LAPORAN PROYEK AKHIR
PRAKTIKUM DATA SCIENCE

Analisis Sentimen Untuk Menentukan Kata-Kata Paling Negatif dan Positif Dari Review Hotel

ANNAS ADHARUQUDNI 123190014
FRISKA EKA KHOIRUNISA 123190031

PROGRAM STUDI INFORMATIKA
JURUSAN TEKNIK INFORMATIKA
FAKULTAS TEKNIK INDUSTRI
UNIVERSITAS PEMBANGUNAN NASIONAL "VETERAN"
YOGYAKARTA

2021

1. PENDAHULUAN

Salah satu lanyanan online pariwisata yang banyak digunakan masyarakat dunia saat ini adalah TripAdvisor. TripAdvisor adalah salah satu situs wisata terbesar di dunia yang membantu wisatawan mengoptimalkan potensi setiap perjalanan. TripAdvisor menawarkan sarana dari jutaan wisatawan serta berbagai pilihan dan fitur perencanaan wisata dengan link praktis ke alat bantu pemesanan yang memeriksa ratusan situs web untuk menemukan harga hotel terbaik. Situs web TripAdvisor merupakan komunitas wisata terbesar di dunia, menjangkau rata-rata 390 juta pengunjung untuk setiap bulannya, serta menampilkan 435 juta ulasan dan opini tentang 6,8 juta akomodasi, restoran, dan objek wisata yang beroperasi di 49 pasar di seluruh dunia (TripAdvisor, 2016).

Situs ini juga menyediakan kolom komentar dari pengguna yang mengulas tentang objek wisata maupun kuliner yang telah dikunjungi dan dapat dibaca oleh wisatawan lain. Data komentar yang ada di situs web tersebut dapat dimanfaatkan bagi pengelola hotel, berdasarkan hasil analisis sentiment yang kelas sentiment positif, negatif, atau netral. Hasil analisi opini itu selanjutnya dapat digunakan sebagai pertimbangan wisatawan dalam penentuan keputusan memilih hotel yang akan dikunjungi.

Untuk membuat sebuah analisis sentimen banyak hal yang harus dipersiapkan terlebih dahulu, salah satunya dengan memilih classifier yang akan digunakan. Clasifier metode yang dapat melakukan klasifikasi data menjadi beberapa kelas. Dalam penelitian ini, classifier yang dipilih adalah K-means, topic modelling, dan wordcloud.

2. METODE

Kami melakukan pengumpulan data yang didapat dari website *kaggle* untuk melihat komentar pelanggan di Tripadvisor. Pengumpulan data ini terbagi menjadi dua yaitu, observasi dan teknik pengolahan data. Observasi digunakan untuk menentukan objek yang nantinya ingin diteliti yatu, kata komentar pelanggan yang positif maupun yang negatif. Teknik pengolahan data yaitu mengambil komentar pelanggan yang memuat kata positif dan negative dari aplikasi Tripadvisor dengan total data 20.491komentar.

Metodologi yang digunakan untuk melakukan analisis sentimen menentukan komentar positif dan negative dari aplikasi tripadvisor menggunakan metode k-*means Clustering*. K-*means* adalah metode pengklasteran secara partitioning yang memisahkan data ke dalam kelompok yang berbeda. K-means mampu meminimalkan rata-rata jarak setiap data ke klasternya.

Pada penelitian ini, algoritma k-means clustering memiliki beberapa tahapan seperti berikut:

1. Menentukan k sebagai jumlah cluster yang ingin dibentuk.
2. Membangkitkan nilai random untuk pusat cluster awal (centroid) sebanyak k.
3. Menghitung jarak setiap data input terhadap masing-masing centroid menggunakan rumus jarak Euclidean (Euclidean Distance) hingga ditemukan jarak yang paling dekat dari setiap data dengan centroid. Berikut adalah persamaan Euclidian Distance:

$$d(xi,\mu j)=\sqrt{\sum}(xi-\mu j)2$$

   Keterangan:

   xi : data kriteria

   μj : centroid pada cluster ke-j
4. Mengklasifikasikan setiap data berdasarkan kedekatannya dengan centroid (jarak terkecil).
5. Memperbaharui nilai centroid. Nilai centroid baru diperoleh dari rata-rata cluster yang bersangkutan dengan menggunakan rumus:

$$\mu j(t+1)= 1/Nsj\sum xjj\in Sj$$

Keterangan:

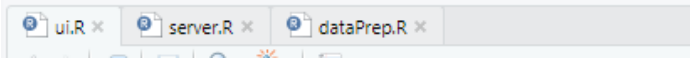
$\mu j(t+1)$: centroid baru pada iterasi ke (t+1)

Nsj : banyak data pada cluster Sj.

6. Melakukan perulangan dari langkah 2 hingga 5, sampai anggota tiap cluster tidak ada yang berubah. Jika langkah 6 telah terpenuhi, maka nilai pusat cluster (μj) pada iterasi terakhir akan digunakan sebagai parameter untuk menentukan klasifikasi data.

Di proyek ini kami menggunakan library shiny, shinycssloaders, shinydashboard, tidyverse, ggtech, tidytext, widyr, stm, e1071, caret, syuzhet, tm, dplyr, wordcloud, cluster, factoexra.

Pembuataannya menggunakan 3 file .R yaitu untuk tampilan interface, untuk persialan data, dan server (**Gambar 2.2**). File ui.R berisi codingan untuk antarmuka di webite, file servere.R digunakan untuk mengirim data ke ui.R, dan dataPrep.R digunakan untuk persiapan data yang akan diolah.

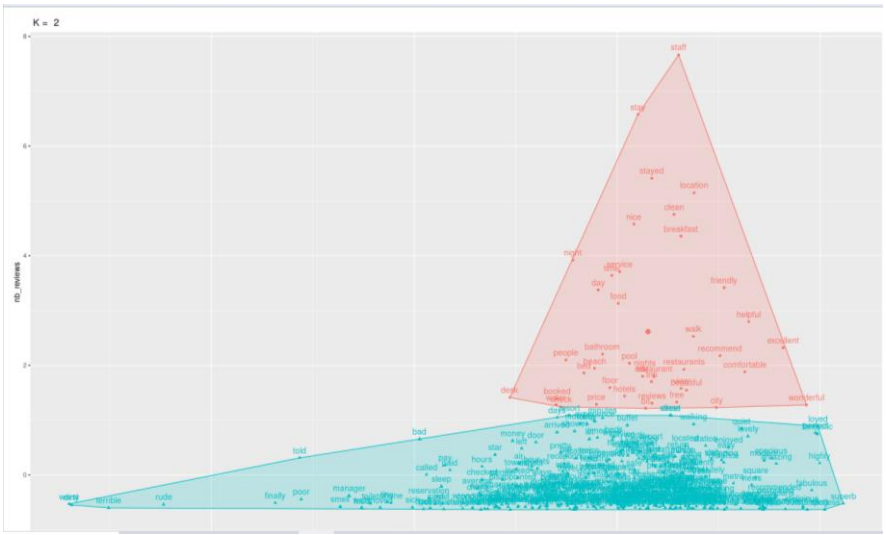


Gambar 2.2 Persiapan File

```
output$kmeans <- renderPlot({
    k2 <- kmeans(indexed, centers =
    input$sizeKmeans)
    fviz_cluster(k2, data = indexed) +
    ggtitle(paste(" K = ",input$sizeKmeans))
   })
```

Tabel 2.1 Menentukan k-Mean

Berdasarkan **Tabel 2.1** nilai k yang direkomendasikan adalah 2 digunakan ntuk melihat pengelompokkan komentar yang berada di aplikasi Tripadvisor. Dengan hasil seperti **Gambar 2.3.**

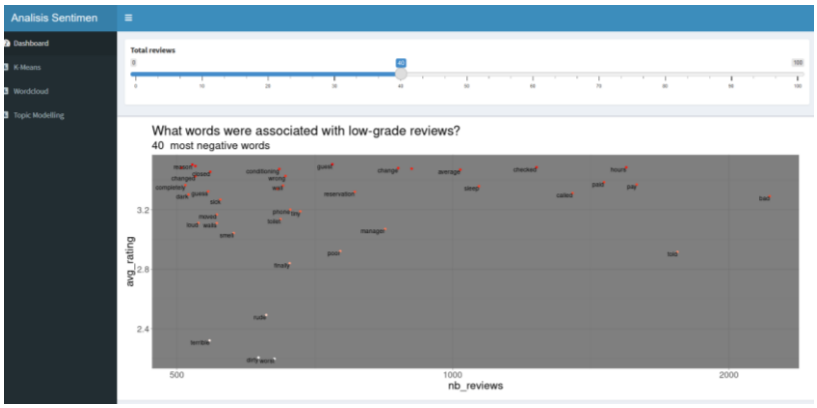


Gambar 2.3 Rekomendasi k-NN

Setelah melakukan pengclusteran dengan k-Mean data akan di visualisasi dengan menggunakan wordcloud, yang akan menghasilkan kata yang paling sering muncul dalam komentar di Tripadvisor.

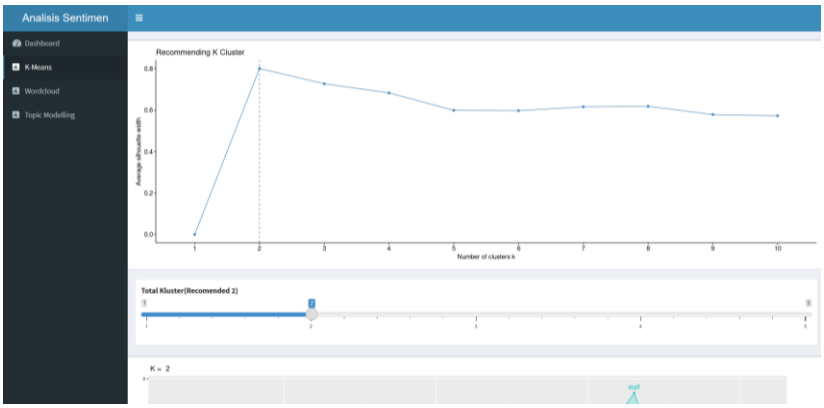
Gambar 2.4 Visualisasi Wordcloud

3. HASIL DAN PEMBAHASAN

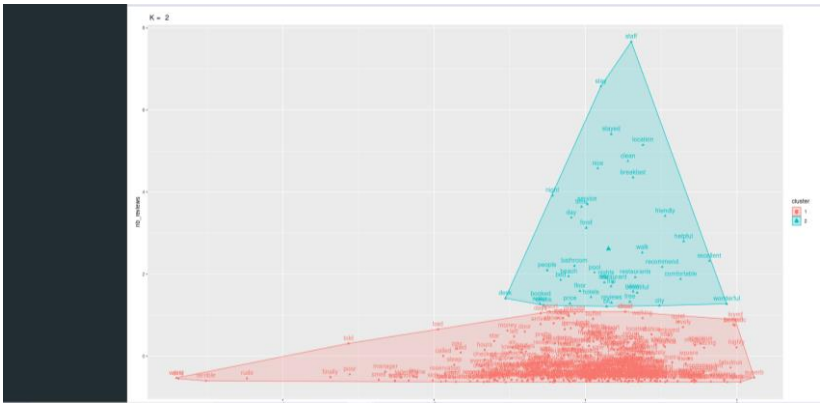


Gambar 3.1 Tampilan Awal

Tampilan awal halaman ketika program dijalankan terdapat tiga fitur disebelah kiri yang digunakan untuk mendapatkan informasi dari komentar yang berada di tripadvisor yaitu k-Mean, wordcloud, dan topic modelling.

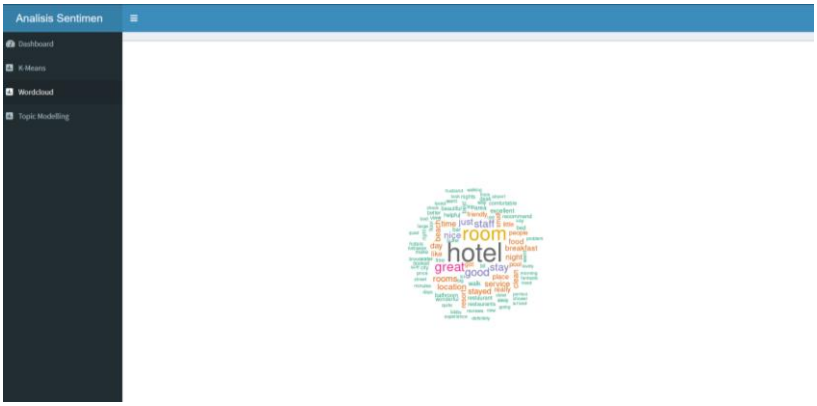


Gambar 3.2 Metode k-Mean



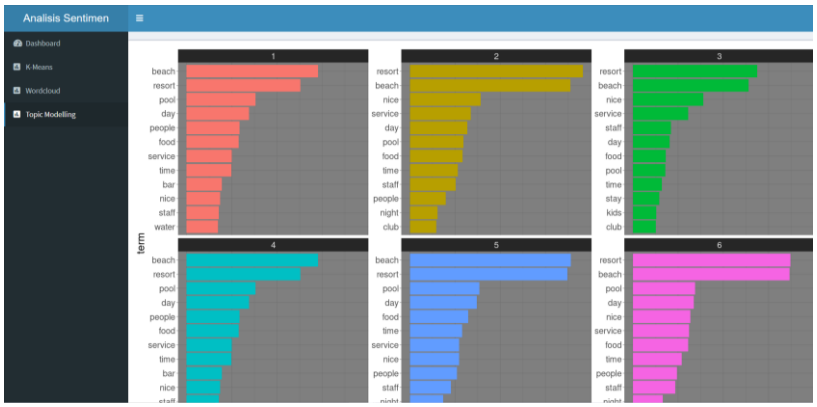
Gambar 3.3 Metode k-Mean

Tampilan selanjunya, pengclusteran kata dengan k-Mean dapat dilihat di **Gambar 3.2** terdapat grafik yang digunakan untuk merekomentasi k-Mean yang digunakan yaitu, 2. Kemudian, di **Gambar 3.3** terdapat pengelompokkan dengan k-Mean.



Gambar 3.4 Worldcloud

Tampilan selanjutnya, memunculkan huruf yang sering keluar di kolomkomentar dengan menggunakan wordcloud. Terdapat kata hotel, staff, good, great, stay, dll.



Gambar 3.5 Topic Modelling

Tampilan terakhir dari adalah topic modeling yaitu Topic modeling atau pemodelan topik merupakan metode clustering yang termasuk dalam unsupervised learning. Dalam unsupervised learning tidak ada label untuk suatu objek.

4. KESIMPULAN

Penelitian ini melakukan pengklasifikasi teks dari komentar trivadsitor dengan pengklasifikasi K-Nearest Neighbor, menggunakan 20.491komentar. Serta kata yang sering muncul dalam komentar adalah sick, moved, loud, walls, smell, toilet, dll.