Sentiment-Analysis

Cadiz, Guion, Jacildo

2024-12-08

LOAD PACKAGES

```
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(tm)
```

```
## Loading required package: NLP
```

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

```
##
## Attaching package: 'lubridate'

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

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

DATA CLEANING

```
tweetsDF$text <- iconv(tweetsDF$text, from = "UTF-8", to = "ASCII//TRANSLIT", sub = "")
tweetsDF$text <- tolower(tweetsDF$text)
tweetsDF$text <- gsub("https\\S+", "", tweetsDF$text)
tweetsDF$text <- gsub("#", "", gsub("\n", " ", tweetsDF$text))
tweetsDF$text <- gsub("([@?]\\S+)", "", tweetsDF$text)
tweetsDF$text <- gsub("\\?", "", tweetsDF$text)
tweetsDF$text <- gsub("\\b\\d{2}\\.\\d{2}\\.\\d{4}\\b", "", tweetsDF$text)
tweetsDF$text <- gsub("<a href=httptwitter.comdownloadiphone rel=nofollow>twitter for iphone<a>", "", tw
tweetsDF$text <- gsub("<a href=([^>]*?) rel=nofollow>([^<]*?)<a>", "", tweetsDF$text)
tweetsDF$text <- gsub("<a href=httptwitter.comdownloadandroid rel=nofollow>twitter for android<a>", "",
tweetsDF$text <- gsub("<a href= rel=nofollow>twitter web app<a>", "", tweetsDF$text)
tweetsDF$text <- gsub("30102022", "", tweetsDF$text)
tweetsDF$text <- gsub("\\s+", " ", tweetsDF$text)
```

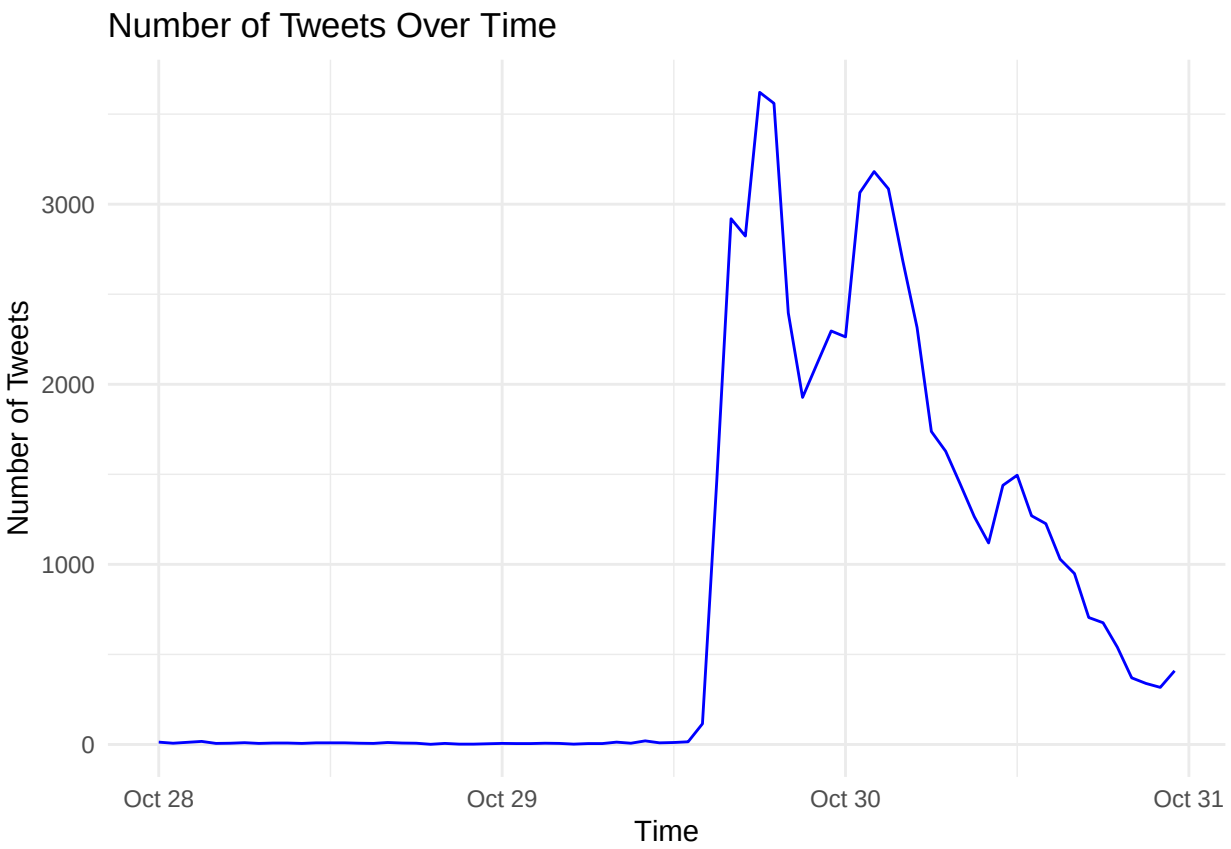
TREND ANALYSIS

For our trend analysis, we used the 'created' column and the number of tweets to generate a line graph. The 'created' column provides the date and time each tweet was posted. Our goal was to identify periods of peak activity and explore the potential reasons behind these trends. We also plan to search for articles online to investigate what might have caused the spikes in tweet activity after the tragedy.

```
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'

## The following object is masked from 'package:NLP':
##
##     annotate
```

```
library(lubridate)
library(dplyr)

file_path <- "tweetsDF.csv"
tweets_df <- read.csv(file_path)

tweets_df$created <- ymd_hms(tweets_df$created)

tweets_per_time <- tweets_df %>%
mutate(hour = floor_date(created, "hour")) %>%
count(hour)

ggplot(tweets_per_time, aes(x = hour, y = n)) +
geom_line(color = "blue") +
labs(title = "Number of Tweets Over Time", x = "Time", y = "Number of Tweets") +
theme_minimal()
```



The graph shows that tweets about Itaewon were few before midnight on October 29. A noticeable spike occurred after midnight, with the highest activity in the early hours of October 30. Activity levels decreased throughout the morning but rose again in the afternoon and evening.

Upon researching the timeline of events, we discovered that the Itaewon tragedy occurred at 10:30 PM on October 29. Reports indicate that the police confirmed the death toll at 2:30 AM, which likely contributed to the early morning peak in tweets.The rise in the evening may be due to more people becoming aware of and talking about the tragedy.

SENTIMENT ANALYSIS

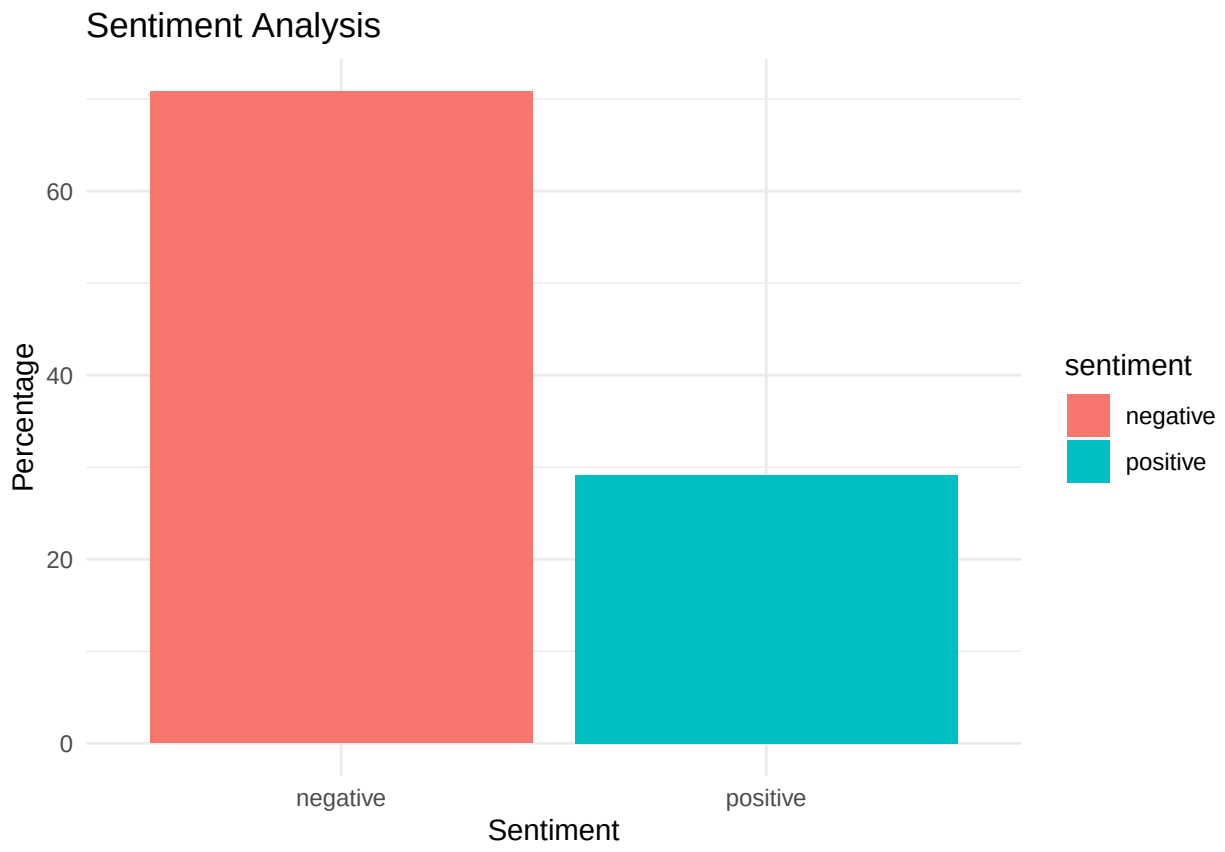
The dataset contains tweets about the Itaewon tragedy in 2022 under the 'text' column. We aim to analyze the sentiment of these tweets to understand public reactions. By categorizing sentiments as either positive or negative and visualizing them, we can determine the overall tone of the conversations and explore how people responded emotionally to the event.

```
library(tidytext)
library(dplyr)

sentiments <- get_sentiments("bing")

tweet_tokens <- tweetsDF %>%
unnest_tokens(word, text) %>%
inner_join(sentiments, by = "word")

sentiment_counts <- tweet_tokens %>%
```

```r
count(sentiment) %>%
mutate(percentage = n / sum(n) * 100)

print(sentiment_counts)
```

```
##   sentiment     n percentage
## 1  negative 49243    70.8298
## 2  positive 20280    29.1702
```

```r
library(ggplot2)
ggplot(sentiment_counts, aes(x = sentiment, y = percentage, fill = sentiment)) +
geom_bar(stat = "identity") +
labs(title = "Sentiment Analysis", x = "Sentiment", y = "Percentage") +
theme_minimal()
```

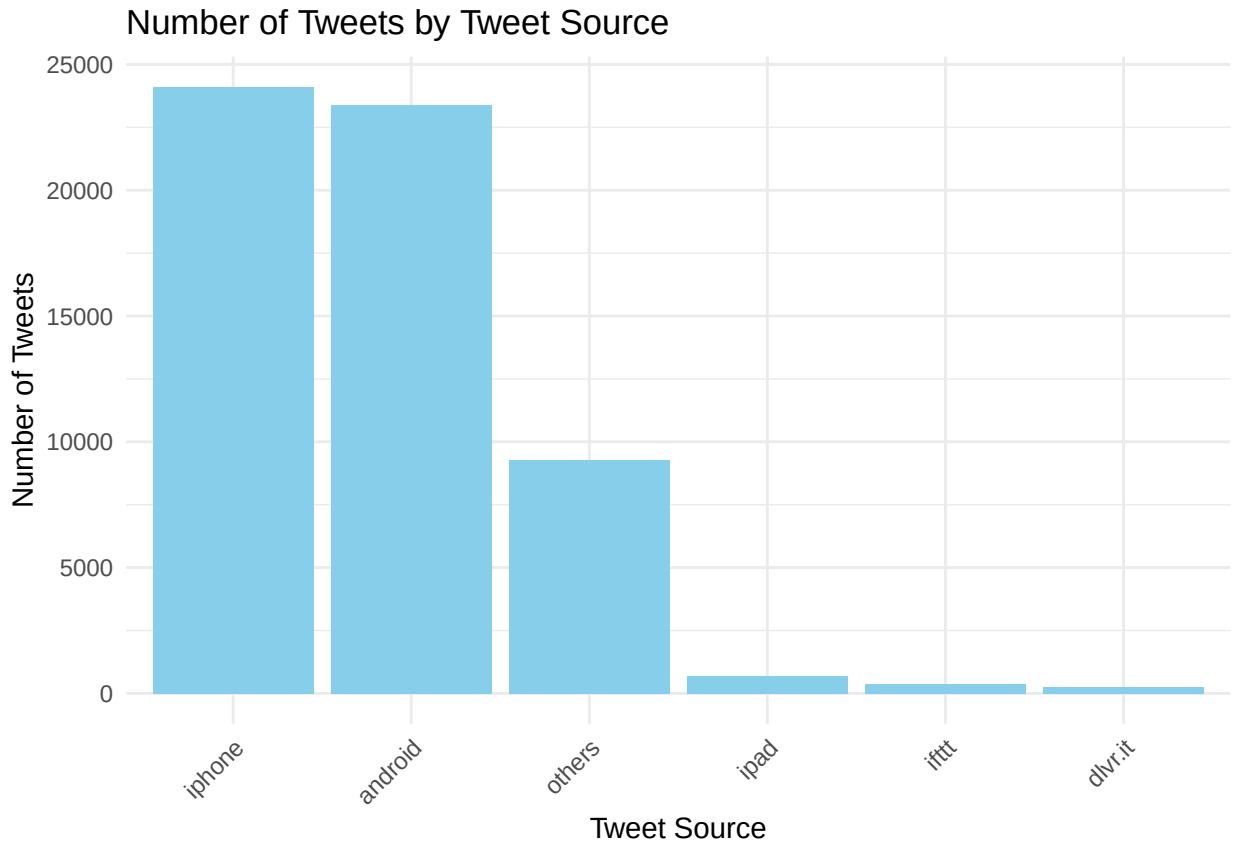


For our graph, the x-axis represents the sentiment categories (negative and positive), while the y-axis represents the percentage of tweets in each category. The bar graph indicates a higher percentage of negative tweets compared to positive ones. That highlights that the majority of tweets are negative. This indicates a strong emotional reaction to the event, possibly driven by distress, outrage, or grief.

ADDITIONAL GRAPH

The dataset contains tweets categorized by their source platforms, such as iPhone, Android, and others.For additional information, we wanted to identify which platforms users mostly used to engage with the content.

```r
tweet_source_analysis <- tweets_df %>%
group_by(tweetSource) %>%
summarise(Number_of_Tweets = n()) %>%
arrange(desc(Number_of_Tweets))


ggplot(tweet_source_analysis, aes(x = reorder(tweetSource, -Number_of_Tweets), y = Number_of_Tweets)) +
geom_bar(stat = "identity", fill = "skyblue") +
labs(
title = "Number of Tweets by Tweet Source",
x = "Tweet Source",
y = "Number of Tweets"
) +
theme_minimal() +
theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



The bar graph visualizes the number of tweets categorized by their source. The x-axis represents the tweet sources (e.g., iPhone, Android, others), while the y-axis shows the total number of tweets from each sourceThe bars indicate the volume of tweets contributed by each platform, with iPhone and Android having significantly higher counts than other sources.