LAPORAN PROYEK AKHIR
PRAKTIKUM DATA SCIENCE

ANALISIS SENTIMEN PADA REVIEW PENGUNJUNG UNIVERSAL STUDIO

SEKAR ARUM K 123190135
ADRIANUS WISNU P 123190148

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

2021

1. PENDAHULUAN

Universal Studio merupakan salah satu perusahan perfilm-an asal Amerika Serikat yang sudah bergerak cukup lama sejak 1992. Banyak film *box movies* seperti Jurassic World, Despicable Me, Fast and Furious, King Kong, dan masih banyak lagi. Mendapatkan keuntungan yang sangat besar, Universal Studio atau dalam pefilm-an lebih dikenal sebagai Universal Pictures, mengekspansi bisnisnya menuju ke ranah yang lebih baik yaitu wahana keluarga. Seperti yang kita tahu, wahana keluarga Universal Studio yang sangat terkenal pula dan sudah sangat mendunia layaknya Disney World dan Lego Land. Menangani sebuah bisnis yang terbilang cukup berbeda dari biasanya, maka Universal Studio memberikan wadah kepada para pengunjung untuk memberikan review demi dijadikan pedoman untuk terus berkembang dan berinovasi menjadi lebih baik.

Namun dari banyaknya pengunjung Universal Studio setiap tahunnya, termasuk juga cabang-cabang yang berada di berbagai belahan dunia, maka digunakan analisis sentimen untuk membantu mengelompokkan hasil review yang telah di dapat. Analisis sentimen adalah bidang penelitian yang digunakan untuk menganalisis pendapat pribadi, emosi, evaluasi, sikap, dan emosi yang berkaitan dengan topik, layanan, produk, individu, organisasi, atau peristiwa tertentu. Analisis sentimen dilakukan untuk mengetahui apakah pendapat atau komentar suatu masalah bersifat positif atau negatif dan dapat dijadikan acuan untuk meningkatkan pelayanan atau meningkatkan kualitas produk.

Tanggapan dapat bermakna sentimen atau mengekspresikan emosi yang dapat dibagi menjadi tiga kategori yaitu kategori positif, kategori netral, dan kategori negatif. Kategori positif mengandung kata-kata yang baik dan mendukung. Kategori netral mengandung dari kategori positif maupun negatif. Sedangkan kategori negatif mengandung kata-kata buruk.

Tujuan dari penelitian ini adalah mengetahui bagaimana review positif hingga negatif pada review yang telah didapatkan melalui dataset dalam bentuk barplot. Selain mengetahui bagaimana review pengunjung, penelitian ini juga bertujuan untuk mengetahui kata apa saja kata yang sering muncul di dalam review Universal Studio dalam bentuk wordcloud.

2. METODE

2.1 Pengambilan data

Proses pengambilan data melalui website kaggle yang berisi ulasan atau review pengunjung dari keseluruhan cabang Universal Studio. Dari website tersebut dapat ditemukan dataset berformat .csv yang akan digunakan dalam penelitian in. Di dalam data set termuat 6 kolom yaitu : Nama Pengulas, Peringkat, Tanggal Tertulis, Judul, Review_Text, dan Cabang. Berikut untuk link dari dataset yang dapat di download dari website kaggle.com :
https://www.kaggle.com/dwiknrd/reviewuniversalstudio

2.2 *Preprocessing*

Setelah mendapatkan data kemudian dilakukan *replacement* terhadap tanda (<, >, /, :, ;, http\\w+, *) atau *emoticon* yang tidak berguna dengan spasi dan melanjutkan dengan taha *preprocessing*. Tujuan dari tahap ini adalah untuk membersihkan data agar tidak menyulitkan proses analisis sentimen nantinya. Yang dibersihkan dari data dalam dataset review Universal Studio ini adalah :

a. *Remove Punctuation,* berfungsi untuk menghilangkan tanda baca yang masih tersisa saat *replacement*.
b. *Remove Numbers,* berfungsi untuk menghilangkan angka.
c. *Remove stripWhiteSpace,* berfungsi untuk menghilangkan ekstra spasi dari *replacement*.
d. *Remove stopWords,* berfungsi untuk menghilangkan kata umum yang tidak memiliki makna.
e. *Content Transformer,* berfungsi untuk mengubah semua huruf menjadi huruf kecil.

2.3 Naïve Bayes

Data frame yang telah dibersihkan akan dibaca dan ditambahkan sebuah variable baru. Variable tersebut befungsi untuk menampung nilai yang telah diberikan pada data review. Bila nilai data review kurang dari 0 maka review tersebut merupakan review negatif dan begitu juga sebaliknya. Tak lupa memberikan bobot pada tiap-tiap kata. Pembobotan dilakukan untuk mendapatkan nilai dari data review yang berhasil dibersihkan. Setelah membuat model Naïve Bayes untuk memprediksi data frame yang telah diberi nilai, yang terakhir menyimpan hasil prediksi ke dalam data frame.

2.4 Barplot

Barplot atau diagram batang adalah diagram yang berguna untuk menyajikan perbandingan data pada satu ata beberapa variabel data. Data pada grafik batang yang disajikan dalam bentuk persegi panjang horizontal, yang panjangnya sesuai nilai masing-masing. Dengan begitu kita bisa melihat dengan cepat dan mudah data mana yang memiliki krinerja atau nilai yang lebih tinggi pada review pengunjung Universaal Studio.

2.5 Wordcloud

Word cloud (disebut juga text cloud atau tag cloud) merupakan salah satu metode untuk menampilkan data teks secara visual. Grafik ini populer dalam text mining karena mudah dipahami. Dengan menggunakan word cloud, gambaran frekuensi kata-kata dapat ditampilkan dalam bentuk yang menarik namun tetap informatif. Semakin sering satu kata digunakan, maka semakin besar pula ukuran kata tersebut ditampilkan dalam word cloud.

2.6 Shiny

Shiny merupakan sebuah paket pada R Studio untuk membuat website. Shiny menggabungkan antara komputasi statistika R dan interaksinya dengan website modern. Shiny sendiri terdiri dari 4 bagian yaitu :

a. Global

berfungsi untuk mencantumkan library-library yang akan digunakan. Selain mencantumkan library, bagian ini juga bisa digunakan untuk memuat data yang akan digunakan, menentukan source file, proses autentikasi, dan semua pengaturan yang bersifat global untuk kepentingan penelitian ini.

b. User Interface

Digunakan untuk mendefinisikan tampilan web dari penelitian ini dan memuat seluruh fungsi input dan output. Pada bagian ini digunakan dashboardpage untuk halamanya yang berisi dashboardheader untuk bagian atas dari halaman (judul), dashboardsidebar untuk menu bagian kiri, dan dashboardbody untuk isi website (analisis sentimen, wordcloud)

c. Server

Merupakan fungsi yang mendefinisikan logika analysis dari sisi server pada penelitian ini.

d. ShinyApp

Merupakan fungsi untuk memanggil UI dan server yang telah dibuat untuk dijalankan.

3. HASIL DAN PEMBAHASAN

Hasil dan pembahasan pada analisis sentimen terhadap review pengunjung menghasilkan data cleansing, output barplot, output wordcloud, dan output aplikasi shiny.

3.1 Library pada Data Cleansing

Pada listing program 3.1 terdapat library yang dibutuhkan seperti llibrary tm untuk membersihkan data, library vrom untuk load dataset dan library here untuk menyimpan dataset.

```
```{r}
# Membersihkan data
library(tm)
# Load dataset
library(vroom)
# Menyimpan dataset
library(here)
```
```

Listing Program 3.1 Library pada Data Cleansing

3.2 Proses Data Cleansing

Pada proses ini dataset text universal_studi_branches.csv akan diproses dan dibersihkan. Tahap data cleansing terdiri dari membersihkan URL, membersihkan new line, membersihkan tanda koma, membersihkan tanda titik dua, membersihkan tanda titik koma, mebersihkan tanda titik tiga, dan lainnya. Proses utama pada data cleansing yaitu agar data bersih dari tanda baca maupun karakter lainnya. Selain itu digunakan juga stopword pada pembersihan data. Setelah proses data cleansing selesai, data baru yang sudah bersih akan disimpan dengan nama ulasan_clean.csv

```
```{r load dataset}
dataset <- vroom(here('universal_studio_branches.csv'))


ulasan <- dataset$review_text
ulasan2 <- Corpus(VectorSource(ulasan))
```
```

```
# Membersihkan URl
removeURL <- function(x) gsub("http[^[:space:]]*", "", x)
reviewclean <- tm_map(ulasan2, removeURL)

# Membersihkan New Line
removeNL <- function(y) gsub("\n", " ", y)
reviewclean <- tm_map(reviewclean, removeNL)

# Membersihkan tanda koma
replacecomma <- function(y) gsub(",", "", y)
reviewclean <- tm_map(reviewclean, replacecomma)

# Membersihkan tanda titik dua
removetitik2 <- function(y) gsub(":", "", y)
reviewclean <- tm_map(reviewclean, removetitik2)

# Membersihkan tanda titik koma
removetitikkoma <- function(y) gsub(";", " ", y)
reviewclean <- tm_map(reviewclean, removetitikkoma)

# Membersihkan tanda titik tiga
removetitik3 <- function(y) gsub("p…", "", y)
reviewclean <- tm_map(reviewclean, removetitik3)

# Membersihkan amp
removeamp <- function(y) gsub("&amp;", "", y)
reviewclean <- tm_map(reviewclean, removeamp)

# Membersihkan karakter
removeUN <- function(z) gsub("@\\w+", "", z)
reviewclean <- tm_map(reviewclean, removeUN)
remove.all <- function(xy) gsub("[^[:alpha:][:space:]]*", "", xy)

# Membersihkan tanda baca
```

```
reviewclean <- tm_map(reviewclean,remove.all)

reviewclean <- tm_map(reviewclean, removePunctuation)

reviewclean <- tm_map(reviewclean, tolower)


myStopwords = readLines("stopwords_en.txt")

reviewclean <- tm_map(reviewclean,removeWords,myStopwords)

dataframe<-data.frame(text=unlist(sapply(reviewclean,          `[`)),
stringsAsFactors=F)

View(dataframe)


# Menyimpan data review_text yang sudah dibersihkan

write.csv(dataframe,file = 'ulasan_clean.csv')

```
```

Listing Program 3.2 Proses Data Cleansing

Pada proses menampilkan barplot dari data yang sudah melakukan cleansing diperlukan library e1071 untuk perhitungan naive bayes, library caret untuk klasifikasi data dan library syuzhet untuk get_nrc. Data ulasan_clean.csv akan diubah menjadi char dan melakukan klasifikasi pada analisis sentimen.

3.3 Output Barplot

```
```{r Barplot}

# Library naive bayes

library(e1071)

# Library klasifikasi data

library(caret)

# Library untuk fungsi get_nrc

library(syuzhet)


datapariwisata  <-  read.csv("ulasan_clean.csv",stringsAsFactors  =
FALSE)

# Mengubah text menjadi char

review <- as.character(datapariwisata$text)

s <- get_nrc_sentiment(review)


# Melakukan klasifikasi data
```

```
review_combine <- cbind(datapariwisata$text,s)

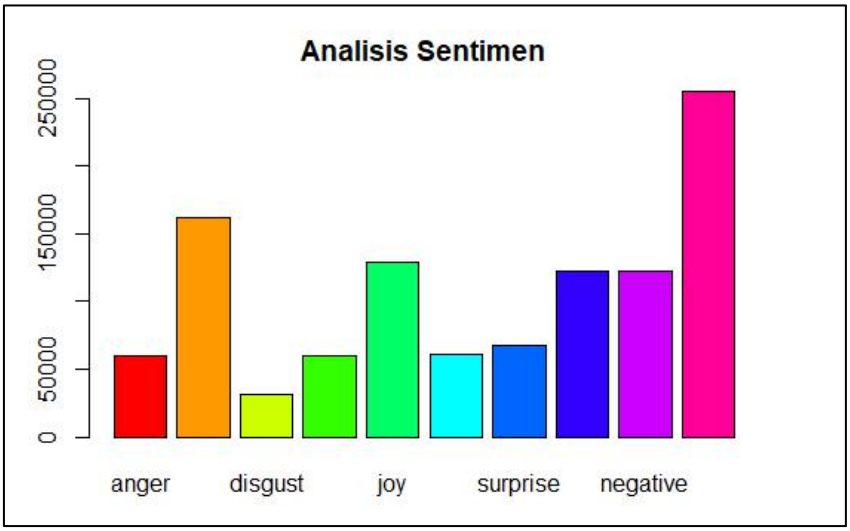
par(mar=rep(3,4))

a     <-     barplot(colSums(s),     col=rainbow(10),     ylab='count',
main='Analisis Sentimen')

brplt <- a

```
```

Listing Program 3.3 Output Barplot



Gambar 3.1 Output Barplot

3.4 Output Wordcloud

Pada proses menampilkan wordcloud dari data yang sudah melakukan cleansing diperlukan library tm, library RTextTools, library e1071, library dplyr dan library caret dan library wordcloud. Data ulasan_clean.csv akan dibersihkan dengan corpus (training, testing, matrix, dll) kemudian data akan ditampilkan.

```
```{r Wordcloud}
# Library untuk corpus dalam cleaning data
library(tm)
library(RTextTools)
# Library algoritma naivebayes
library(e1071)
library(dplyr)
library(caret)
df <- read.csv("ulasan_clean.csv",stringsAsFactors = FALSE)
```

```
glimpse(df)


# Set the seed of R‘s random number generator, which is useful for
creating simulations or random objects that can be reproduced.
set.seed(20)
df <- df[sample(nrow(df)),]
df <- df[sample(nrow(df)),]
glimpse(df)


corpus <- Corpus(VectorSource(df$text))
corpus
inspect(corpus[1:10])


# Membersihkan data yang tidak dibutuhkan
corpus.clean <- corpus%>%
    tm_map(content_transformer(tolower))%>%
    tm_map(removePunctuation)%>%
    tm_map(removeNumbers)%>%
    tm_map(removeWords,stopwords(kind="en"))%>%
    tm_map(stripWhitespace)
dtm <- DocumentTermMatrix(corpus.clean)


inspect(dtm[1:10,1:20])


df.train <- df[1:50,]
df.test <- df[51:100,]


dtm.train <- dtm[1:50,]
dtm.test <- dtm[51:100,]


corpus.clean.train <- corpus.clean[1:50]
corpus.clean.test <- corpus.clean[51:100]


dim(dtm.train)
```

```
fivefreq <- findFreqTerms(dtm.train,5)

length(fivefreq)


dtm.train.nb   <-   DocumentTermMatrix(corpus.clean.train,control   =
list(dictionary=fivefreq))


#dim(dtm.train.nb)


dtm.test.nb    <-    DocumentTermMatrix(corpus.clean.test,control    =
list(dictionary=fivefreq))


dim(dtm.test.nb)


convert_count <- function(x){

    y <- ifelse(x>0,1,0)

    y <- factor(y,levels=c(0,1),labels=c("no","yes"))

    y

}

trainNB <- apply(dtm.train.nb,2,convert_count)

testNB <- apply(dtm.test.nb,1,convert_count)

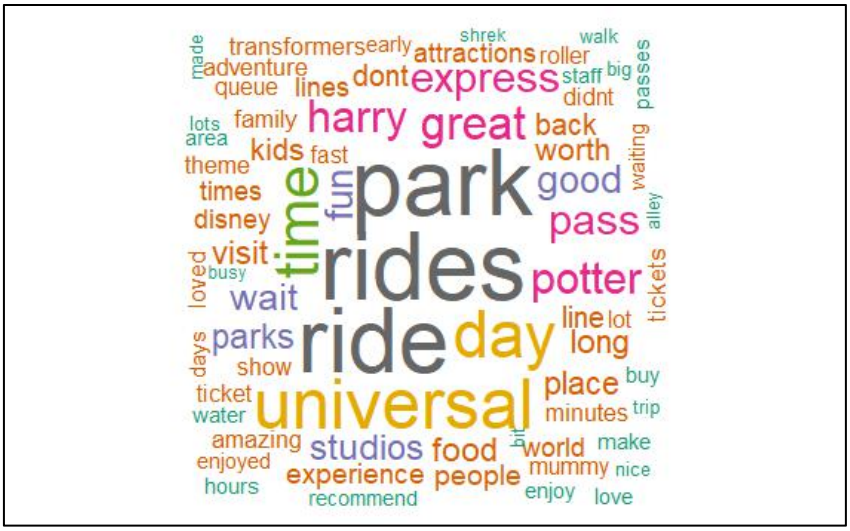



library(wordcloud)

wordcloud(corpus.clean,min.freq                                          =
4,max.words=100,random.order=F,colors=brewer.pal(8,"Dark2"))


```
```

Listing Program 3.4 Output Wordcloud
```



Gambar 3.2 Output Wordcloud

3.5 Output Aplikasi Shiny

Pada proses menampilkan wordcloud dari data yang sudah melakukan cleansing diperlukan library shiny, library here, library vroom, library dplyr, library ggplot2 dan library plotly. Data ulasan_clean.csv akan diproses dan ditampilkan pada aplikasi shiny berupa 4 kolom yaitu dataset asli, data cleansing, barplot dan wordcloud.

```
```{r Shiny App}
library(shiny)
library(here)
library(vroom)
library(dplyr)
library(ggplot2)
library(plotly)
library(syuzhet)
library(wordcloud)


universal_studio<- vroom(here("ulasan_clean.csv"))
universal_studio <- data.frame(universal_studio)
ui <- fluidPage(
    titlePanel("Analisis Sentimen pada Review Universal Studio"),
        mainPanel(
            tabsetPanel(type = "tabs",
```
```

```
                tabPanel("Data Universal Studio", DT::dataTableOutput('tbl2')), # Data Review

                tabPanel("Data Clean", DT::dataTableOutput('tbl')), # Data Clean

                tabPanel("Barplot", plotOutput("scatterplot")), # Plot


                tabPanel("Wordcloud", plotOutput("Wordcloud"))

              )

        )

    )

# SERVER

server <- function(input, output) {




    # Output Data Tabel

    output$tbl2 = DT::renderDataTable({

        DT::datatable(dataset <- vroom(here('universal_studio_branches.csv')), options = list(lengthChange = FALSE))

    })


    # Output Data Clean

    output$tbl = DT::renderDataTable({

        DT::datatable(universal_studio, options = list(lengthChange = FALSE))

    })


    # Output Plot

    output$scatterplot <- renderPlot({

par(mar=rep(3,4))

a  <-  barplot(colSums(s),col=rainbow(10),ylab='count',main='Sentimen Analisis')}, height=400)


    df <- read.csv("ulasan_clean.csv",stringsAsFactors = FALSE)

glimpse(df)


# Set the seed of R‘s random number generator, which is useful for
```

```
creating simulations or random objects that can be reproduced.
set.seed(20)
df <- df[sample(nrow(df)),]
df <- df[sample(nrow(df)),]
glimpse(df)

corpus <- Corpus(VectorSource(df$text))
corpus
inspect(corpus[1:10])

# Membersihkan data yang tidak dibutuhkan
corpus.clean <- corpus%>%
    tm_map(content_transformer(tolower))%>%
    tm_map(removePunctuation)%>%
    tm_map(removeNumbers)%>%
    tm_map(removeWords,stopwords(kind="en"))%>%
    tm_map(stripWhitespace)
dtm <- DocumentTermMatrix(corpus.clean)

inspect(dtm[1:10,1:20])

df.train <- df[1:50,]
df.test <- df[51:100,]

dtm.train <- dtm[1:50,]
dtm.test <- dtm[51:100,]

corpus.clean.train <- corpus.clean[1:50]
corpus.clean.test <- corpus.clean[51:100]

dim(dtm.train)
fivefreq <- findFreqTerms(dtm.train,5)
length(fivefreq)
```

```
dtm.train.nb   <-   DocumentTermMatrix(corpus.clean.train,control   =
list(dictionary=fivefreq))


#dim(dtm.train.nb)


dtm.test.nb    <-    DocumentTermMatrix(corpus.clean.test,control    =
list(dictionary=fivefreq))


dim(dtm.test.nb)


convert_count <- function(x){

    y <- ifelse(x>0,1,0)

    y <- factor(y,levels=c(0,1),labels=c("no","yes"))

    y

}

trainNB <- apply(dtm.train.nb,2,convert_count)

testNB <- apply(dtm.test.nb,1,convert_count)


    # Output WordCloud

    output$Wordcloud <- renderPlot({

    wordcloud(corpus.clean,min.freq                                    =
4,max.words=100,random.order=F,colors=brewer.pal(8,"Dark2"))

  })

}


# Program Shiny

shinyApp(ui = ui, server = server)

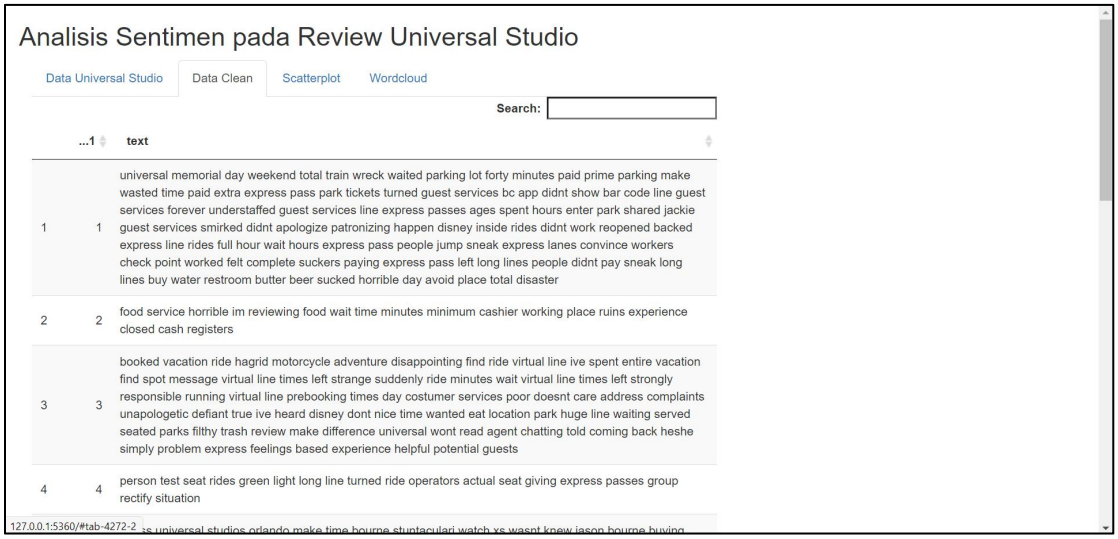
```

Listing Program 3.5 Program Shiny

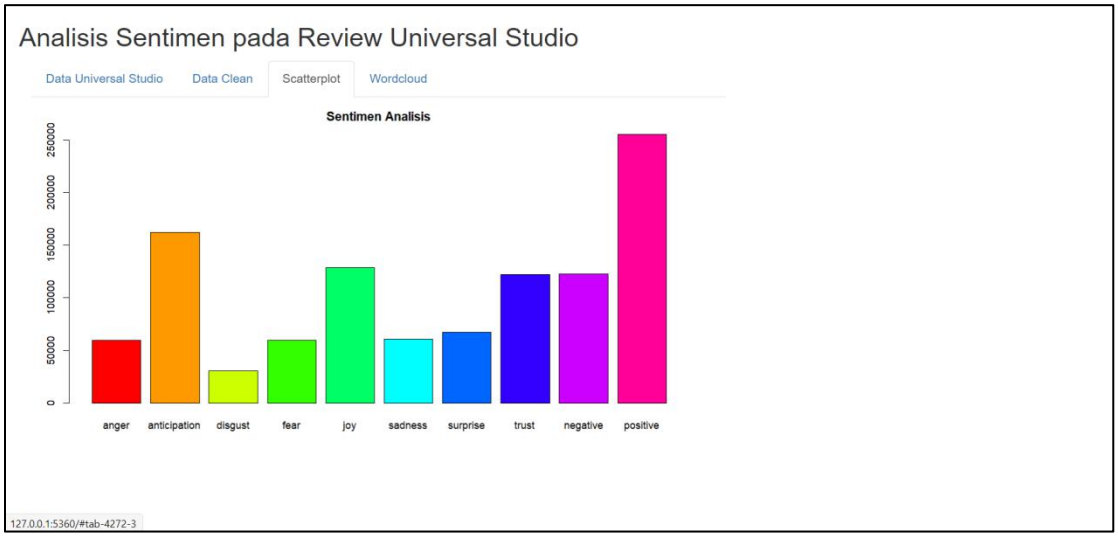
10	Jen	4	May 26, 2021	Good first time visit with kids	delivered, no one came by to refill drinks or ask us about our food. Thankfully, the food and experience at the Leaky Cauldron was 100% better. -- The "quick service" dining locations where you mobile ordered and they brought food to your table was incredibly slow- usually about a 45-60 minute wait for food. There was no way to order cups of water on the mobile device, and when we asked for it when the waiter brought our food they seemed quite put out by the request. Maybe this wait is normal, but it was unexpected for us. It was so hard to find water in the park and you always felt like you were the biggest annoyance when asking for it. Many of the freestyle machines (I had been told you could get water from for free) did not offer free water. --I stood in quite a few lines where the employees were much more interested in	Universal Studios Florida

127.0.0.1:5360/#tab-4272-1

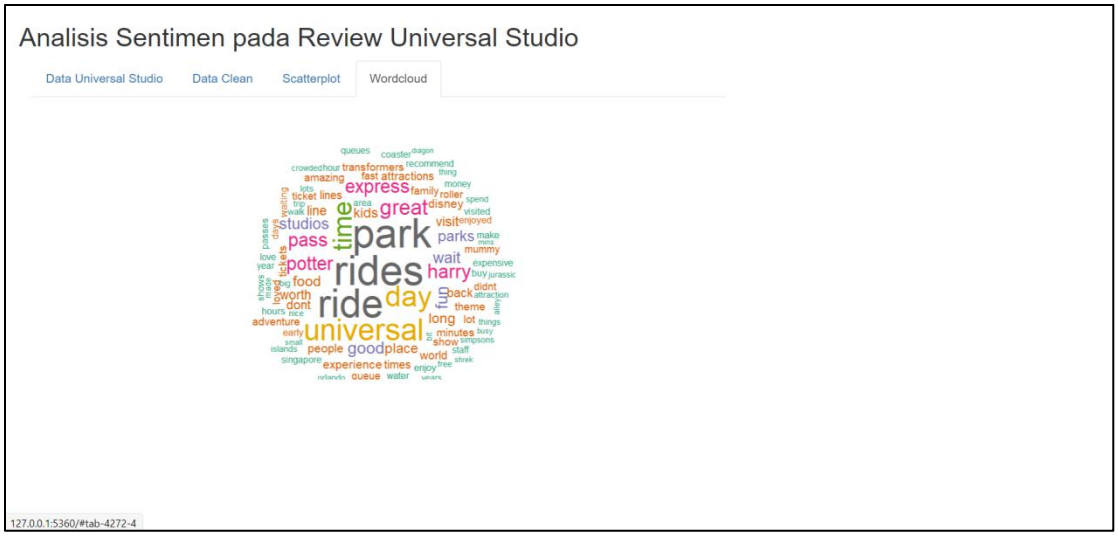
Gambar 3.3 Output Dataset Universal Studio Review Shiny



Gambar 3.4 Output Data Cleansing Shiny



Gambar 3.5 Output Barplot Shiny



Gambar 3.6 Output Wordcloud Shiny

4. KESIMPULAN

Berdasarkan dari hasil analisis, perancangan dan pembahasan yang telah dilakukan, dapat diambil kesimpulan antara lain :

1. Sistem ini dapat berjalan dengan baik, dapat menyeleksi data dari dataset dan diambil reviewnya saja, menghilangkan maupun mengganti karakter-karakter yang tidak dibutuhkan, dan menghasilkan analisis sentimen yang bernilai positif maupun negatif sesuai yang diharapkan.
2. Banyak-sedikitnya data yang digunakan akan mempengaruhi hasil analisis sentimen yang dihasilkan. Perbedaan jumlah data yang memungkinkan perbedaan hasil, sehingga dibutuhkan proses pengambilan data ulang dan pengujian ulang.
3. Dari hasil pengujian data review mengenai Universal Studio, didapatkan review paling besar adalah positif dan kata yang paling sering muncul adalah park, ride dan rides.