sentiment analysis project

2024-12-14

```
#install.packages("ggplot2")
#install.packages("dplyr")
library(ggplot2)
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
```

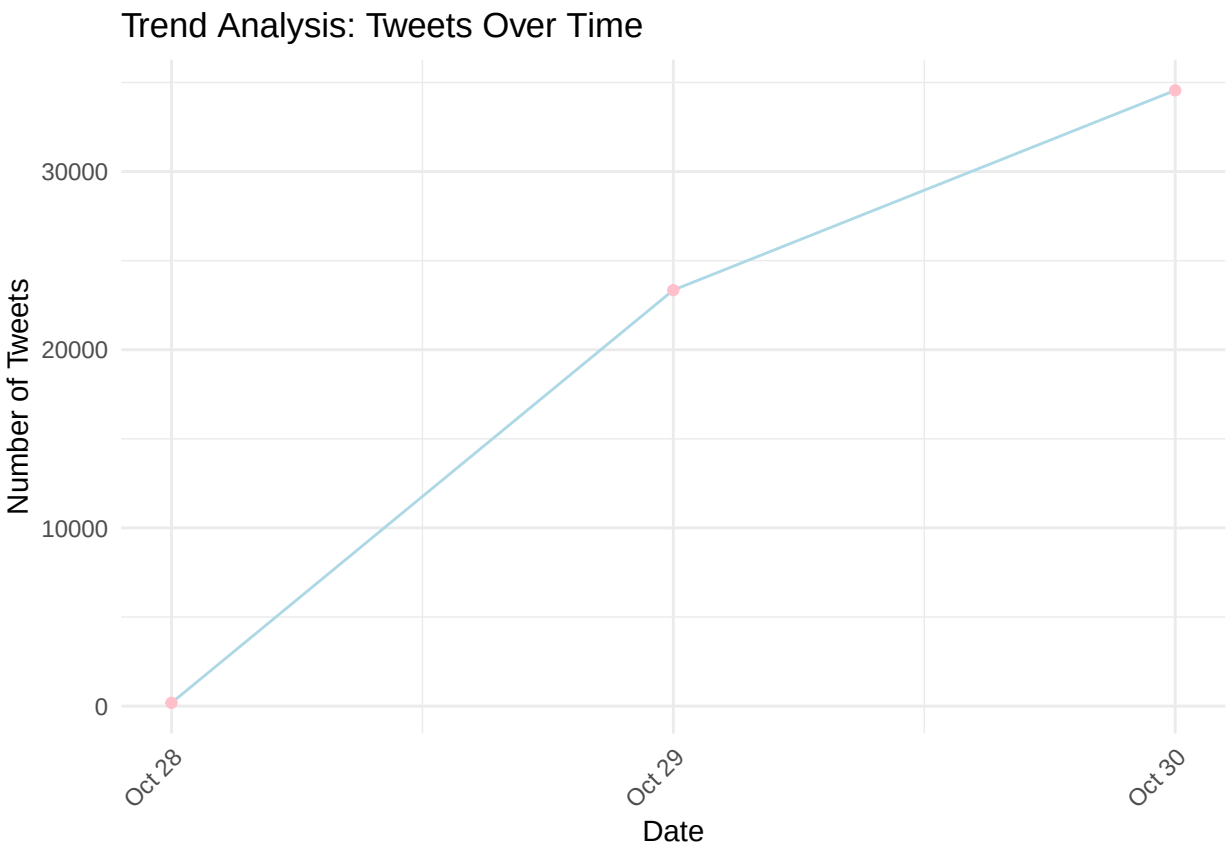
```
tweets_df <- read.csv("D:/Files/SentimentAnalysisProject/tweetsDF.csv")
```

Trend Analysis

```
tweets_df$created <- as.POSIXct(tweets_df$created, format = "%Y-%m-%d %H:%M:%S", tz = "UTC")
tweets_df$date <- as.Date(tweets_df$created)
```

```
tweetsPerDay <- tweets_df %>%
  group_by(date) %>%
  summarise(tweetCount = n())
```

```
ggplot(data = tweetsPerDay, aes(x = date, y = tweetCount)) +
  geom_line(color = "lightblue") +
  geom_point(color = "pink") +
  labs(title = "Trend Analysis: Tweets Over Time",
       x = "Date",
       y = "Number of Tweets") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Insights:

The graph labeled "Trend Analysis: Tweets Over Time" displays the amount of tweets on October 28, 29, and 30. The x-axis indicates the dates, while the y-axis shows the tweet volume, which ranges from 0 to 35,000. The data points are linked by a light blue line and highlighted in light pink. On October 28, tweet activity was low, but it increased dramatically to roughly 25,000 on October 29 and over 30,000 on October 30, showing increased involvement. The rapid increasing trend shows that the Itaewon incident on October 29 sparked heightened public attention and conversation, culminating the next day.

Sentiment Analysis

```
#install.packages("tidytext")
#install.packages("dplyr")
#install.packages("ggplot2")
#install.packages("stringr")
library(tidytext)
```

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

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

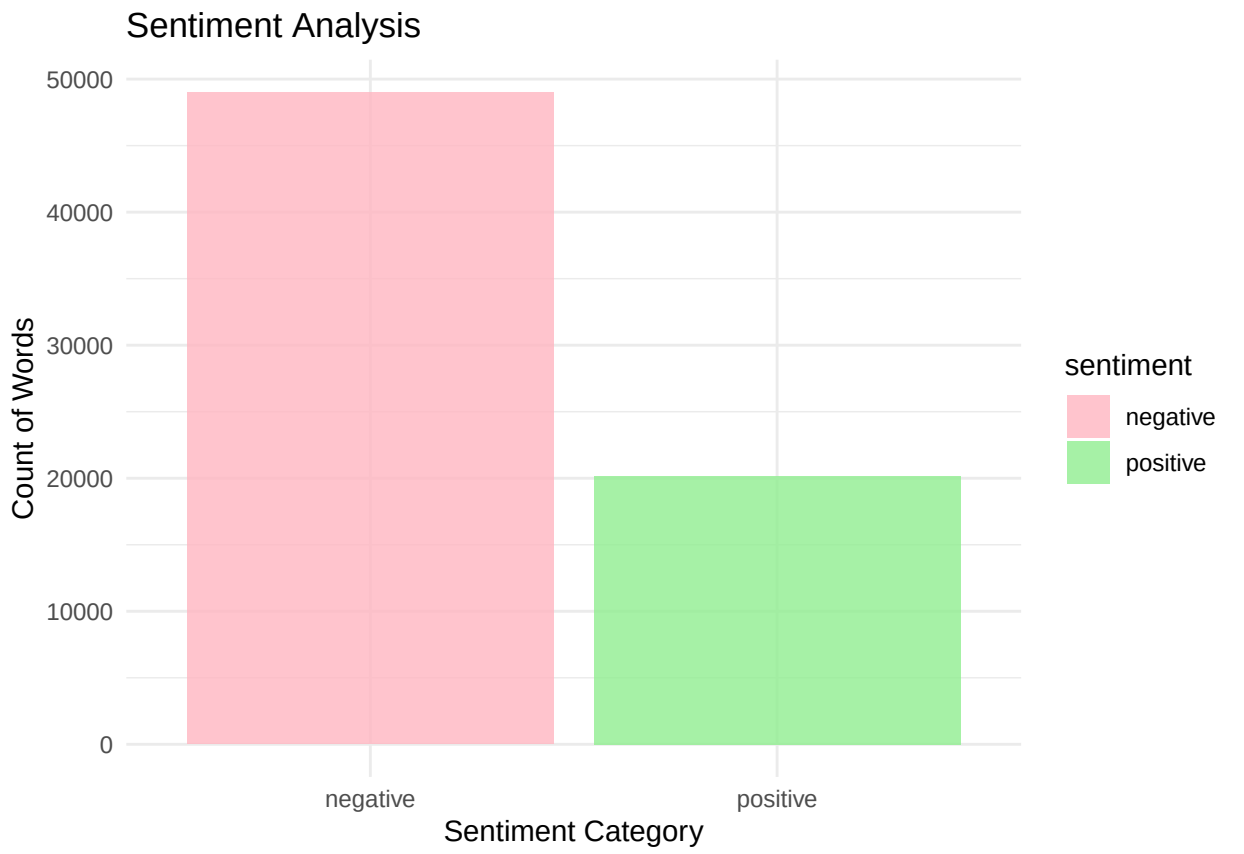
```
tweets_df$text <- tweets_df$text %>%
  str_to_lower() %>%
  str_replace_all("http\\S+|www\\S+", "") %>%
  str_replace_all("@\\w+", "") %>%
  str_replace_all("[^a-zA-Z\\s]", "")
```

```
tweetsWords <- tweets_df %>%
  unnest_tokens(word, text)
```

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

```
tweetsSentiments <- tweetsWords %>%
  inner_join(bingSentiments, by = "word") %>%
  count(sentiment) %>%
  mutate(sentiment = factor(sentiment, levels = c("negative", "positive")))
```

```
ggplot(data = tweetsSentiments, aes(x = sentiment, y = n, fill = sentiment)) +
  geom_bar(stat = "identity", alpha = 0.8) +
  scale_fill_manual(values = c("lightpink", "lightgreen")) +
  labs(title = "Sentiment Analysis",
       x = "Sentiment Category",
       y = "Count of Words") +
  theme_minimal()
```



Insights:

The graph depicts a Sentiment Analysis that compares two sentiment categories—negative and positive—based on the number of words. Negative emotion, indicated in pink, has a significantly higher word count, roughly 50,000 words, whereas good sentiment, represented in green, has a much lower word count, around 20,000 words. This suggests that the negative sentiment category dominates the analysis, with more than twice as many words as positive sentiment. This difference shows that the dataset of tweets about the Itaewon incident has a larger proportion of negative vocabulary, indicating a negative tone or attitude towards the tragedy.

Use Case:

This dataset mainly draws attention to the devastating Itaewon tragedy. Using this dataset, we can draw an analysis to identify at least three key aspects: first, several key public emotional trends before, during, plus after the incident; second, many community concerns or support, along with key discussion points, which require careful analysis to fully grasp their meaning; in addition to, actionable recommendations from policymakers, media outlets, plus event organizers.