SentimentAnalysisProject

Condag, Gagante, Gonzaga

10/12/2024

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(syuzhet)
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.4     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1
## v readr     2.1.5
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(stringr)
library(readr)
library(lubridate)

iniTweets <- read.csv("tweetsDF.csv")

tweetsDf <- iniTweets
```

— Cleaning Data —

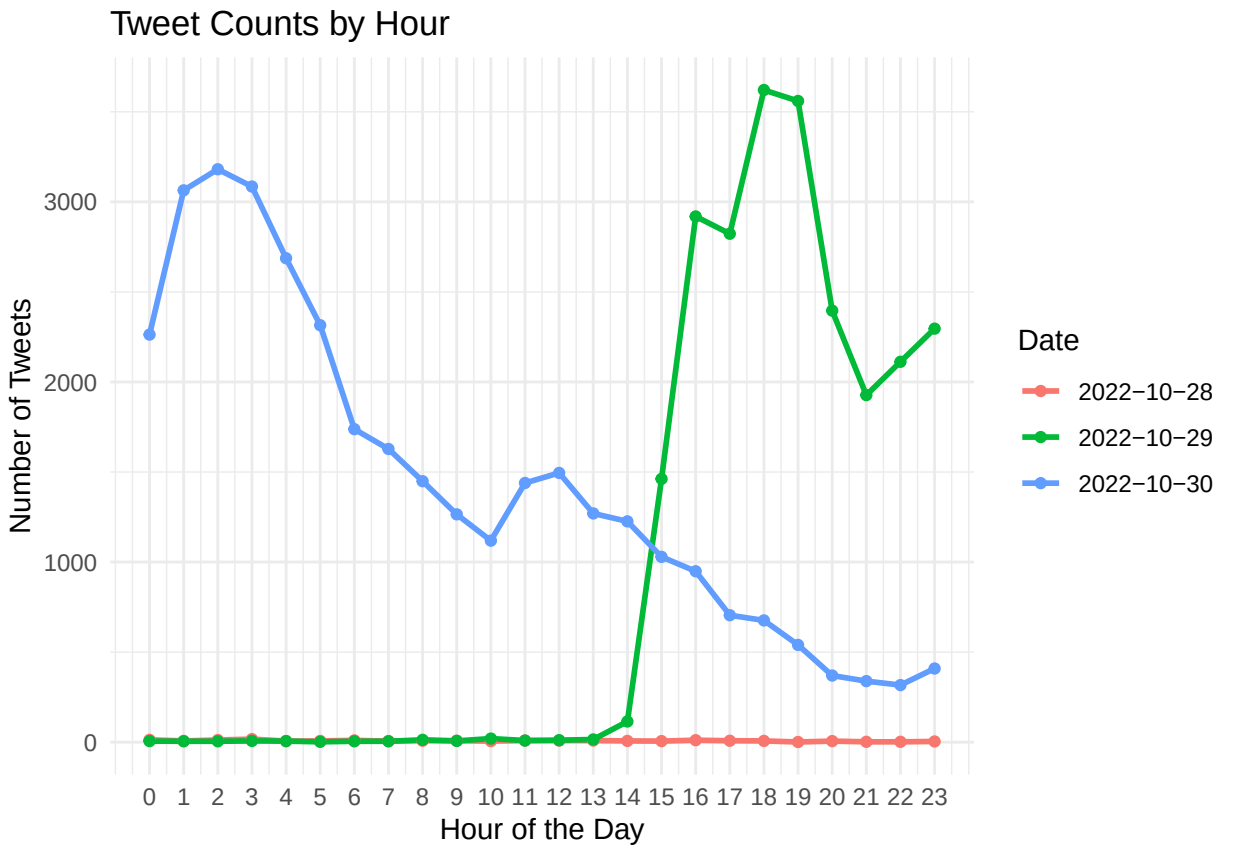
```
cleanTweets <- tweetsDf %>%
  mutate(text = str_to_lower(text),
         text = str_remove_all(text, "http\\S+"),
         text = str_remove_all(text, "[^\\w\\s]"),
         text = str_remove_all(text, "#\\w+"),
         text = str_remove_all(text, "@\\w+"),
         text = str_remove_all(text, "\\d+"),
```

```
         text = str_squish(text),
         sentiment = get_sentiment(text, method = "bing"))

write.csv(cleanTweets, "cleaned_tweets.csv", row.names = FALSE)
```

— Trend Analysis —

```
cleanTweets <- cleanTweets %>%
  mutate(date = ymd_hms(created)) %>%
  mutate(hour = hour(date))

byHour <- cleanTweets %>%
  group_by(day = as.Date(date), hour) %>%
  summarise(tweet_count = n(), .groups = "drop")

day1 <- byHour %>% filter(day == unique(day)[1])
day2 <- byHour %>% filter(day == unique(day)[2])
day3 <- byHour %>% filter(day == unique(day)[3])

all_days <- bind_rows(
  day1 %>% mutate(day_label = paste("Day", 1)),
  day2 %>% mutate(day_label = paste("Day", 2)),
  day3 %>% mutate(day_label = paste("Day", 3))
)


ggplot(all_days, aes(x = hour, y = tweet_count, group = day, color = as.factor(day))) +
  geom_line(size = 1) +
  geom_point() +
  scale_x_continuous(breaks = seq(0, 23, by = 1)) +
  labs(
    title = "Tweet Counts by Hour",
    x = "Hour of the Day",
    y = "Number of Tweets",
    color = "Date"
  ) +
  theme_minimal()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



The graph shows Hourly Tweet Counts from Day 1 to Day 3, with tweet data represented across different times of the day.

- In Day 1, The red line remains consistently flat at the bottom, indicating almost no tweeting activity throughout the 24 hours.
- In Day 2, The green line shows no activity until around 14:00 (2 PM), after which there is a sharp increase in tweet counts. The number of tweets spikes significantly between 14:00 and 18:00. Afterward, the tweet count drops but remains moderately high until the end of the day.
- In Day 3, The blue line shows consistent tweeting activity throughout the day, starting with a high volume (around 2000-3000 tweets) during the early morning hours (0:00 - 5:00). The tweet count gradually decreases throughout the morning and early afternoon. There is a slight dip and, followed by a minor uptick before declining again in the late afternoon.

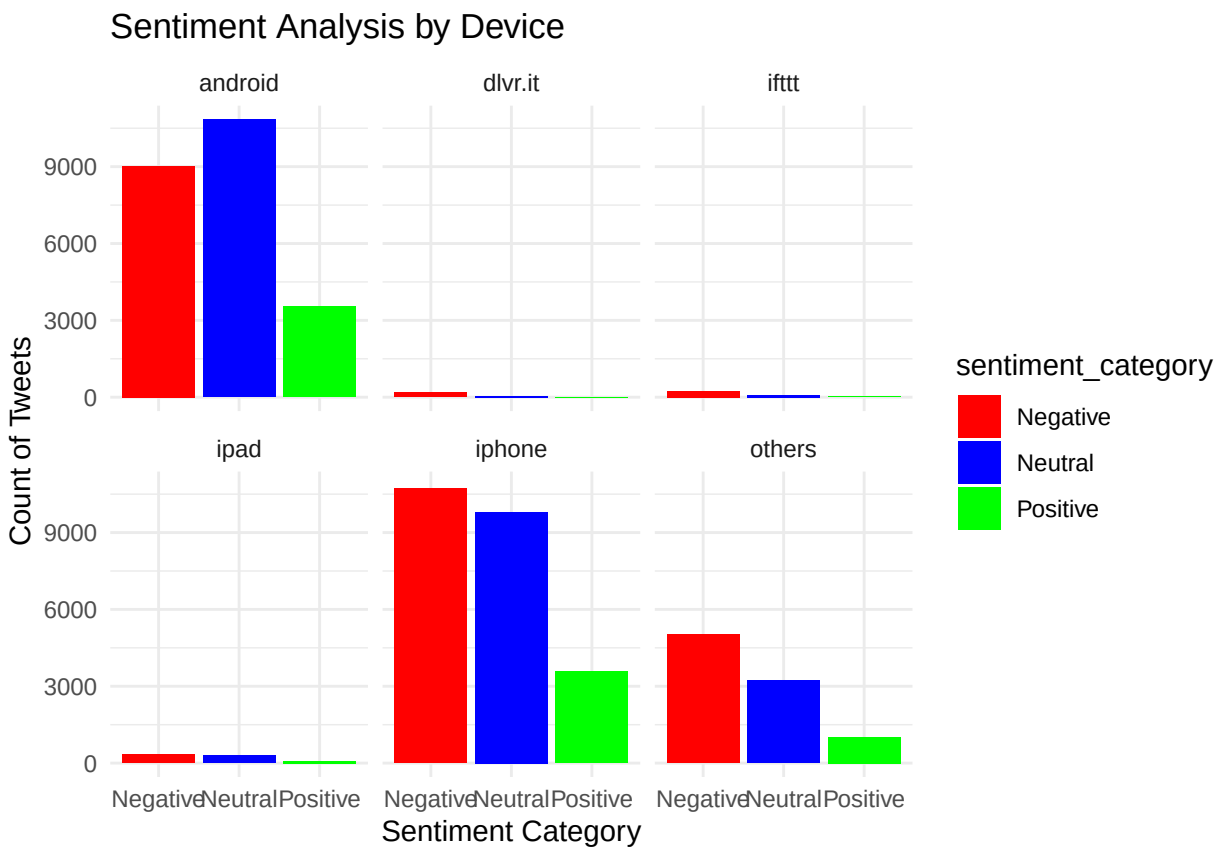
#INSIGHTS

- Day 1 had no significant activity, which could suggest either downtime, a lack of engagement, or a deliberate pause.
- Day 2 saw a dramatic surge in tweets during the afternoon, that triggered widespread engagement during that period.
- Day 3 maintained consistent tweeting, especially during early morning hours, indicating a sustained level of interest or activity before gradually tapering off.

— Sentiment Analysis —

```
cleaned_tweets <- read.csv("cleaned_tweets.csv")
```

```r
cleaned_tweets <- cleaned_tweets %>%
  mutate(sentiment_category = case_when(
    sentiment > 0 ~ "Positive",
    sentiment < 0 ~ "Negative",
    TRUE ~ "Neutral"
  ))

sentiment_by_device <- cleaned_tweets %>%
  group_by(tweetSource, sentiment_category) %>%
  summarise(count = n(), .groups = "drop")


sentiment_plot <- ggplot(sentiment_by_device, aes(x = sentiment_category, y = count, fill = sentiment_ca
  geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~ tweetSource) +
  labs(
    title = "Sentiment Analysis by Device",
    x = "Sentiment Category",
    y = "Count of Tweets"
  ) +
  theme_minimal() +
  scale_fill_manual(values = c("Positive" = "green", "Neutral" = "blue", "Negative" = "red"))


print(sentiment_plot)
```



The graph displays the sentiment analysis of tweets categorized by the device source. Tweets are grouped

into three sentiment categories—Positive, Neutral, and Negative. Their counts are compared across devices.

- Android: The majority of tweets are Neutral (blue), followed by Negative (red) and then Positive (green), with significantly higher tweet counts compared to other devices.
- Iphone: Negative tweets dominate slightly over Neutral tweets, with Positive tweets being the least frequent. The overall volume is lower than Android but still substantial.
- Others: Neutral tweets are the most frequent, followed by Negative and Positive. The total tweet count is considerably lower than Android and iPhone.
- Ipad: Negative tweets dominate, with almost no Neutral or Positive tweets.
- dlvr.it and IFTTT: These sources have very low tweet volumes, and most tweets are Neutral or Negative, with Positive tweets nearly absent.

#INSIGHTS

- Android has the highest volume of tweets, suggesting it might be the most popular device for tweeting. iPhone users have a relatively higher proportion of Negative tweets, indicating potential dissatisfaction or criticism from this group.Devices like iPad, dlvr.it, and IFTTT have minimal tweet activity, which might indicate limited user engagement or niche usage. Neutral tweets are consistently higher across most devices, implying a tendency for users to share non-opinionated or factual content.