Analysis and Sentiment Analysis

Aguirre,Benajiba,Delatina

2024-12-14

Data Cleaning

```
library(dplyr)
```

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

```
##
## 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(stringr)
```

```
tweets_df <- read.csv("C:/Users/HP/Downloads/tweetsDF.csv")
```

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

```
missing_values <- colSums(is.na(tweets_df))
```

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

```
tweets_df$text <- tweets_df$text %>%
str_replace_all("http\\S+|www\\.\\S+", "") %>%
str_replace_all("[^[:alnum:][:space:]]", "") %>%
str_squish()
```

```
print(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 Itaewo
## 2                                                            Nah this Itaewon tragedy really has me
## 3                                                                                      JlN Pray for
## 4   TRANSLATION Seoul residents lay flowers at a makeshift memorial near the site of the crush in Ita
## 5 The Itaewon stampede incident really caught me off guard Makes me notice how important it is to kn
## 6         What to do about my child What to do about my child Park Gayoungs mother Choi Seonmi said
##               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
```

```
library(dplyr)
library(ggplot2)
```

```
## Warning: package ’ggplot2’ was built under R version 4.4.2
```

```
library(lubridate)
```

```
## Warning: package ’lubridate’ was built under R version 4.4.2
```

```
##
## Attaching package: ’lubridate’
```

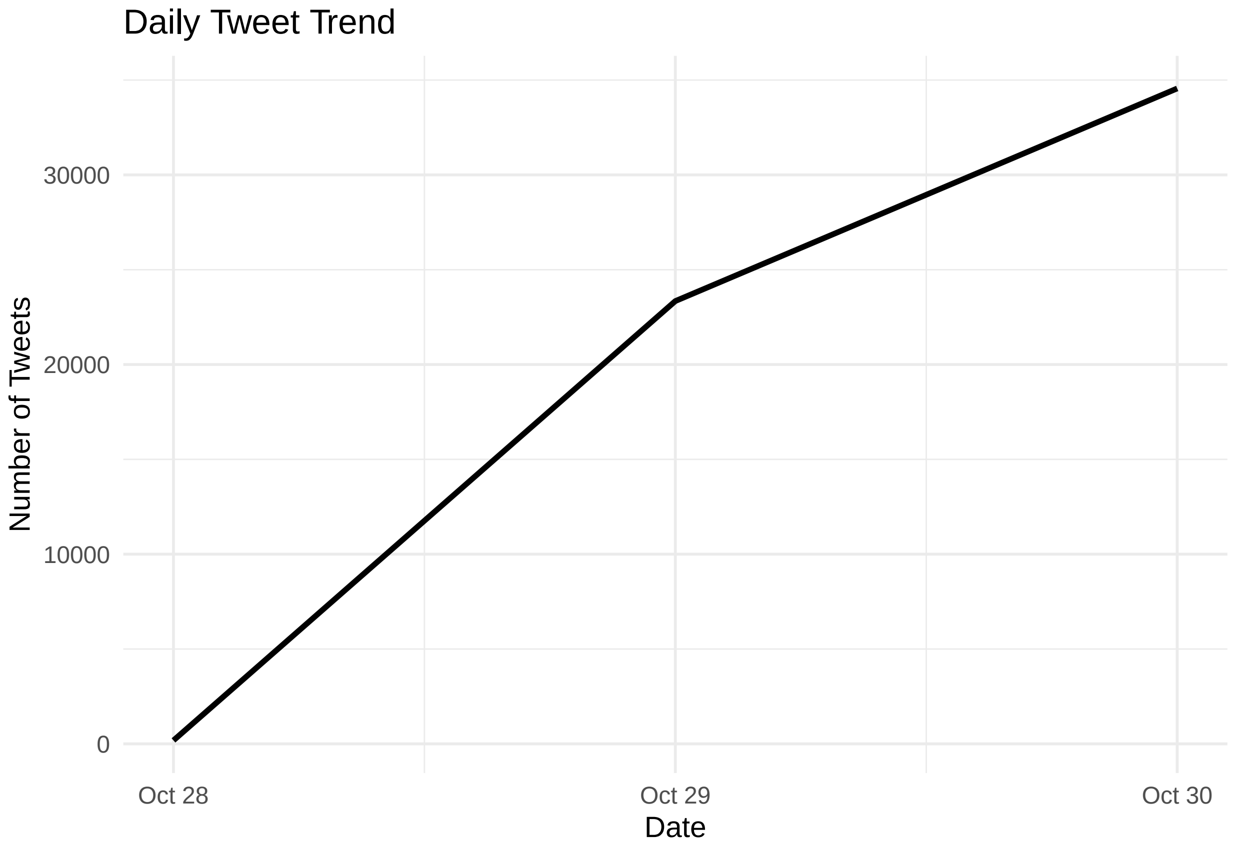
```
## The following objects are masked from ’package:base’:
##
##     date, intersect, setdiff, union
```

```
tweets_df$created <- ymd_hms(tweets_df$created)
```

```
daily_trend <- tweets_df %>%
mutate(date = as_date(created)) %>%
group_by(date) %>%
summarise(tweet_count = n())
```

```
ggplot(daily_trend, aes(x = date, y = tweet_count)) +
geom_line(color = "black", size = 1) +
labs(
title = "Daily Tweet Trend",
x = "Date",
y = "Number of Tweets"
) +
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.
```



Sentimental Analysis

```
library(dplyr)
library(tidytext)
```

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

```
library(ggplot2)
```

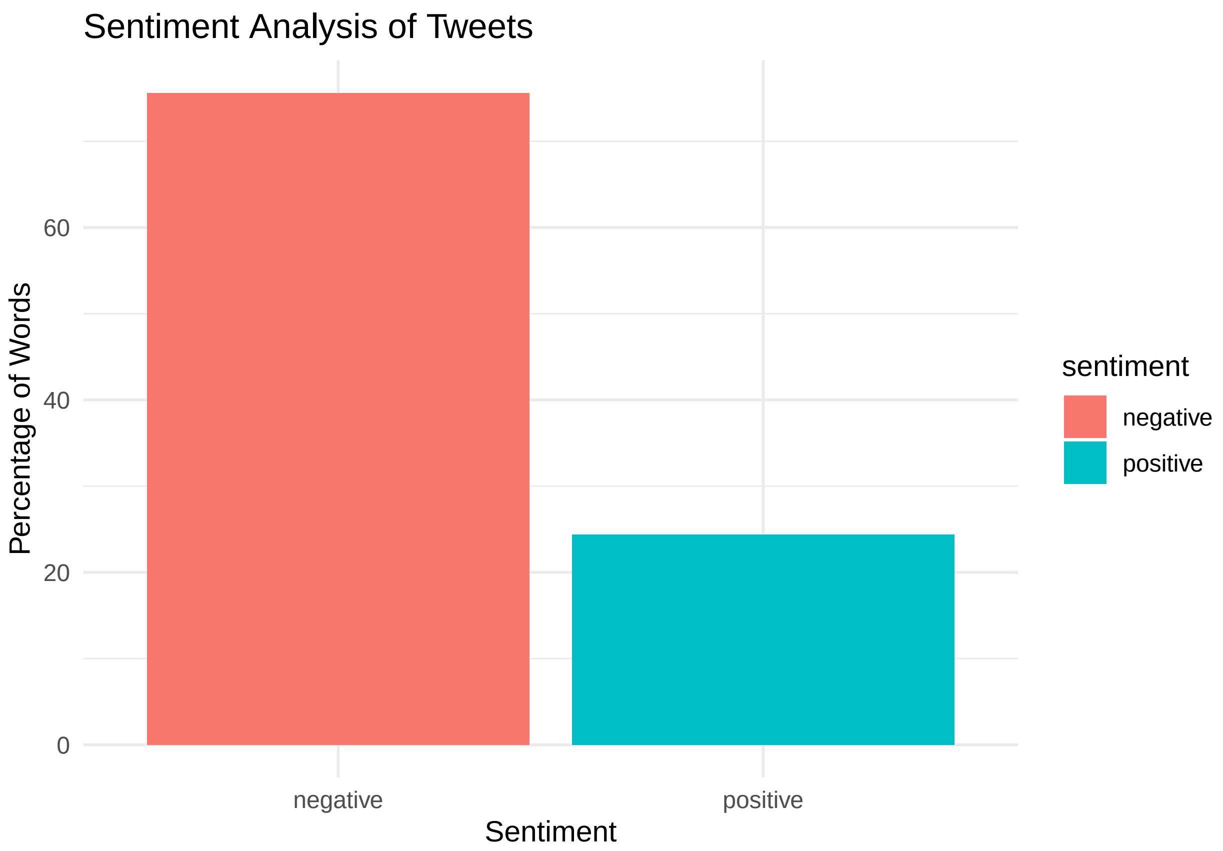
```
bing_lexicon <- get_sentiments("bing")
```

```
tokenized_tweets <- tweets_df %>%
unnest_tokens(word, text) %>%
anti_join(stop_words)
```

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

```
sentiment_analysis <- tokenized_tweets %>%
inner_join(bing_lexicon, by = "word") %>%
count(sentiment) %>%
mutate(percent = n / sum(n) * 100)
```

```
ggplot(sentiment_analysis, aes(x = sentiment, y = percent, fill = sentiment)) +
geom_bar(stat = "identity") +
labs(
title = "Sentiment Analysis of Tweets",
x = "Sentiment",
y = "Percentage of Words"
) +
theme_minimal()
```



2.Present your Use Case on what you will do with the dataset.

Use Case: Tracking Public Opinion on Social Issues
This project analyzes public sentiment and trends in 58,085 tweets to understand opinion shifts, engagement peaks, and emotional reactions to events. By examining tweet content, timestamps, and sources (e.g., Android/iOS), we can identify patterns over time.
The analysis benefits governments, brands, media, and social movements by providing insights into public reactions, helping refine strategies, and guiding actions. Key outcomes include trend visualizations, sentiment distribution (positive, neutral, negative), and insights linking sentiment changes to significant events. This data helps stakeholders respond effectively and predict future trends.

3.Make sure to describe each graph you have created. Give insights.

- Trend Analysis: Daily Tweet Count Graph Description: Type of Graph: Line graph X-axis: Date Y-axis: Number of Tweets Purpose: This graph shows the daily count of tweets over a specific period. We can visualize how tweet activity fluctuates over time, identifying spikes or dips that might correspond to certain events or key moments. Insights: Spike in Activity: Look for days where the tweet count drastically increases. This could indicate a major event, like a news break or a viral moment that caused a surge in engagement. Trends Over Time: Are there consistent periods of high activity (e.g., weekends, certain days of the week) or spikes tied to specific events? This helps understand the rhythm of public engagement. Overall Volume: If tweet counts remain consistent over time, it suggests steady engagement with the topic, while drastic fluctuations might indicate public interest following key moments.
- Sentiment Analysis: Sentiment Distribution (Positive, Negative, Neutral) Graph Description: Type of Graph: Bar chart X-axis: Sentiment categories (Positive, Negative, Neutral) Y-axis: Percentage of Words Purpose: This bar chart displays the proportion of words in the tweets categorized into three sentiments: positive, negative, and neutral. It helps us understand the emotional tone of the public

discourse. Insights: Dominant Sentiment: If one sentiment (e.g., negative) dominates, this suggests that the public is reacting strongly with negative emotions to the event or issue. Balance of Sentiment: If the chart shows a balanced distribution across positive, negative, and neutral, it indicates a nuanced or mixed reaction. Sentiment Shifts: Comparing sentiment across different time periods (e.g., before and after a major event) can reveal shifts in public opinion.