# **LAPORAN PROYEK AKHIR**

**PRAKTIKUM DATA SCIENCE**

**Analisis Sentimen Tweet Masyarakat Indonesia Tentang Covid dan Pemerintah dengan Metode Naive Bayes**



|  |  |
| --- | --- |
| Monica Elisabet | 123200082 |
| Rifqi Maulana | 123200128 |

**PROGRAM STUDI INFORMATIKA**

**JURUSAN TEKNIK INFORMATIKA**

**FAKULTAS TEKNIK INDUSTRI**

**UNIVERSITAS PEMBANGUNAN NASIONAL “VETERAN” YOGYAKARTA**

**2022**

**HALAMAN PENGESAHAN LAPORAN PROYEK AKHIR**



Disusun oleh :

*Monica Elisabet 123200082*

*Rifqi Maulana 123200128*

Telah Diperiksa dan Disetujui oleh Asisten Praktikum Data Science

Pada Tanggal : ................................

.

|  |  |  |
| --- | --- | --- |
| **Asisten Praktikum**  **Dhea Anggita**  123190046. |  | **Asisten Praktikum**  **Rico Aminanda**  123190079 |

# **PENDAHULUAN**

Akhir tahun, 2019 dunia digemparkan oleh kemunculan sebuah virus yang penularannya sangat cepat, yang berawal dari Negara China, tepatnya di Wuhan. Virus tersebut ialah virus corona (covid-19). Dan di awal tahun 2020, virus tersebut masuk ke Indonesia.   
Virus Corona (COVID 19) adalah penyakit menular yang disebabkan oleh virus SARS-CoV-2. Sebagian besar orang yang tertular COVID 19 akan mengalami gejala ringan hingga sedang, dan akan pulih tanpa penanganan khusus di rumah sakit. Namun, sebagian orang lagi akan mengalami sakit parah dan memerlukan bantuan medis dengan penanganan khusus. Gejala paling umum yang dialami ialah demam, batuk, kelelahan, dan kehilangan rasa atau bau. Virus ini dapat menyebar dari mulut atau hidung orang yang terinfeksi melalui partikel kecil ketika orang tersebut batuk, bersin, berbicara atau bernafas. Seseorang dapat tertular dengan menghirup udara yang mengandung virus jika berada di dekat orang yang terinfeksi COVID 19.

Tercatat hingga 29 November 2022, kasus covid di Indonesia mencapai angka 6.653.469 kasus dengan 159.735 kematian, sedangkan untuk kasus dunia 641.560.304 kasus dengan 6.631.765 kasus meninggal. Namun, seiring berjalannya waktu, sudah ditemukan vaksin yang tujuannya untuk mencegah dan mengurangi penularan virus covid 19 ini, bukan berarti dengan vaksin tersebut dapat menghentikan penularan covid-19, vaksin tersebut hanya menambah kekebalan tubuh agar tidak mudah menerima penularan. Ditemukan juga banyak kasus orang-orang yang sudah divaksin pun masih terjangkit virus tersebut. Saat ini sudah ada 3 tahapan vaksin, vaksin 1, vaksin 2, dan vaksin booster. Selain dengan vaksin, pemerintah Indonesia juga sempat memberlakukan PPKM di beberapa waktu, untuk membatasi pergerakan masyarakat Indonesia.   
Beberapa hal yang terjadi diatas, tidak luput dari pandangan warganet terhadap pemerintahan Indonesia, termasuk berpendapat di social media, salah satunya twitter. Twitter adalah layanan jejaring sosial dan mikroblog daring yang memungkinkan penggunanya untuk mengirim dan membaca pesan berbasis teks hingga 140 karakter akan tetapi pada tanggal 07 November 2017 bertambah hingga 280 karakter yang dikenal dengan sebutan *kicauan* (tweet).  
Twitter menjadi social media yang banyak digunakan untuk menuliskan cuitan. Dari data-data Komentar, pendapat dan kicauan yang ada pada twitter mengenai perjalanan covid dari awal hingga sekarang, kita dapat melakukan analisis. Dimana komentar dan pendapat tersebut sifatnya yaitu positif dan negatif, ada yang pro dan ada yang kontra. Namun tentunya kita akan mengalami kesulitan, karena banyak symbol dan jenis-jenis penulisan yang dilakukan tentunya berbeda-beda. Untuk mengatasi persoalan tersebut dapat dilakukan perubahan format yang tidak terstruktur menjadi terstruktur, dan pengklasifikasiannya dapat dilakukan dengan analisis sentimen. Salah satu algoritma yang bisa digunakan untuk analisis sentimen adalah Naïve Bayes. Kelebihan menggunakan Naïve Bayes adalah memiliki model yang sederhana serta penerapannya cepat dan ada iterasi.

# **METODE** Data yang kami ambil berasal dari dataset pada kaggle.com dengan judul, tweet masyarakat tentang covid-19. Dan metode yang dilakukan untuk menyelesaikan project akhir Data Science ini ialah metode naïve bayes. Naive Bayes adalah algoritma machine learning untuk masalah klasifikasi. Ini didasarkan pada teorema probabilitas Bayes. Hal ini digunakan untuk klasifikasi teks yang melibatkan set data pelatihan dimensi tinggi. Beberapa contohnya adalah penyaringan spam, analisis sentimental, dan klasifikasi artikel berita. Metode bayes merupakan salah satu metode pengambilan keputusan yang telah banyak dipakai, metode ini dikembangkan untuk menyelesaikan masalah pengambilan keputusan dengan menentukan nilai peluang dari kejadian dan nilai evidence (bukti) yang didapat dari fakta tentang objek yang diteliti. Tidak hanya dikenal karena kesederhanaannya, tetapi juga karena keefektifannya. Sangat cepat untuk membangun model dan membuat prediksi dengan algoritma Naive Bayes. Naive Bayes adalah algoritma pertama yang harus dipertimbangkan untuk memecahkan masalah klasifikasi teks.

Project ini menggunakan analisis sentiment untuk menentukan hasil klasifikasi terhadap tweet masyarakat Indonesia mengenai covid dan pemerintahan. Sentiment analysis merupakan salah satu bidang dari Natural Languange Processing (NLP) yang membangun sistem untuk mengenali dan mengekstraksi opini dalam bentuk teks. Informasi berbentuk teks saat ini banyak terdapat di internet dalam format forum, blog, media sosial, serta situs berisi review. Dengan bantuan sentiment analysis, informasi yang tadinya tidak terstruktur dapat diubah menjadi data yang lebih terstruktur.  
Sentiment analysis adalah proses penggunaan text analytics untuk mendapatkan berbagai sumber data dari internet dan beragam platform media sosial. Tujuannya adalah untuk memperoleh opini dari pengguna yang terdapat pada platform tersebut.

Setiap hari, internet dibanjiri oleh miliaran data dari berbagai sumber. Sentiment analysis berperan sebagai alat yang dapat menghubungkan seluruh data tersebut. Dengan begitu, perusahaan dapat memperoleh masukan inti dari pengguna atau konsumen secara efisien.

# **HASIL DAN PEMBAHASAN**

3.1 Listing Program

|  |
| --- |
| # Load Dataset  ```{r}  rawData = read.csv("tweet-pemerintah-dan-covid.csv", header = T)  ```  # Ubah Data (Tabel Tweet) Menjadi Vector  ```{r}  tweets = rawData$tweet  tweets.text = Corpus(VectorSource(tweets))  ```  # Cleaning Data  ```{r}  removeURL <- function(x) gsub("http[^[:space:]]\*", "",x)  clean <- tm\_map(tweets.text, removeURL)  removeNL <- function(y) gsub("\n", " ", y)  clean <- tm\_map(clean, removeNL)  removepipe <- function(z) gsub("<[^>]+>", "", z)  clean <- tm\_map(clean, removepipe)  remove.mention <- function(z) gsub("@\\S+", "", z)  clean <- tm\_map(clean, remove.mention)  remove.hashtag <- function(z) gsub("#\\S+", "", z)  clean <- tm\_map(clean, remove.hashtag)  removeamp <- function(y) gsub("&amp;", "", y)  clean <- tm\_map(clean, removeamp)  removetitik3 <- function(y) gsub("[[:punct:]]", "", y)  clean <- tm\_map(clean, removetitik3)  remove.all <- function(xy) gsub("[^[:alpha:][:space:]]\*", "", xy)  clean <- tm\_map(clean,remove.all)  clean <- tm\_map(clean, tolower)  ## Menghilangkan Extra Whitespace (Spasi)  clean <- tm\_map(clean, stripWhitespace)  ## Load Stopword-ID  stopwordID <- "ID-stopwords.txt"  ## Membaca Stopword-ID per-baris  cStopwordID <- readLines(stopwordID)  ## Load Slangword (Bahasa Gaul)  slang <- read.csv("Slangword.csv", header = T)  old\_slang <- as.character(slang$old)  new\_slang <- as.character(slang$new)  ## Load Stemming (Kata Imbuhan)  stemm <- read.csv("Stemming.csv", header = T)  old\_stemm <- as.character(stemm$old)  new\_stemm <- as.character(stemm$new)  ## Load Lemmatization (Pengelompokan Infleksi Kata)  lemma <- read.csv("Lemmatization.csv", header = T)  old\_lemma <- as.character(stemm$old)  new\_lemma <- as.character(stemm$new)  stemmword <- function(x) Reduce(function(x,r) gsub(stemm$old[r],stemm$new[r],x,fixed=T),  seq\_len(nrow(stemm)),x)  clean <- tm\_map(clean,stemmword)  slangword <- function(x) Reduce(function(x,r) gsub(slang$old[r],slang$new[r],x,fixed=T),  seq\_len(nrow(slang)),x)  clean <- tm\_map(clean,slangword)  lemmatization <- function(x) Reduce(function(x,r) gsub(lemma$old[r],lemma$new[r],x,fixed=T),  seq\_len(nrow(lemma)),x)  clean <- tm\_map(clean,lemmatization)  clean <- tm\_map(clean, removeWords, cStopwordID)  writeLines(strwrap(clean[[2]]$content, 100))  ```  # Save Data  ```{r}  dataframe = data.frame(text = unlist(sapply(clean, `[`)), stringsAsFactors = F)  View(dataframe)  write.csv(dataframe, file = 'dataTweetBersih.csv')  ``` |

**Listing Program** 3.1.1 Cleaning Data

|  |
| --- |
| # Library Awal  ```{r}  library(vroom)  library(here)  library(sentimentr)  library(tidytext)  library(tidyr)  library(textclean)  library(tibble)  library(tm)  ```  # Library Analisis Sentimen  ```{r}  library(e1071)  library(caret)  library(syuzhet)  ```  # Library WordCloud  ```{r}  library(wordcloud) # Library Wordcloud  library(tm) # Library Penggunaan Corpus dalam Cleaning Data  library(RTextTools) # Library Penggunaan Corpus dalam Cleaning Data  library(e1071) # Library yang di dalamnya terdapat sebuah Algoritma Naive Bayes  library(dplyr) # Library yang di dalamnya terdapat sebuah Algoritma Naive Bayes  library(caret) # Library yang di dalamnya terdapat sebuah Algoritma Naive Bayes  ```  # Library Shiny  ```{r}  library(shiny)  ``` |

**Listing Program** 3.1.2 Library yang digunakan

|  |
| --- |
| # SENTIMENT ANALYSIS  # Library  ```{r}  library(e1071)  library(caret)  library(syuzhet)  ```  # Membaca File Data .csv yang sudah melewati proses Cleaning Data  ```{r}  data <- read.csv("dataTweetBersih.csv", stringsAsFactors = FALSE)  ## Deklarasi Variabel Teks Kolom (Column Text) Menjadi Char  tweetss <- as.character(data$text)  ## Memanggil NRC Sentiment Dictionary untuk menghitung berbagai emosi  s <- get\_nrc\_sentiment(tweetss)  tweetss\_combine <- cbind(data$text,s)  par(mar=rep(3,4))  a <- barplot(colSums(s), col = rainbow(10), ylab = 'count', main = 'Sentiment Analisis')  iki\_ba <- a  ``` |

**Listing Program** 3.1.3 Sentiment Analysis

|  |
| --- |
| # WORDCLOUD  # Library  ```{r}  library(wordcloud) # Library Wordcloud  library(tm) # Library Penggunaan Corpus dalam Cleaning Data  library(RTextTools) # Library Penggunaan Corpus dalam Cleaning Data  library(e1071) # Library yang di dalamnya terdapat sebuah Algoritma Naive Bayes  library(dplyr) # Library yang di dalamnya terdapat sebuah Algoritma Naive Bayes  library(caret) # Library yang di dalamnya terdapat sebuah Algoritma Naive Bayes  ```  # Membaca Data yang sudah dibersihkan  ```{r}  df <- read.csv("dataTweetBersih.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 dari R random number generator, yang mana berguna untuk membuat simulasi atau obyek random yang bisa diproduksi ulang  ```{r}  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-data yang tidak dibutuhkan  ```{r}  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)  ```  # Generate Wordcloud  ```{r}  wordcloud(corpus.clean,min.freq = 4,max.words=100,random.order=F,colors=brewer.pal(8,"Dark2"))  ``` |

**Listing Program** 3.1.4 Wordcloud

|  |
| --- |
| # SHINY (Membuat Tampilan)  # Library  ```{r}  library(shiny)  ```  # Script UI  ```{r}  ui <- fluidPage(    ## Judul Aplikasi  titlePanel("Analisis Sentimen Terhadap Tweet Masyarakat Indonesia Tentang Covid dan Pemerintah"),  mainPanel(  ##textOutput("selected\_var"),    ##plotOutput("asPlot"),  tabsetPanel(type = "tabs",  tabPanel("Data Asli", DT::dataTableOutput('tbl')),  tabPanel("Data Bersih", DT::dataTableOutput('tbl2')),  tabPanel("Scatterplot", plotOutput("asPlot")),  tabPanel("Wordcloud", plotOutput("wordcl"))  )  )  )  ```  # Script Server  ```{r}  ## Membuat fungsi input, output, session  ## Di dalam badan fungsi berisi seluruh kode pemrosesan data  ## Membuat input dan menampilkan hasil pada output  server <- function(input, output, session) {  as\_data <- reactive({    input$Update  isolate({  withProgress({  setProgress(message = "Processing analisis...")  as\_file <- input$as  if(!is.null(as\_file)){  as\_text <- readLines(as\_file$datapath)  }  else  {  as\_text <- "A Barplot is an immage made of words that..."  }    })  })  })    barplot\_rep <- repeatable(barplot)    # Menampilkan data asli  output$tbl = DT::renderDataTable({  DT::datatable(rawData, options = list(lengthchange = FALSE))  })  # Menampilkan Data Bersih  output$tbl2 = DT::renderDataTable({  DT::datatable(data, options = list(lengthchange = FALSE))  })  # Menampilkan Grafik Sentimen  output$asPlot <- renderPlot({ withProgress({  setProgress(message = "Creating barplot...")  barplot(colSums(s),col = rainbow(10),ylab = 'count',main = 'Sentiment Analysis')  })  })  # Menampilkan Wordcloud  output$wordcl <- renderPlot({  wordcloud(corpus.clean,min.freq = 4,max.words=100,random.order=F,colors=brewer.pal(8,"Dark2"))  })    }  ```  # Running Shiny  ```{r}  shinyApp(ui = ui, server = server, options = list(height = "600px"))  ``` |

**Listing Program** 3.1.5 Pemodelan dan Tampilan Shiny

3.2 Output Program

|  |
| --- |
|  |

**Output Gambar** 3.2.6 Tampilan data sebelum cleaning

|  |
| --- |
|  |

**Output Gambar** 3.2.7 Tampilan data setelah cleaning

|  |
| --- |
|  |

**Output Gambar** 3.2.8 Tampilan Grafik Analisis Sentimen

|  |
| --- |
|  |

**Output Gambar** 3.2.9 Tampilan Data Asli pada Shiny

|  |
| --- |
|  |

**Output Gambar** 3.2.10 Tampilan Data Bersih pada Shiny

|  |
| --- |
|  |

**Output Gambar** 3.2.11 Tampilan Grafik Analisis Sentimen pada Shiny

|  |
| --- |
|  |

**Output Gambar** 3.2.12 Tampilan Wordcloud pada Shiny

|  |
| --- |
|  |

**Output Gambar** 3.2.13 Tampilan Wordcloud pada Shiny

# **KESIMPULAN**

Kesimpulan dari analisis sentimen terhadap tweet masyarakat Indonesia mengenai covid dan pemerintahan lebih menitikberatkan ke hasil yang negatif, sedangkan ekspresi lainnya bermacam-macam, dengan hasil positif, santai, takut,dan lainnya.

Dari sentiment analisis ini juga kita dapat mengetahui bagaiamana penilaian masyarakat mengenai covid-19, baik dari penularan, dampak,serta upaya-upaya yang dilakukan pemerintah untuk menangaani hal tersebut. Serta pemerintah juga dapat mengetahui ulasan masyarakat mengenai program-program yang sudah dilakukan , sehingga dapat diperbaiki kedepannya jika hal tersebut menuai banyak kritikan dan hal. negatif.  
Project ini juga masih jauh dari kata sempurna dan membutuhkan evaluasi dan perbaikan lebih lanjut lagi