# Twit tentang Gojek

## Contents



## Pendahuluan

Data yang digunakan di sini diambil dengan menggunakan beberapa parameter, yaitu:

1. tagar #savegojek
2. tagar #savedrivergojek
3. tagar #saveojekonline
4. ceritan transportasi online timeline

Tujuan ekplorasi adalah:

1. Memisahkan data yang dapat dianalisis dan yang tidak (duplicate dan atau kontennya tidak relevan)
2. Mengetahui utama (jika ada)
3. Mengetahui isi yang dibahas

## Library

```
# Runing RJava
if (Sys.info()['sysname'] == 'Darwin') {
  libjvm <- paste0(system2('/usr/libexec/java_home',stdout = TRUE)[1],'/jre/lib/server/libjvm.dylib')
  message (paste0('Load libjvm.dylib from: ',libjvm))
  dyn.load(libjvm)
}
```

```
library(lubridate)
library(tidyverse)
library(tidytext)
library(stringr)
library(tm)
library(reshape2)
library(scales)
library(AnomalyDetection)
library(igraph)
library(ggraph)
library(topicmodels)
library(SnowballC)
library(RWeka)
```

## Import Data

```
dirwd <- paste(getwd(),"/wrangled data proj-2/",sep='')
twit_gojek <- read.csv(paste(dirwd,"twit-gojek.csv",sep=''),
                       header = TRUE, sep = ",", stringsAsFactors = FALSE)
```

## Eksplorasi 1 - Perbandingan Sumber data

Data yang digunakan di sini bersumber dari beberapa parameter yang digunakan untuk scrapping. Gambar di bawah ini, menunjukkan perbandingan jumlah twit yang dihasilkan oleh masing-masing parameter.

```
twit_gojek %>%
  group_by(parameter) %>%
  count(parameter) %>%
  ggplot(aes(parameter, n)) + geom_col() +
  labs( x = "Parameter Pencarian", y = "Jumlah Twit")
```

Sebagian besar (mayoritas data) berasal dari twit yang didapat dengan menggunakan parameter #savegojek

Namun, data di atas masih kotor (ada yang duplicate dan tidak relevan). Salah satunya adalah twit yang diposting oleh akun @poocongs, di mana akun tersebut menggunakan tagar #savegojek dalam twitnya, namun isinya tidak membahas tentang gojek. Untuk itu, dalam gambar data akan dibersihkan terlebih dahulu dari:

1. twit duplicate
2. twit dari akun @poocongs

```
twit_gojek %>%
  filter(is_duplicate == FALSE) %>%
  filter(!user == "@poocongs") %>%
  group_by(parameter) %>%
  count(parameter) %>%
  ggplot(aes(parameter, n)) + geom_col() +
  labs( x = "Parameter Pencarian", y = "Jumlah Twit")
```

Gambar di atas menunjukkan bahwa distribusi twit masih didominasi oleh tagar #savegojek. Namun dilihat dari segi jumlahnya, twit dengan tagar tersebut sudah menunurun dibanding sebelumnya.

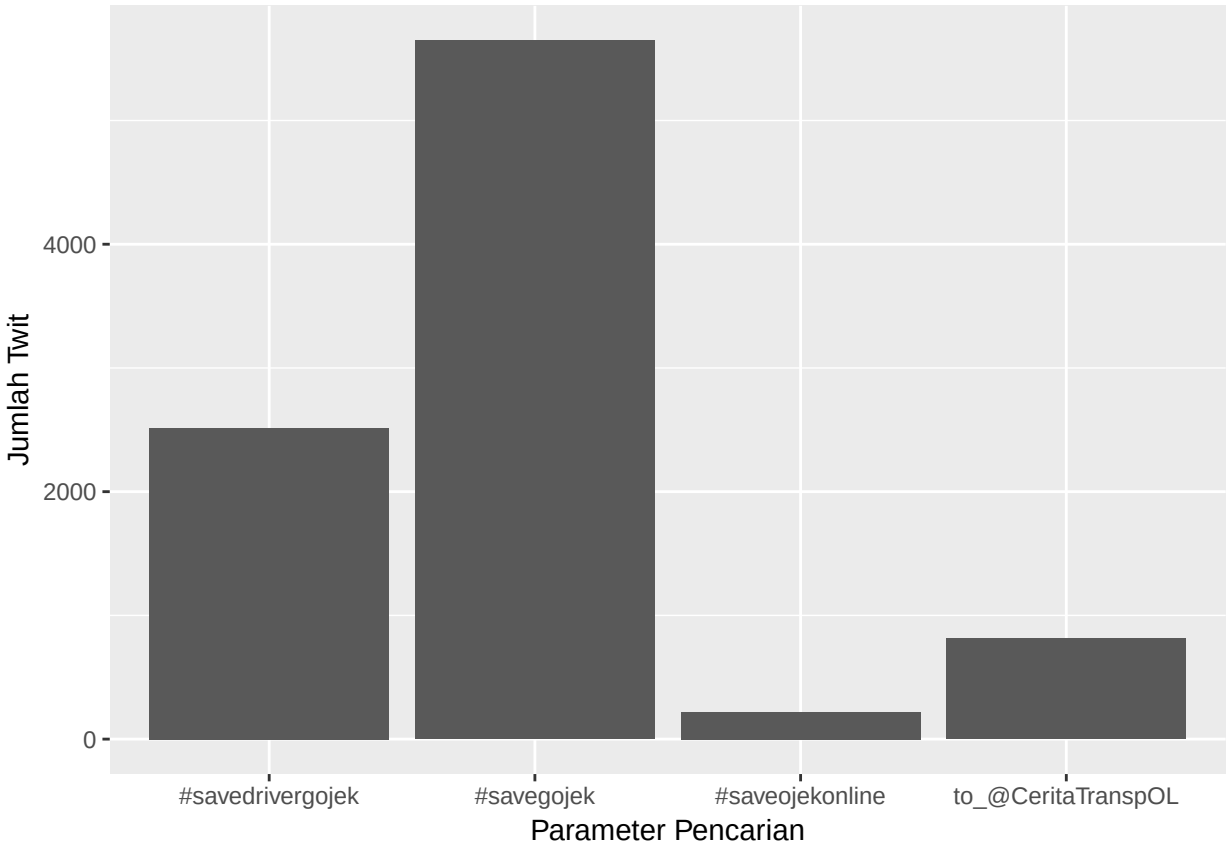


Figure 1: Perbandingan twit antar parameter

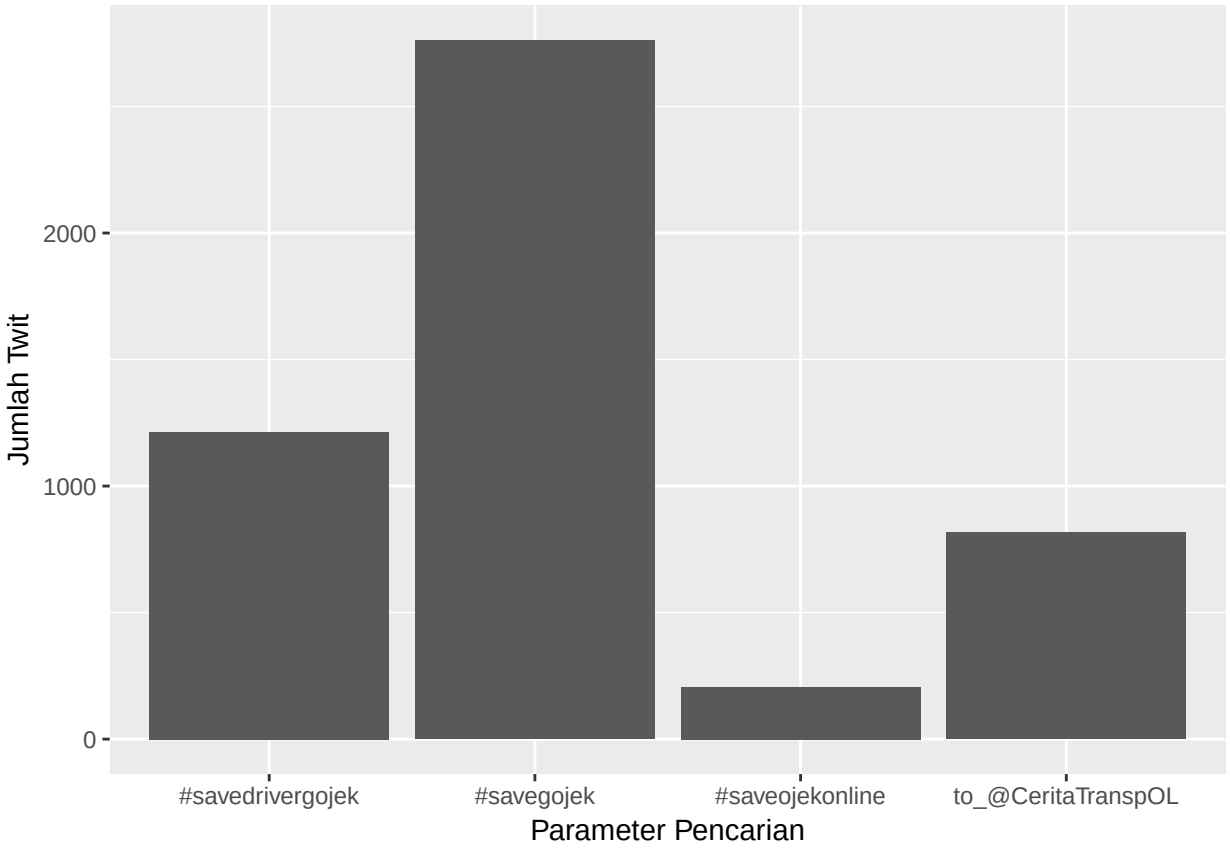


Figure 2: Perbandingan twit antar parameter setelah cleaning

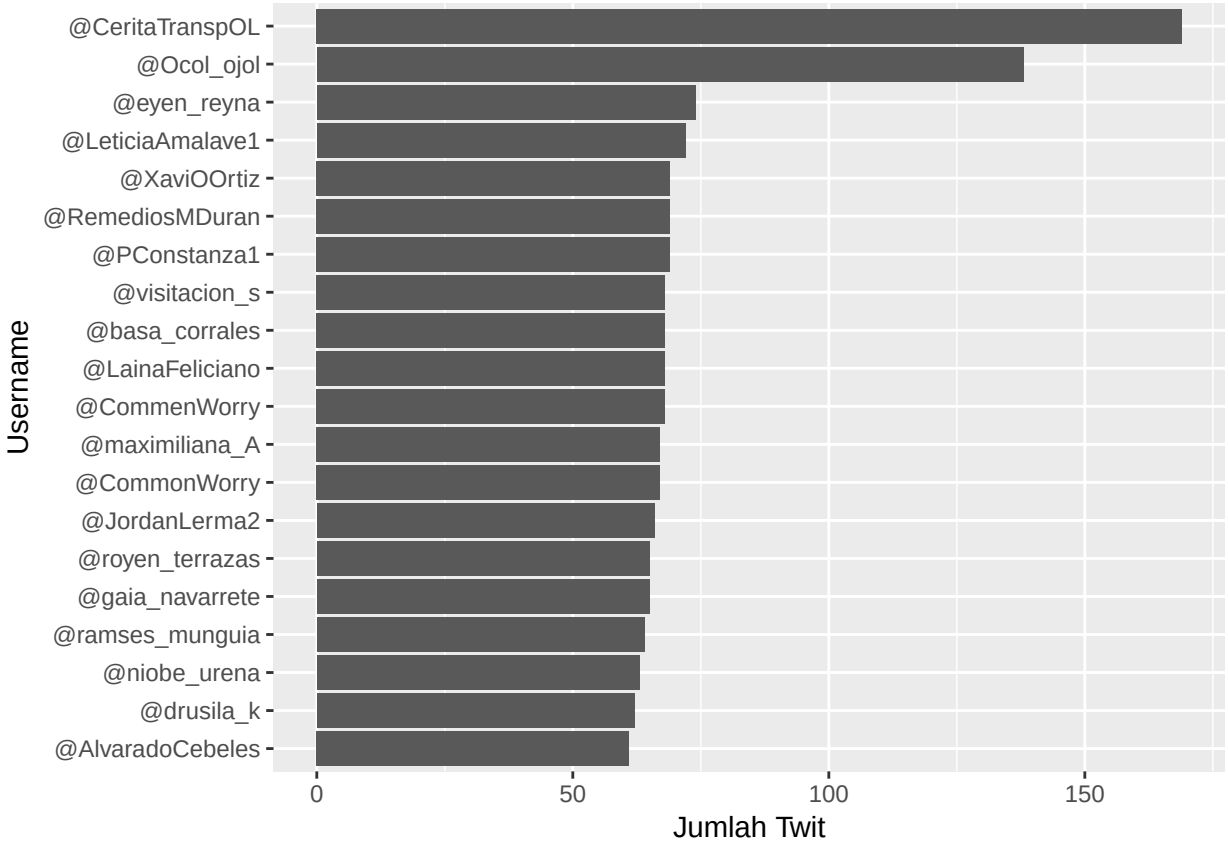


Figure 3: Username pengirim twit terbanyak

## Eksplorasi 2 - Username pengirim twit

```
twit_gojek %>%
  filter(is_duplicate == FALSE) %>%
  filter(!user == "@poocongs") %>%
  group_by(user) %>%
  count(user, sort = TRUE) %>%
  head(n = 20) %>%
  ggplot(aes(x = reorder(user, n), y = n)) + geom_col() + coord_flip() +
  labs( x = "Username", y = "Jumlah Twit")
```

Setelah dibersihkan total masih tersisi **4995** twit dengan daftar username yang paling banyak mengirim twit di atas. Selanjutnya, ekploasi dilakukan terhadap asal twit berdasarkan waktu

## Eksplorasi 3 - Distribusi twit

```
twit_gojek$date <- as.Date(twit_gojek$date)

twit_gojek %>%
  filter(is_duplicate == FALSE) %>%
  filter(!user == "@poocongs") %>%
```

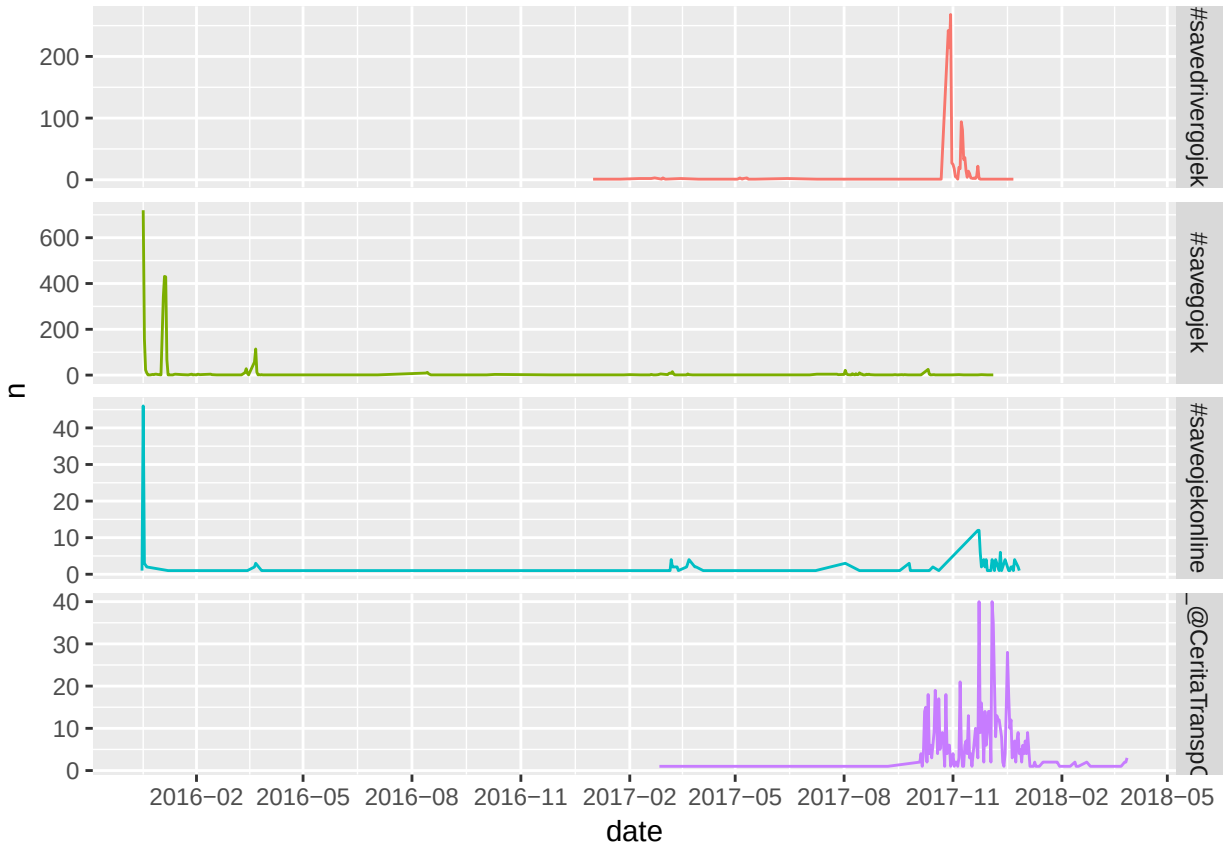


Figure 4: Perbandingan distribusi twit antar parameter

```
  group_by(parameter) %>%
  count(date, sort = TRUE) %>%
  ggplot(aes(x=date, y=n, colour=parameter)) +
  geom_line(show.legend = FALSE) +
  scale_x_date(labels = date_format("%Y-%m"),
               breaks = date_breaks("3 months")) +
  facet_grid(parameter~., scales="free") +
  theme(legend.position="top")
```

Gambar di atas menunjukkan asal waktu untuk masing-masing paramter pencarian twit. Twit dengan tagar #savedrivergojek berasal dari akhir tahun 2016 hingga akhir 2017, dengan puncaknya sekitar bulan 11 tahun 2017. Di sini parameter yang lain juga dapat dilihat tren penggunaan dan postingan twitnya.

## Eksplorasi 4 - Tagar

```
twit_gojek %>%
  filter(is_duplicate == FALSE) %>%
  filter(!user == "@poocongs") %>%
  select(hashtag) %>%
  unnest_tokens(tagar, hashtag) %>%
  count(tagar, sort = TRUE) %>%
  filter(!tagar == "savegojek") %>%
```

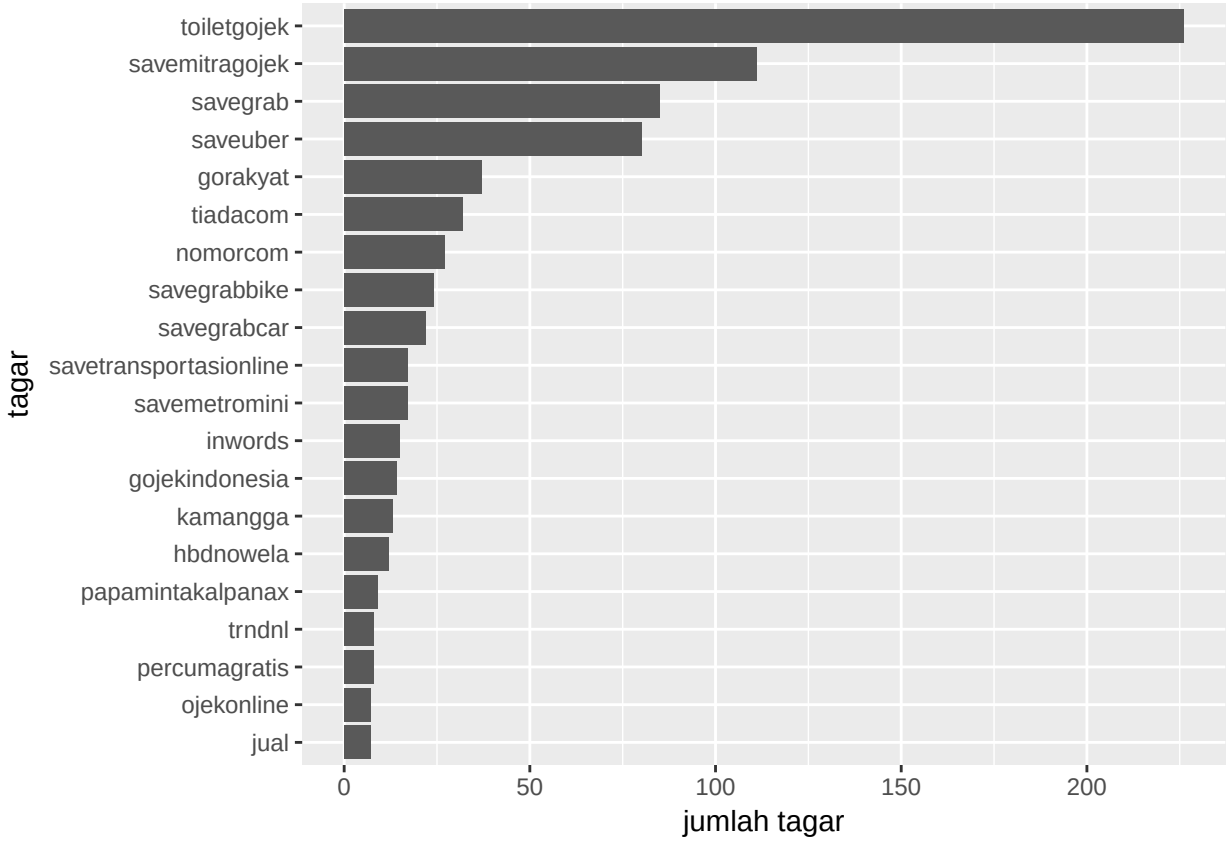


Figure 5: Tagar paling sering digunakan dalam keseluruhan twit

```
  filter(!tagar == "savedrivergojek") %>%
  filter(!tagar == "saveojekonline") %>%
  filter(!tagar == "gojek") %>%
  filter(!tagar == "akukaukita") %>%
  head(n = 20) %>%
  ggplot(aes(reorder(tagar, n), n)) + geom_col() + coord_flip() +
  labs(x = "tagar", y = "jumlah tagar")
```

Gambar di atas menunjukkan tagar yang paling sering digunakan setelah dilakukan filter terlebih dahulu. Tagar yang tidak diikutsertakan di antaranya adalah, tagar yang menjadi parameter pencarian, tagar gojek, dan akukaukita.

## Eksplorasi 5 - Term

Term yang digunakan dalam twit setelah di cleaning dapat dilihat pada gambar di bawah ini. Gambar menunjukkan 20 pasangan kata (bigram) yang paling sering muncul dalam data.

```
twit_gojek %>%
  filter(is_duplicate == FALSE) %>%
  filter(!user == "@poocongs") %>%
  select(clean_text) %>%
  unnest_tokens(kata, clean_text, token = "ngrams", n = 2) %>%
```

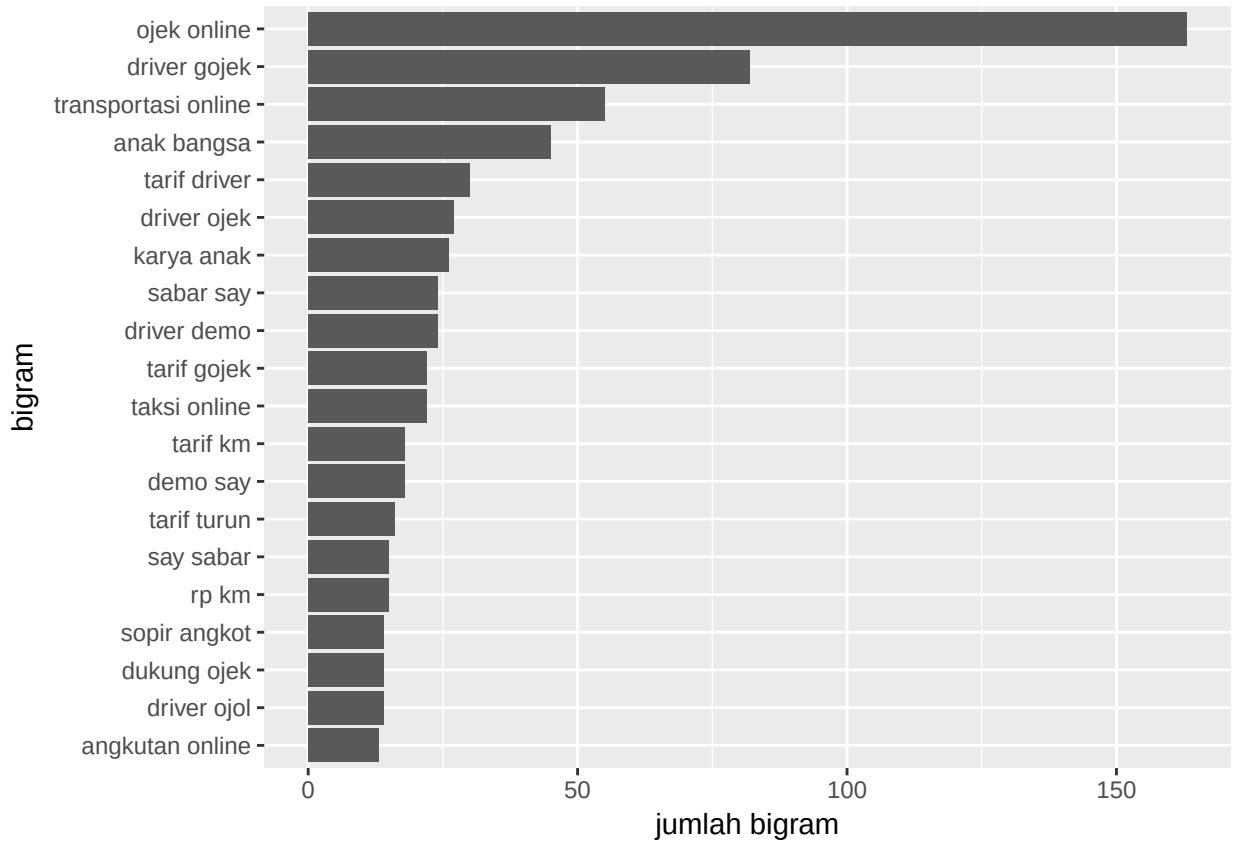


Figure 6: Bigram dengan frekuensi tertinggi

```
  count(kata, sort = TRUE) %>%
  head(n = 20) %>%
  ggplot(aes(reorder(kata, n), n)) + geom_col() + coord_flip() +
  labs(x = "bigram", y = "jumlah bigram")
```

Bigram yang paling sering muncul, yaitu “ojek online” menunjukkan bahwa dokumen teks sebagian besar membahas tentang ojek online atau transportasi online. Hal tersebut ditunjukkan dengan tingginya coocurrance kata ojek dengan online.

## Eksplorasi 6 - TF-IDF

Dari twit dengan tagar savedrivergojek, diketahui bahwa kata yang memiliki nilai tf-idf tertinggi adalah tarif driver. Hal tersebut merujuk pada tagar itu sendiri yang digunakan untuk menyarakan pendapat tentang kasus menyelamatkan dirver gojek. Diselamatkan dari apa? dan apa yang terjadi?

Dengan menggunakan data di atas, juga dapat diketahui bahwa salah satu isu yang dibahas adalah tentang tarif driver. Hal ini kemungkinan berkaitan dengan kebijakan tarif yang diterapkan oleh PT GI dan berdampak pada driver sebagai mitra, serta mendapatkan respons. Tagar lain juga dapat diinterpretasikan seperti itu, dengan jumlah kata dengan nilai tf-idf tertinggi yang diubah-ubah.

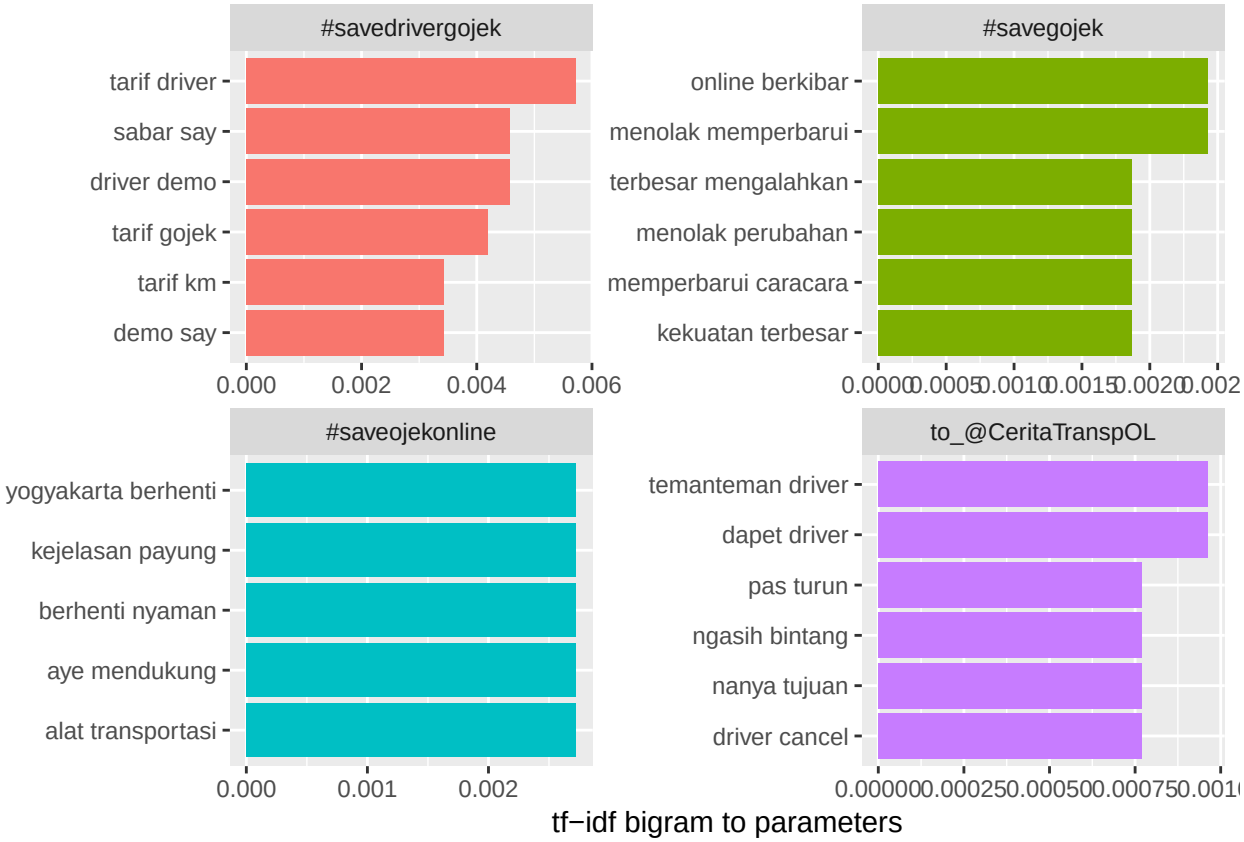


Figure 7: TF-IDF untuk masing parameter

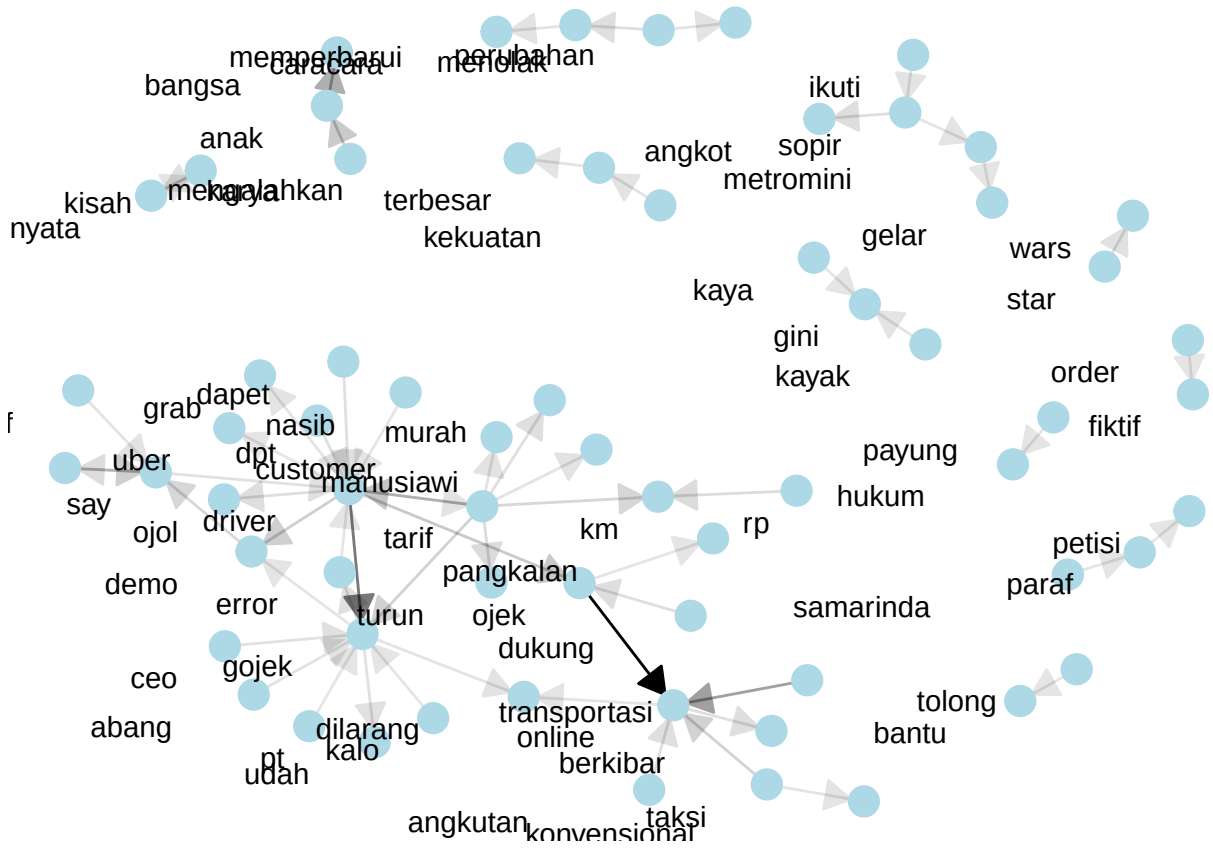


Figure 8: Semantic Network

## Eksplorasi 7 - Semantic Network

```
## [1] "parameter" "kata"      "n"         "total"     "tf"        "idf"
## [7] "tf_idf"
```

```
## IGRAPH a3f8746 DN-- 66 63 --
## + attr: name (v/c), n (e/n)
## + edges from a3f8746 (vertex names):
##  [1] ojek        ->online driver      ->gojek  transportasi->online
##  [4] anak        ->bangsa tarif       ->driver driver      ->ojek
##  [7] karya       ->anak   driver      ->demo   sabar       ->say
## [10] taksi       ->online tarif       ->gojek  demo        ->say
## [13] tarif       ->km     tarif       ->turun  rp          ->km
## [16] say         ->sabar  driver      ->ojol   dukung      ->ojek
## [19] sopir       ->angkot angkutan    ->online dpt         ->driver
## + ... omitted several edges
```

## Eksplorasi 8 - Topic Modelling

Menyiapkan file

```
# execution
tm_gojek <- twit_gojek %>%
```

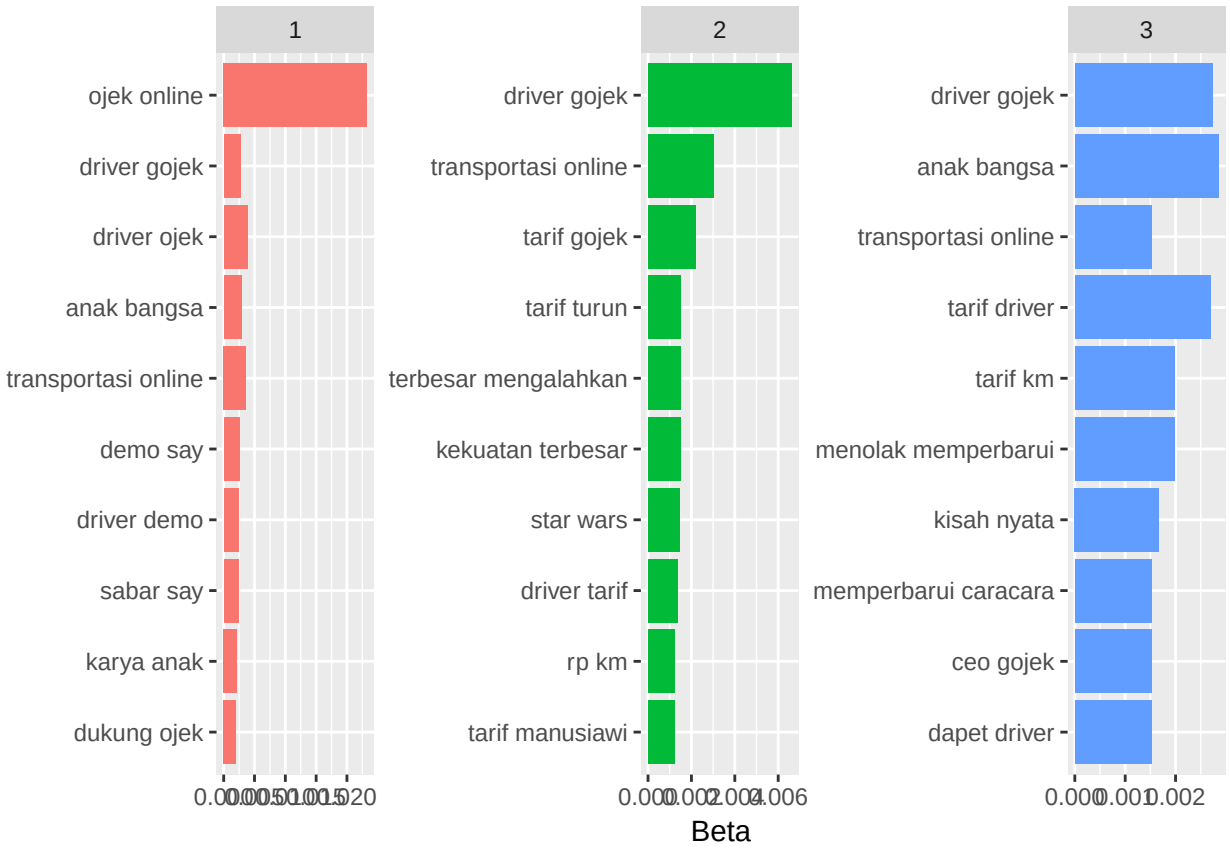


Figure 9: Hasil topic modelling

```
  filter(is_duplicate == FALSE) %>%
  filter(!user == "@poocongs") %>%
  select(date, user,clean_text, word_count, parameter) %>%
  filter(word_count >= 2)

tm_twit_gojek <- bigram_tm(tm_gojek$clean_text, number_of_topics = 3)
tm_twit_gojek
```