

LAPORAN PROYEK AKHIR
PRAKTIKUM DATA SCIENCE
SENTIMEN ANALISIS MENGENAI ULASAN HOTEL DI TWITTER



RIVANKA DESYA FAWWAZ A 123190018

AMELIA PUTRI CHRISTANTI 123190032

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

2021

1. PENDAHULUAN

Pada era teknologi saat ini pertumbuhan teknologi semakin canggih dan modern, pertumbuhan pesat juga terjadi pada internet dan media social yang saat ini menyediakan berbagai informasi opini orang lain. Dalam dunia bisnis, khususnya pada bidang perhotelan ulasan yang positif dari para konsumen akan memberikan gambaran yang bagus untuk para konsumen lainnya. Ulasan yang positif dapat mempertahankan pelayanan pada hotel tersebut, namun jika ulasan tersebut berupa hal yang negatif, maka bisa menjadi bahan evaluasi bagi pihak hotel.

Untuk membantu sebuah perusahaan hotel dalam melihat bagaimana penilaian masyarakat terhadap hotelnya, perlu adanya sistem yang dapat merangkum bagaimana arah ulasan tersebut, yaitu dengan analisis sentimen. Analisis Sentimen adalah sebuah cara yang dilakukan untuk mendapatkan Opinion atau pendapat public. Dengan Analisis Sentimen kita akan mendapatkan feedback yang cepat dari masyarakat dengan memanfaatkan komentar dari masyarakat yang berkunjung sehingga pihak hotel dapat melakukan tindakan cepat dalam laporan konsumen.

Dengan banyaknya konsumen yang sering datang menginap di hotel cukup menyulitkan pihak hotel dalam memantau opinion konsumen, dengan adanya sistem yang menggunakan metode Naïve Bayes ini diharapkan mempermudah pihak hotel dalam mendapatkan informasi dan cepat dalam mengambil tindakan.

2. METODE

Data yang digunakan didapat dari tweet masyarakat yang berada di social media Twitter. Dengan menggunakan aplikasi Rstudio, kami membuat akun twitter developer untuk mendapatkan twitter API yang akan dihubungkan kedalam aplikasi Rstudio dan memanfaatkan library (rtweet) sehingga memungkinkan untuk mendapatkan data dari twitter. Dan dibawah ini bebrapa library yang digunakan untuk mengambil data dan melakukan proses sentimen analisis.

```
library(rtweet)
library(dplyr)
library(tidyr)
library(tidytext)
library(ggplot2)
library(textdata)
library(purrr)
library(twitter)
library(wordcloud)
library(shiny)

## ambil API dari developer Twitter
api_key <- "nTAKExJAWaREXLXvA3dRaqIPm"
api_secret_key <- "xQKLmpMjiKuFwLVIOxOYzQd89FzLi5yBC4zrog26gF1jxxaJwt"

## authenticate via web browser
token <- create_token(
  app = "project_rivanka_lia",
  consumer_key = api_key,
  consumer_secret = api_secret_key)
```

Metode yang digunakan dalam project ini adalah Naïve Bayes. Metode Naïve Bayes adalah metode klasifikasi dalam penambangan teks yang digunakan dalam analisis sentimen. Metode ini berpotensi baik dalam klasifikasi dalam hal presisi dan komputasi data. Naïve Bayes banyak digunakan dalam teknik klasifikasi terutama Twitter menggunakan beberapa metode seperti Unigram Naïve Bayes, Multinomial Naïve Bayes, dan Maximum Entropy Classification. Fitur utama klasifikasi Naïve Bayes untuk mendapatkan hipotesis yang kuat dari setiap kondisi atau peristiwa. Perhitungan kategori probabilitas dalam Naïve Bayes menggunakan pendekatan algoritma Bayes menggunakan persamaan:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \quad (1)$$

Persamaan (1) menunjukkan Y suatu kelas spesifik, X merupakan data pada kelas yang belum diketahui, sedangkan $P(Y|X)$ adalah probabilitas hipotesis berdasarkan kondisi, sedangkan $P(Y)$ dan $P(X|Y)$ merupakan probabilitas sebelumnya dari kelas berdasarkan kondisi hipotesis, sedangkan $P(X)$ merupakan probabilitas Y. Dalam

klasifikasi Naïve Bayes, persamaan 1 kemudian dikembangkan lagi menjadi persamaan (2) berikut:

$$P(Y|X_1, X_2, \dots, X_n) = \frac{P(X_1, X_2, \dots, X_n|Y)P(Y)}{P(X_1, X_2, \dots, X_n)} = \frac{P(X_1|Y)P(X_2|Y) \dots P(X_n|Y)P(Y)}{P(X_1, X_2, \dots, X_n)}$$

Dimana persamaan diatas hasil hitung dari semua probabilitas posterior pada nilai X untuk semua nilai di Y, sehingga klasifikasi Naïve Bayes akan membuat prediksi berdasarkan probabilitas maksimum dari probabilitas posterior laplace yang ditunjukkan pada persamaan (3) berikut dimana c merupakan jumlah nilai pada Y[28]:

$$P(X_i|Y) = \frac{N_{ic}+1}{N_c+c}$$

```
# Naive bayes
library(e1071)
library(caret)
library(syuzhet)

#digunakan untuk membaca file csv yang sudah di cleaning data
hotel_dataset <- read.csv("cleaningdata\\clean_data.csv", stringsAsFactors = FALSE)

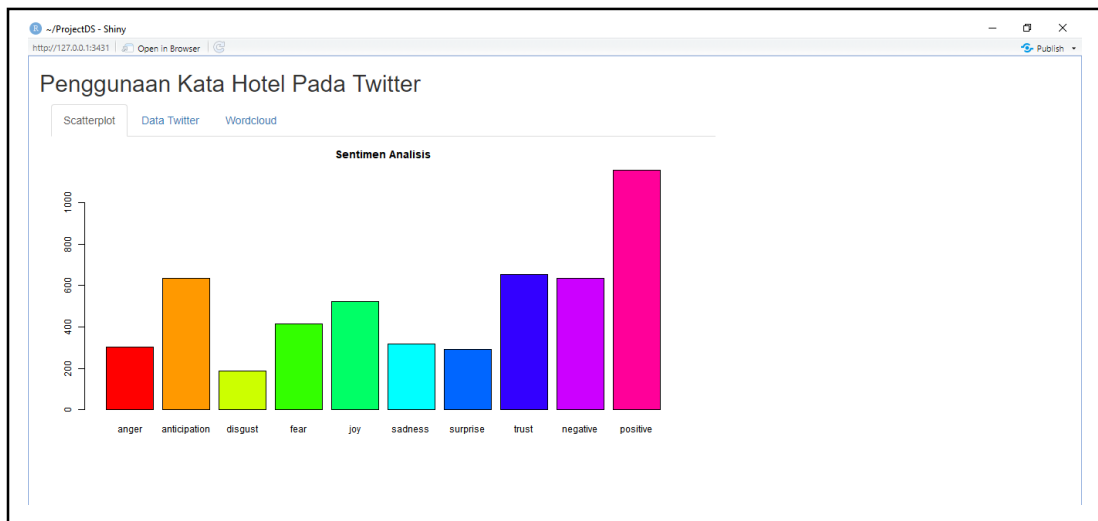
#digunakan untuk mengeset variabel cloumn text menjadi char
review <- as.character(hotel_dataset$stripped_text1)

#memanggil sentimen dictionary untuk menghitung presentasi dari beberapa emotion dan mengubahnya ke dalam text f
get_nrc_sentiment('happy')
get_nrc_sentiment('excitement')
s<-get_nrc_sentiment(review)
review_combine<-cbind(hotel_dataset$stripped_text1,s)
par(mar=rep(3,4))
barplot(colSums(s),col=rainbow(10),ylab='count',main='sentiment analisis hotel')

#20 kata teratas di tweet hotel
cleaned_tweets.Kata %>%
  count(word, sort = TRUE) %>%
  top_n(20) %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(x=word, y = n)) +
  geom_col() +
  coord_flip() +
  theme_classic() +
  labs(x="Count",
       y="Unique word",
       title="Unique word counts found in hotel Tweets")
```

Code pada R Studio yang menggunakan metode Naïve Bayes untuk memanggil sentimen dictionary yang akan dihitung presentasi dari beberapa emotion dan mengubahnya ke dalam text file. Selain itu metode ini juga digunakan untuk menampilkan 20 kata teratas pada tweet hotel.

3. HASIL DAN PEMBAHASAN



Gambar 3.1 Tampilan Scatterplot

Gambar diatas merupakan tampilan awal shiny saat di lakukan proses run. Pada tampilan awal terdapat Scatterplot yang merupakan grafik dari sentimen analisis hotel. pada grafik tersebut dapat dilihat bahwa terdapat kata yang mengungkapkan pendapat tentang hotel dengan kisaran angka seperti yang terlihat pada gambar.

The figure shows the "Data Twitter" tab of the Shiny application. It displays a table with 10 entries of tweet data. The table has columns for user_id, status_id, created_at, screen_name, text, source, display_text_width, reply_to_status_id, and reply_to_user_id. The first two entries are visible, showing tweets from users with IDs x1278888189143334912 and x757759857629941760.

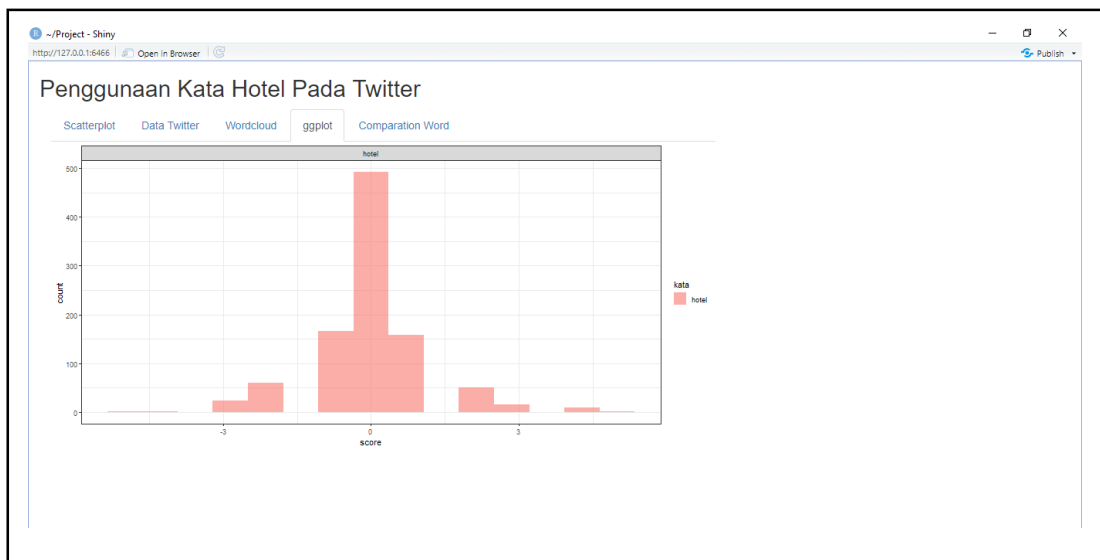
| | user_id | status_id | created_at | screen_name | text | source | display_text_width | reply_to_status_id | reply_to_user_id |
|---|----------------------|----------------------|---------------------|----------------|---|-----------------------|--------------------|----------------------|------------------|
| 1 | x1278888189143334912 | x1465766843436523523 | 2021-11-30 19:37:02 | sigh_lisalouse | <U+2729> <U+0001F50A> neutral mlk hotel ðŸŒ¸ where youâ€™ll find me now https://t.co/CqUAWwkyq https://t.co/zmeCarETSM | Twitter for iPhone | 76 | | |
| 2 | x757759857629941760 | x1465766831453470722 | 2021-11-30 19:36:59 | knicks_fan333 | @MBonvoyAssist Thanks! It is a Luxury Collection hotel, where your website mentions \$100 USD compensation if a welcome gift is not offered at check in. It would be great if you could follow through on the guarantee and offer the compensation! | Twitter Web App | 227 | x1465765839613960192 | x2541510835 |

Gambar 3.2 Tampilan Data Twitter

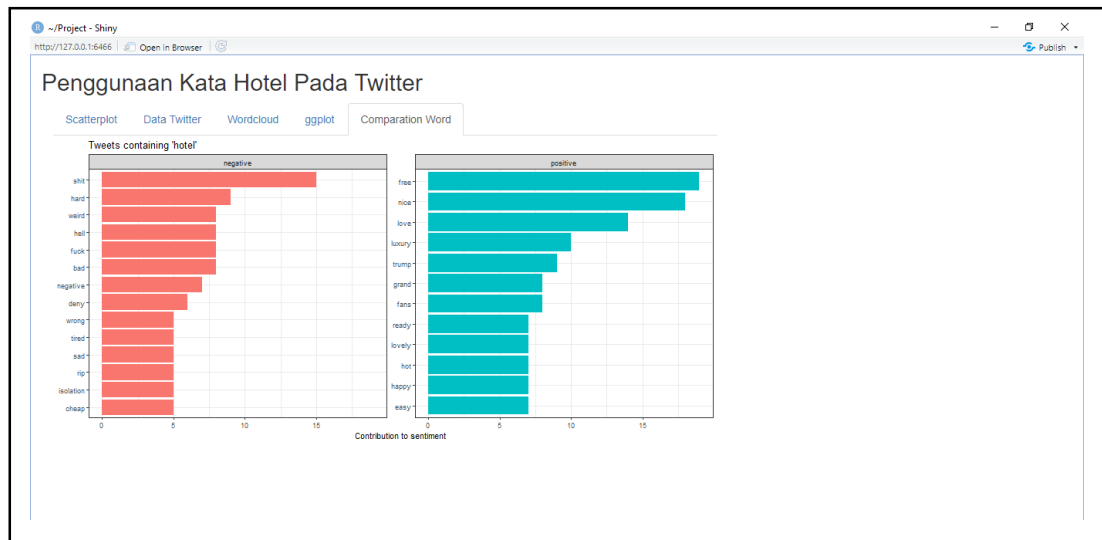
Gambar diatas merupakan data twitter yang telah diambil untuk dilakukan sentimen analisis. Pada gambar tersebut menunjukkan 10 data teratas yang terdapat pada dataset hotel.



Pada tampilan Wordcloud terdapat 125 kata yang sering muncul pada ulasan tentang hotel. Dan kata 'free' dan 'nice' menunjukkan kata yang sering digunakan untuk ulasan mengenai hotel. Pada gambar tersebut menunjukkan semakin kecil tulisan katanya maka kata tersebut jarang digunakan atau sedikit digunakan.



Tampilan ggplot menunjukkan kata yang menjadi ulasan tentang hotel di twitter paling banyak bersifat netral lalu diikuti kata yang bersifat negatif dan yang terakhir bersifat positif.



Gambar 3.5 Tampilan Comparison Word

Tampilan pada Comparison Word ini terdapat 2 tabel yaitu tabel negatif dan tabel positif. Masing-masing tabel terdapat kata yang menduduki peringkat terendah maupun tertinggi. Pada tabel negatif kata “shit” berada paling atas maksudnya kata tersebut sering digunakan untuk ulasan hotel di twitter. Sedangkan pada tabel positif kata “free” berada paling atas yang berarti kata tersebut sering digunakan.

4. KESIMPULAN

Proyek Akhir ini menerapkan algoritma Naïve Bayes untuk klasifikasi sentimen. Analisis sentimen dalam study ini menggunakan data Twitter dengan kata kunci hotel. Beberapa twit menunjukkan kearah positif dengan kata ‘free’ merupakan kata yang memiliki frekuensi tinggi dalam percakapan pada periode tersebut. Dengan pengolahan data ini diharapkan pengguna lebih mudah untuk mengerti isi data tersebut dan agar lebih efisien.

Link Github : https://github.com/AmeliaPutriChristanti/Project_Akhir

Link Video :

<https://drive.google.com/file/d/12tH9N-uQgZOMsUxqouRtDptG2gI6YyJR/view?usp=drivesdk>