SentimentAnalysis_Magallanes-Trongoy-Nava

Killy Magallanes

2024-12-13

Load Libraries

```r
library(tidyverse)
```

```
## Warning: package 'dplyr' was built under R version 4.4.2
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- 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
```

```r
library(lubridate)
library(tidytext)
```

```
## Warning: package 'tidytext' was built under R version 4.4.2
```

```r
library(ggplot2)
library(scales)
```

```
##
## Attaching package: 'scales'
##
## The following object is masked from 'package:purrr':
##
##     discard
##
## The following object is masked from 'package:readr':
##
##     col_factor
```

Load Data

```
# Load the dataset
tweets_df <- read.csv("TweetsDF.csv", stringsAsFactors = FALSE)

# Display the first few rows of the dataset
head(tweets_df)
```

```
##   X    screenName
## 1 1      whourj31
## 2 2       nnainot
## 3 3   febry_sri_M
## 4 4 telehuntwatch
## 5 5    Typing0824
## 6 6   niccijsmith
##
## 1          A soldier angry at the support fund consolation money for the bereaved family of the Itaew
## 2                                                                                        Nah this Itae
## 3
## 4 TRANSLATION :\nSeoul residents lay flowers at a makeshift memorial near the site of the crush in It
## 5  The Itaewon stampede incident really caught me off guard. Makes me notice how important it is to
## 6  "What to do about my child? What to do about my child?" Park Ga-young’s mother, Choi Seon-mi, sai
##               created
## 1 2022-10-30 23:59:43
## 2 2022-10-30 23:59:32
## 3 2022-10-30 23:59:31
## 4 2022-10-30 23:59:28
## 5 2022-10-30 23:59:20
## 6 2022-10-30 23:59:04
##                                                                            statusSource
## 1                <a href="https://www.fs-poster.com/" rel="nofollow">FS_Poster_App</a>
## 2 <a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>
## 3 <a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>
## 4                            <a href="https://ruprop.live" rel="nofollow">telehunt</a>
## 5 <a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>
## 6   <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
##      Created_At_Round tweetSource
## 1 2022-10-31 00:00:00      others
## 2 2022-10-31 00:00:00     android
## 3 2022-10-31 00:00:00     android
## 4 2022-10-31 00:00:00      others
## 5 2022-10-31 00:00:00     android
## 6 2022-10-31 00:00:00      iphone
```

Data Cleaning

```
# Remove duplicates
tweets_df <- tweets_df %>% distinct()

# Convert created column to date-time
tweets_df$created <- as.POSIXct(tweets_df$created, format="%Y-%m-%d %H:%M:%S")

# Remove rows with missing text
tweets_df <- tweets_df %>% filter(!is.na(text))
```

```r
# Display the cleaned dataset
glimpse(tweets_df)
```

```
## Rows: 58,086
## Columns: 7
## $ X                <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16~
## $ screenName       <chr> "whourj31", "nnainot", "febry_sri_M", "telehuntwatch"~
## $ text             <chr> "A soldier angry at the support fund consolation mone~
## $ created          <dttm> 2022-10-30 23:59:43, 2022-10-30 23:59:32, 2022-10-30~
## $ statusSource     <chr> "<a href=\"https://www.fs-poster.com/\" rel=\"nofollo~
## $ Created_At_Round <chr> "2022-10-31 00:00:00", "2022-10-31 00:00:00", "2022-1~
## $ tweetSource      <chr> "others", "android", "android", "others", "android", ~
```

Sentiment Analysis

```r
# Unnest tokens and perform sentiment analysis
sentiment_df <- tweets_df %>%
  unnest_tokens(word, text) %>%
  inner_join(get_sentiments("bing")) %>%
  count(screenName, sentiment) %>%
  spread(sentiment, n, fill = 0) %>%
  mutate(sentiment_score = positive - negative)
```

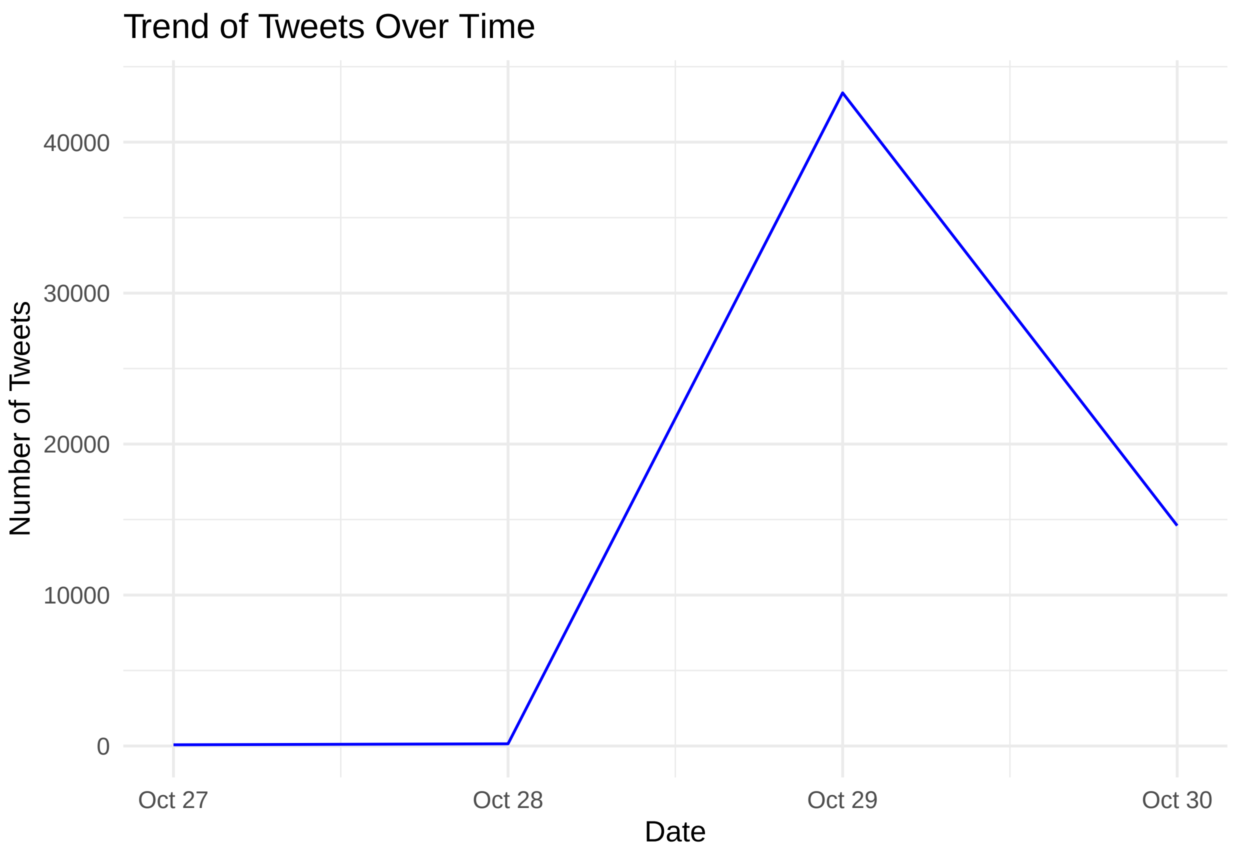
```
## Joining with ‘by = join_by(word)‘
```

```r
# Display sentiment scores
head(sentiment_df)
```

```
##        screenName negative positive sentiment_score
## 1        _____BMF        1        2               1
## 2 _____yourdreams        0        1               1
## 3         ____ljn        1        0              -1
## 4       ____nonfn        1        0              -1
## 5      ___00ff7f_        0        1               1
## 6   ___BLACKSWAN_        2        0              -2
```

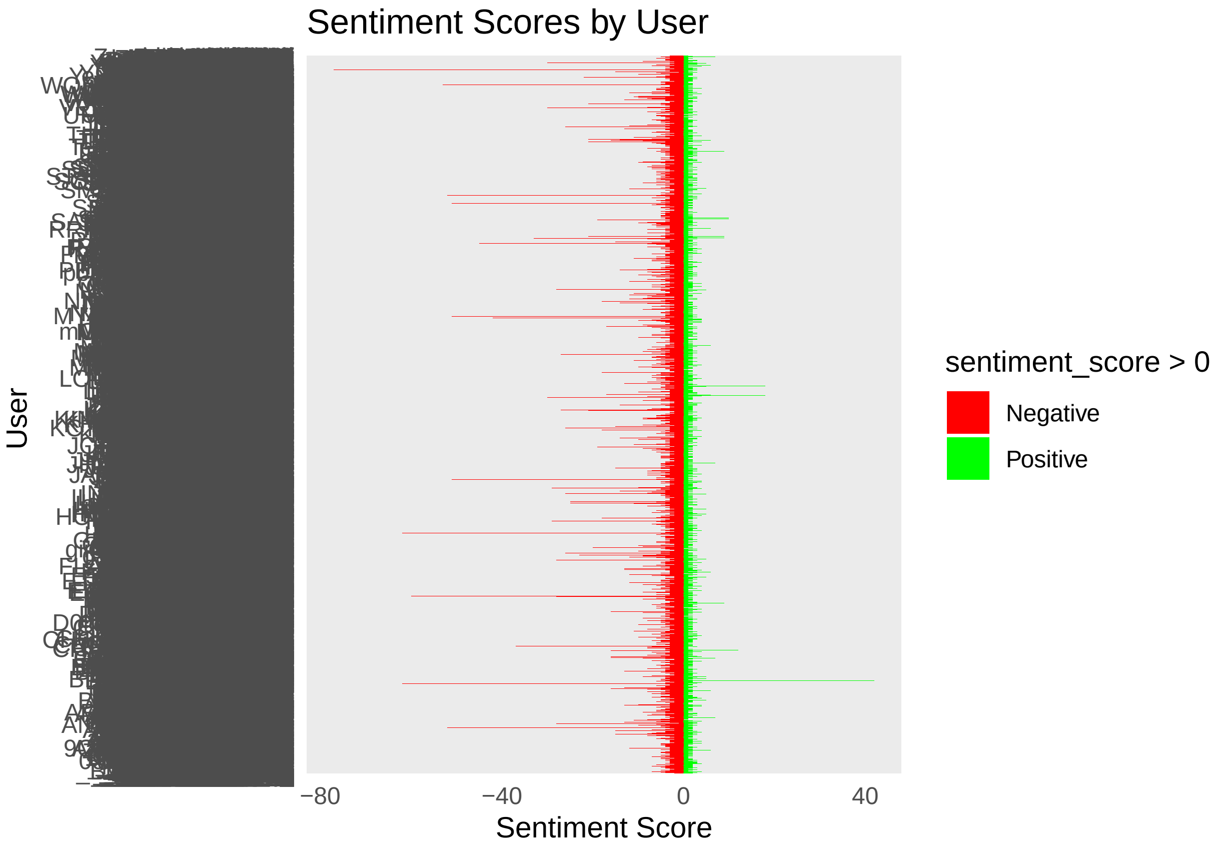
Trend analysis

```r
# Create a new column for date
tweets_df$date <- as.Date(tweets_df$created)

# Count tweets per day
daily_tweets <- tweets_df %>%
  group_by(date) %>%
  summarise(tweet_count = n())

# Plot the trend of tweets over time
ggplot(daily_tweets, aes(x = date, y = tweet_count)) +
  geom_line(color = "blue") +
  labs(title = "Trend of Tweets Over Time",
       x = "Date",
       y = "Number of Tweets") +
  theme_minimal()
```



Sentiment Score Validation

```r
# Plot sentiment scores
ggplot(sentiment_df, aes(x = screenName, y = sentiment_score, fill = sentiment_score > 0)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Sentiment Scores by User",
       x = "User ",
       y = "Sentiment Score") +
  scale_fill_manual(values = c("red", "green"), labels = c("Negative", "Positive")) +
  theme_minimal()
```



The sentiment score visualization shows the sentiment of tweets by user. Positive scores indicate a higher number of positive words in their tweets, while negative scores indicate the opposite. This can help identify which users are more positive or negative about a topic.

In this analysis, we performed sentiment analysis and trend analysis on a dataset of tweets. We cleaned the data, analyzed the sentiment of the tweets, and visualized trends over time. The insights gained can be useful for understanding public sentiment and engagement on social media platforms.