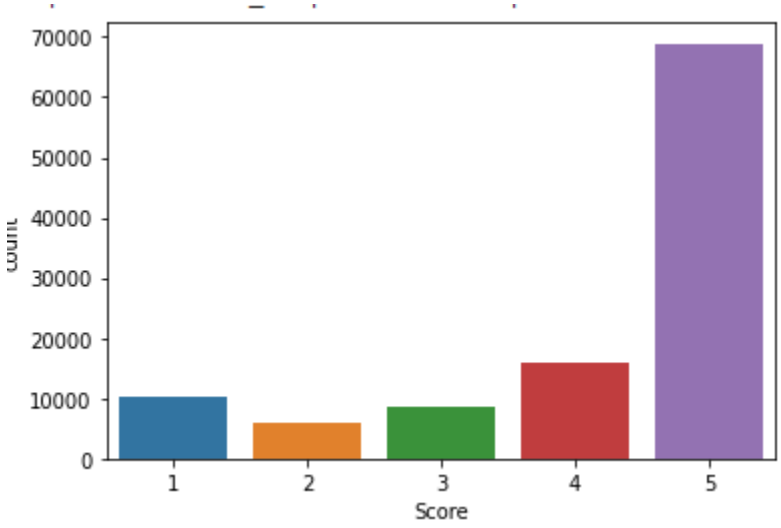


Gambar.1

III. HASIL DAN PEMBAHASAN

Dalam permasalahan ini kami melakukan pengolahan dan pengambilan data amazon melalui website Kaggle, dimana datasetnya diambil sebagai data base ulasan untuk percobaan permasalahan kami ini. Dataset ini awalnya terdiri dari 500 ribu data, tetapi sulit untuk menentukan kualitas review baik pada makanannya. Maka dari itu dataset ini kami ekstraksi menjadi 110 ribu data set yang terdiri dari 9 kolom ulasan lengkap.

Kolom HelpfulnessNumerator adalah jumlah pengguna yang menganggap ulasan ini bermanfaat terus-menerus. Kolom HelpfulnessDenominator adalah jumlah pengguna yang menunjukkan apakah mereka menganggap ulasan tersebut bermanfaat atau tidak bermanfaat secara terus-menerus. Skor adalah peringkat antara 1 dan 5. Sedangkan Teks adalah teks mengenai ulasan. Berikut adalah grafik untuk melihat score peringkat pelanggan didistribusikan.



Gambar.2. Graifik score

Dilanjutkan dengan melakukan cleaning data, dimana ini bertujuan untuk memastikan kebenaran dari data set yang didapatkan. Pada cleaning data kita akan mengubah ulasan numerik menjadi ulasan kategorik tentang kondisi diatas 3 adalah positif dan dibawah 3 merupakan negatif, karena peringkat ulasan dengan nilai 3 tidak banyak berguna dengan mencari tahu seberapa banyak ulasan dari pengguna yang bernilai positif dan negatif berdasarkan hasil dari skor yang didapat. Didapatkan output dari skor data yaitu 93644 ulasan data bernilai positif, dan untuk ulasan yang bernilai negatif yaitu 16356. Jadi dapat disimpulkan dari para pengguna yang memberikan ulasan sebagian besar memberikan ulasan positif .

Mengesampingkan fitur dengan data teks, kami sekarang menemukan korelasi antara fitur lain yang berbeda dalam kumpulan data dan pengaruhnya terhadap skor

ulasan diberikan sebagai masukan preprocessing dimana nantinya akan dilakukan penghapusan kata dan stemming.Pada dataset ini digunakan null stemming untuk stemming kata-kata dan diproses tanpa penghapusan kata berhenti. Pada permasalahan ini sebagian besar kata adalah jenis kata sifat karena ulasan sebagian besar menggambarkan konten produk.

Sehubungan dengan makanan,pengulas lebih banyak mengomentari makanan pada platform Amazon tersebut dengan kualitas dengan cara menggambarkan subjek dan lebih terorganisir. Stemming pada dataset ini untuk memahami makna dasar komentar reviewnya tetapi tidak mengevaluasi ulasan pengguna, kemudian lakukan ekstraksi kata yang sering muncul di komentar pengulas. Disini jarak matriks Euclidean digunakan untuk mengevaluasi kesamaan antara kata-kata. Jika dibandingkan dengan tf-idf, Euclidean dapat menemukan kalimat serupa di keseluruhan komentar ulasan. Secara umum persamaan yang digunakan untuk mencari jarak antara dua vektor dalam ruangan n-dimensi adalah D.

IV. KESIMPULAN

V. Referensi

Sulthana, A. Razia, and Ramasamy Subburaj. “An Improvised Ontology based K-Means Clustering Approach for Classification of Customer Reviews.” 2016.

VI. Lampiran

sumber:
https://www.kaggle.com/code/ameeamin/using-k-means-clustering-to-predict-helpfulness/notebook

colab:
https://colab.research.google.com/drive/1Oiu4CAT_oGvnxXPmt9cJDY0hJu8exedj#scrollTo=diXzli12CmB_