Analyzing-Social-Media-Sentiment

Bhavesh P. Purohit

1. Introduction

I contributed to a project where we analyzed data from 1.6 million Twitter users, uncovering valuable insights by examining various patterns. Our methods involved text mining, sentiment analysis, probability analysis, constructing time series data, and employing hierarchical clustering on text and words to extract meaningful information from the dataset.

1.1 Data Description

1. The dataset *tweets.csv* comprises 1.6 million tweets, structured with six fields:
    - **Target:** Indicates the polarity of the tweet (0 = negative, 2 = neutral, 4 = positive)
    - **IDs:** Unique identification number assigned to each tweet (e.g., 2087)
    - **Date:** Timestamp indicating the date and time of the tweet (e.g., Sat May 16 23:58:44 UTC 2009)
    - **Flag:** Query associated with the tweet, if any (e.g., lyx). If absent, denoted as NO_QUERY.
    - **User:** Username of the Twitter user who posted the tweet (e.g., robotickilldozr)
    - **Text:** Actual text content of the tweet (e.g., Lyx is cool)
2. The dataset *daily-website-visitors.csv* encompasses five years of daily time series data, consisting of 2167 records across eight columns:
    - **Row:** Unique identifier for each record
    - **Day:** Day of the week represented in text format (e.g., Sunday, Monday)
    - **Day of Week:** Day of the week represented in numeric form (1-7)
    - **Date:** Date in mm/dd/yyyy format
    - **Page Loads:** Daily count of pages loaded
    - **Unique Visits:** Daily count of visitors whose IP addresses haven’t generated hits on any page in over 6 hours
    - **First Time Visits:** Number of unique visitors identified without a cookie as a previous customer
    - **Returning Visits:** Number of unique visitors excluding first-time visitors

1.2 Data Acquisition

We obtain both datasets from Kaggle:

The first dataset is sourced from https://www.kaggle.com/kazanova/sentiment140 (https://www.kaggle.com/kazanova/sentiment140).

The second dataset is sourced from https://www.kaggle.com/bobnau/daily-website-visitors (https://www.kaggle.com/bobnau/daily-website-visitors).

```
# Previewing few columns of Twitter user data set
tweets_preview <- tweetsDataRaw %>%
  select(date, text) %>%
  slice(1:5)  # Corrected slice index

kable(tweets_preview, caption = "Previewing few columns of Twitter user dataset")
```

Previewing few columns of Twitter user dataset

date	text
Mon Apr 06 22:19:45 PDT 2009	@switchfoot http://twitpic.com/2y1zl (http://twitpic.com/2y1zl) - Awww, that’s a bummer. You shoulda got David Carr of Third Day to do it. ;D
Mon Apr 06 22:19:49 PDT 2009	is upset that he can’t update his Facebook by texting it… and might cry as a result School today also. Blah!
Mon Apr 06 22:19:53 PDT 2009	@Kenichan I dived many times for the ball. Managed to save 50% The rest go out of bounds
Mon Apr 06 22:19:57 PDT 2009	my whole body feels itchy and like its on fire
Mon Apr 06 22:19:57 PDT 2009	@nationwideclass no, it’s not behaving at all. i’m mad. why am i here? because I can’t see you all over there.

```
# Reading the daily website visitors dataset
page <- read.csv('daily-website-visitors.csv', header = TRUE, sep = ',')

# Previewing few columns of Daily time series dataset
page_preview <- page %>%
  select(Row, Day, Date, Page.Loads, Unique.Visits) %>%
  slice(1:5)  # Corrected slice index

kable(page_preview, caption = "Previewing few columns of Daily time series dataset")
```

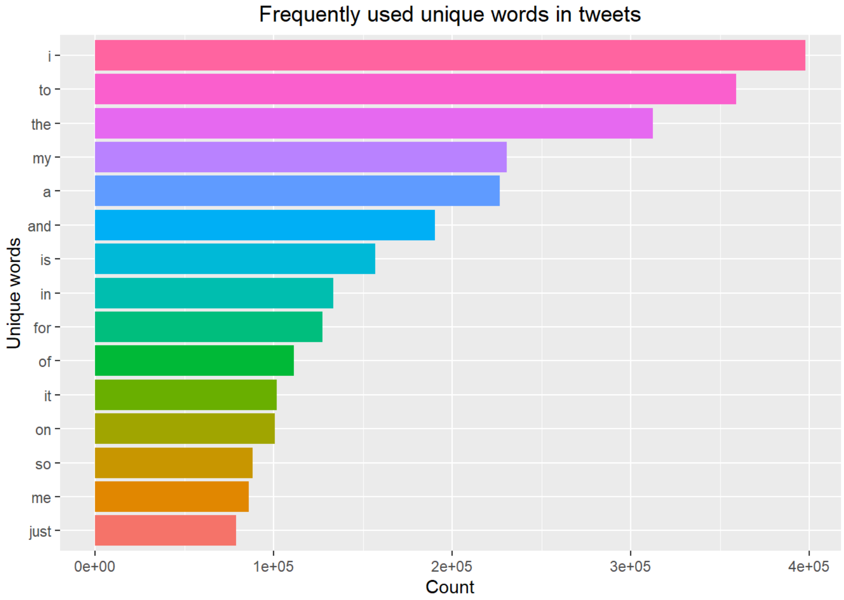
Previewing few columns of Daily time series dataset

Row	Day	Date	Page.Loads	Unique.Visits
1	Sunday	9/14/2014	2,146	1,582
2	Monday	9/15/2014	3,621	2,528
3	Tuesday	9/16/2014	3,698	2,630
4	Wednesday	9/17/2014	3,667	2,614
5	Thursday	9/18/2014	3,316	2,366

2.Analytical Questions

2.1 Text Mining

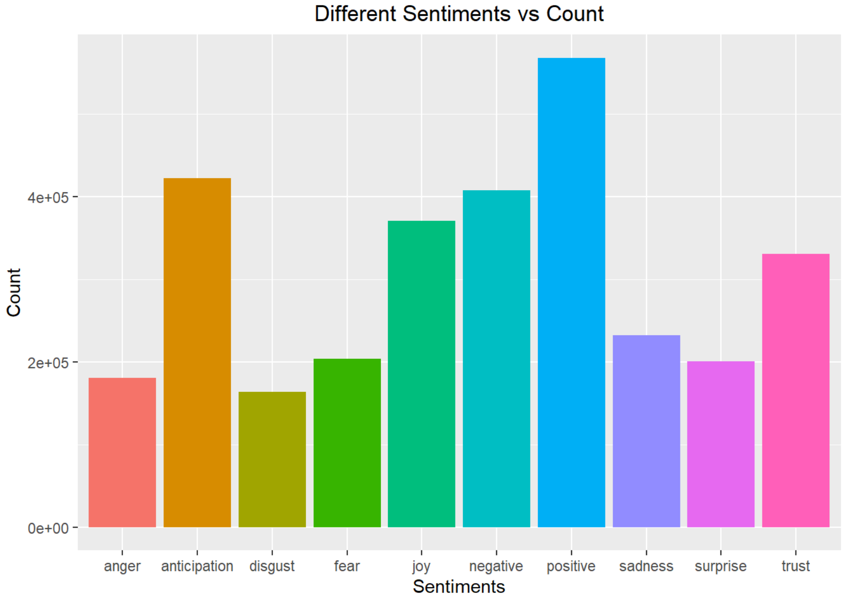
2.1.1 Finding the frequently used unique words



For this analysis, we focus solely on the unique ideas expressed by the users/authors. We eliminate stop words, user mentions, replies, and retweets to isolate the “original” tweets and present our findings visually.

Finding: The word *Day* stands out as the most commonly used term, appearing approximately 63,000 times among the 1.6 million tweets analyzed. Subsequently, words such as *Time*, *Home*, *love*, and *night* are also notable, each being utilized roughly 30,000 times.

2.1.2 Sentimental Trends of Tweets

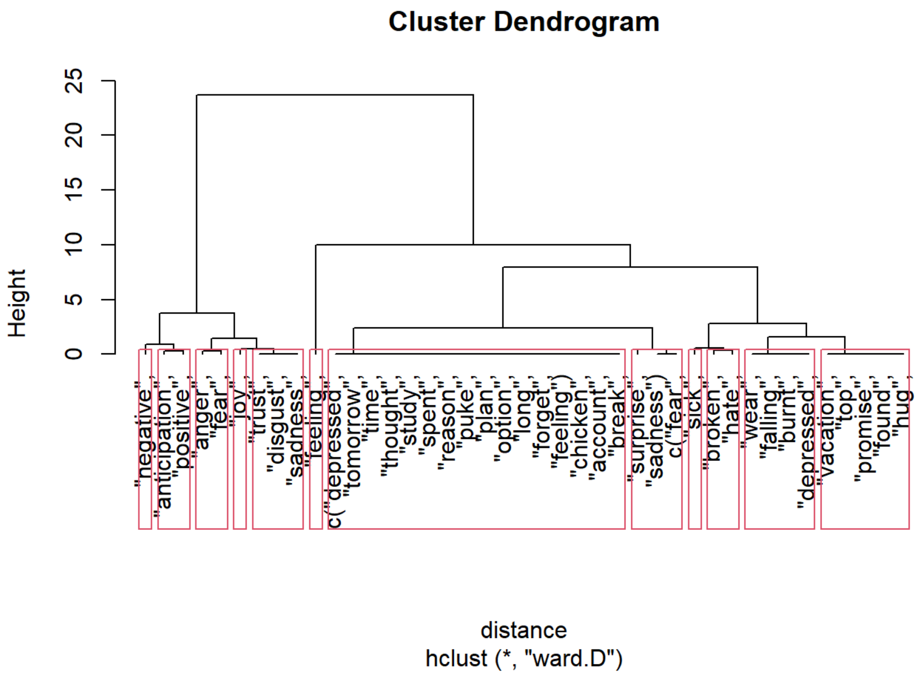


We utilize the NRC lexicon to identify various sentiments conveyed in each tweet and visualize their occurrences.

Finding: The most frequently tweeted sentiments include *Positive*, *Negative*, and *Anticipation*. Additionally, an interesting observation reveals an equal distribution of tweets expressing *Anger*, *Disgust*, and *Surprise*. Furthermore, a significant number of users have shared tweets concerning both *Fear* and *Trust* issues.

2.2 Clustering Analysis

Hierarchical clustering words by sentiments



As our dataset consists of textual data, we construct a corpus and apply the hierarchical clustering technique. This approach provides us with a dendrogram displaying various words clustered together based on sentiments. During the plotting process, a range of cluster numbers is suggested. After evaluating these ranges, we opt for 12 clusters as the most suitable choice.

Finding: The dendrogram presented above organizes our sample space into 12 distinct clusters, each categorized by sentiments. The height of the dendrogram indicates the distance between these clusters, providing insights into the clustering patterns based on sentiment similarities.

2.3 Probability

2.3.1. Calculating the PMF and CDF

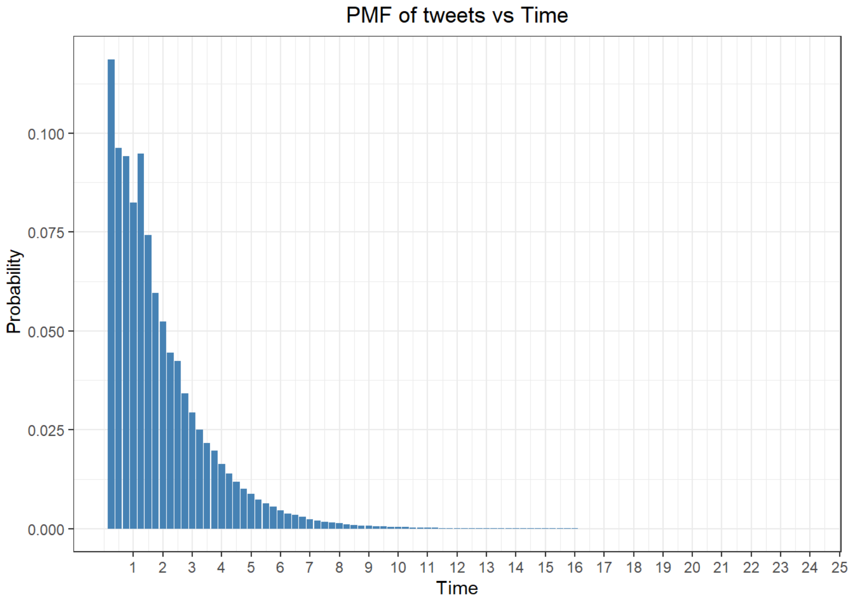
First 5 records of PMF of the tweet frequency.

pickup_pmf
0.1186604
0.0961854
0.0941705
0.0823568
0.0947725

First 5 records of CDF of the tweet frequency

pickup_cdf
0.1186604
0.2148458
0.3090163
0.3913731
0.4861457

2.3.2. Probability Mass Function over Time



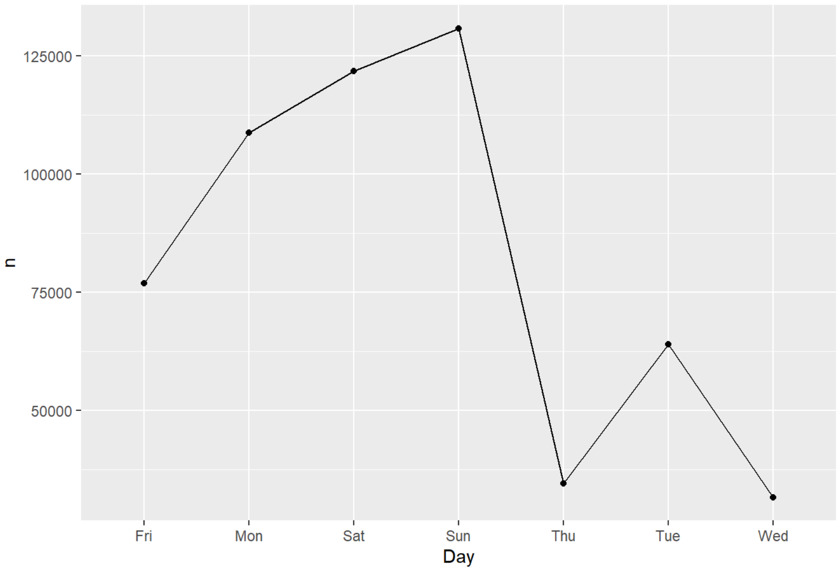
Finding: Over the specified period, there is an exponential decrease in the probability of tweets. Initially, the probability is highest at the beginning of the chosen time period, gradually diminishing as time progresses.

2.4 Time Series

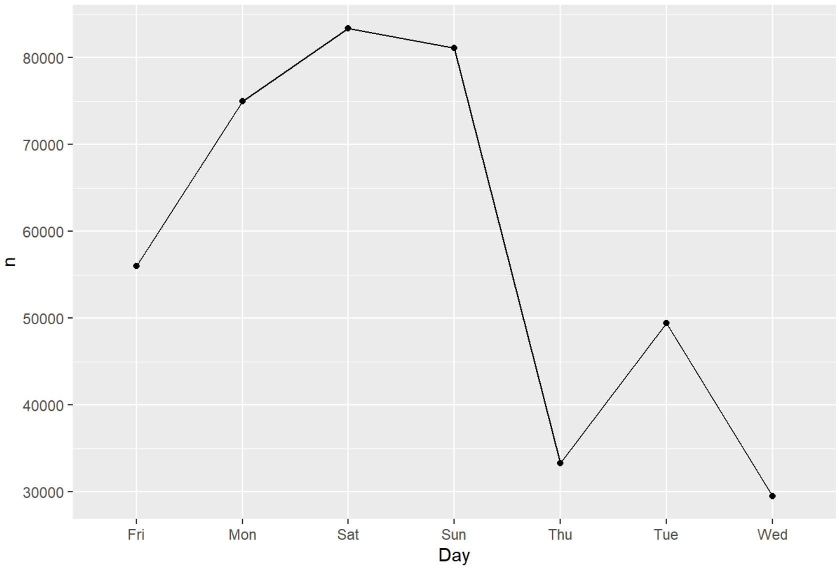
2.4.1. Trend analysis for different sentiments for each day of the week.

To extract all sentiments from the sentiments and date columns and determine the sentiments related to each day, we'll create visualizations to represent the counts of each sentiment. Here are the graphs for easier readability. **These graphs illustrate the distribution of sentiments over time, providing insights into the emotional trends observed throughout the analyzed period.**

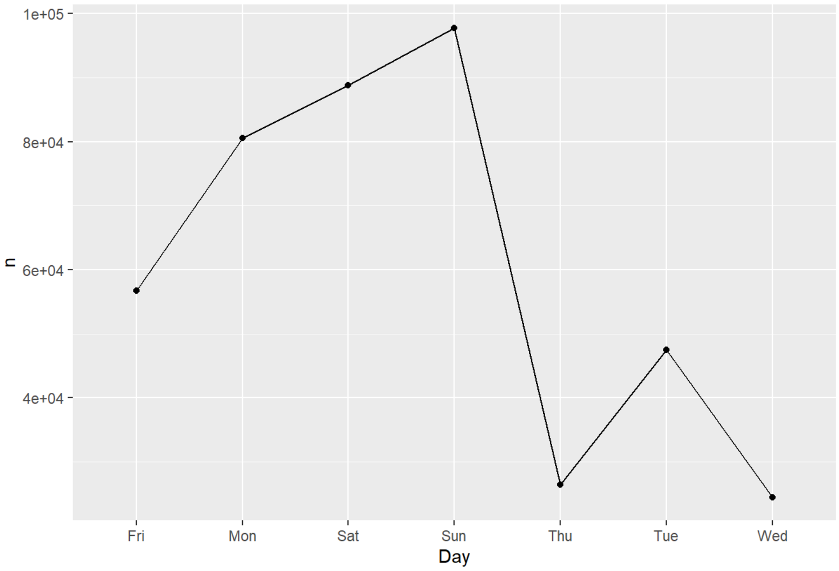
Positive Sentiment over the days

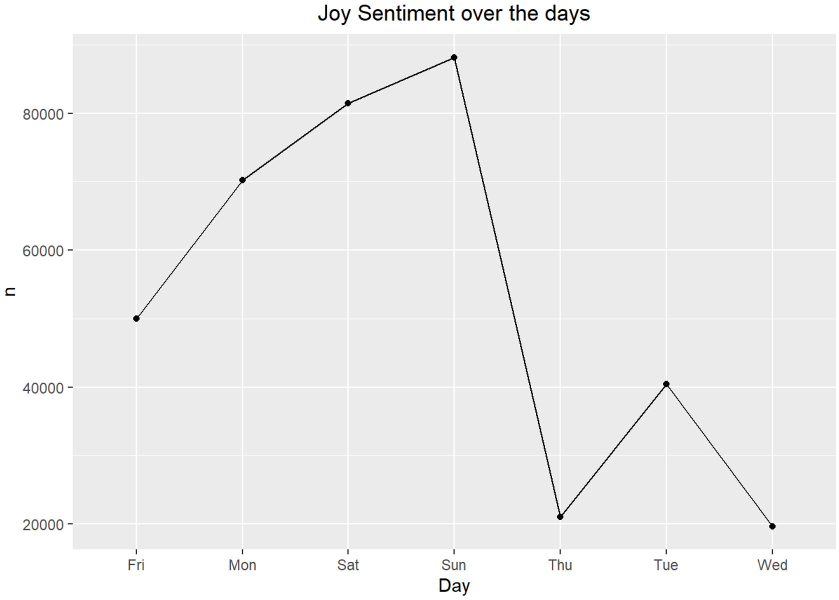


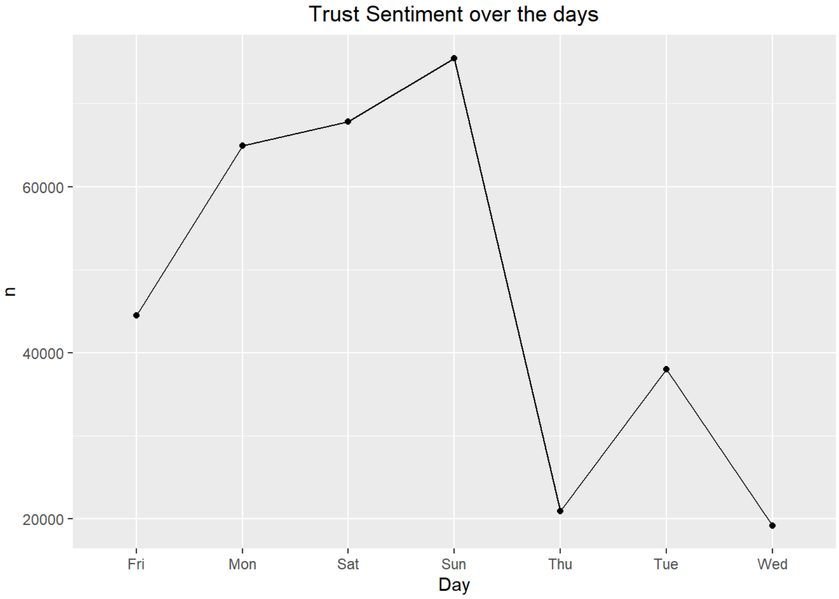
Negative Sentiment over the days



Anticipation Sentiment over the days



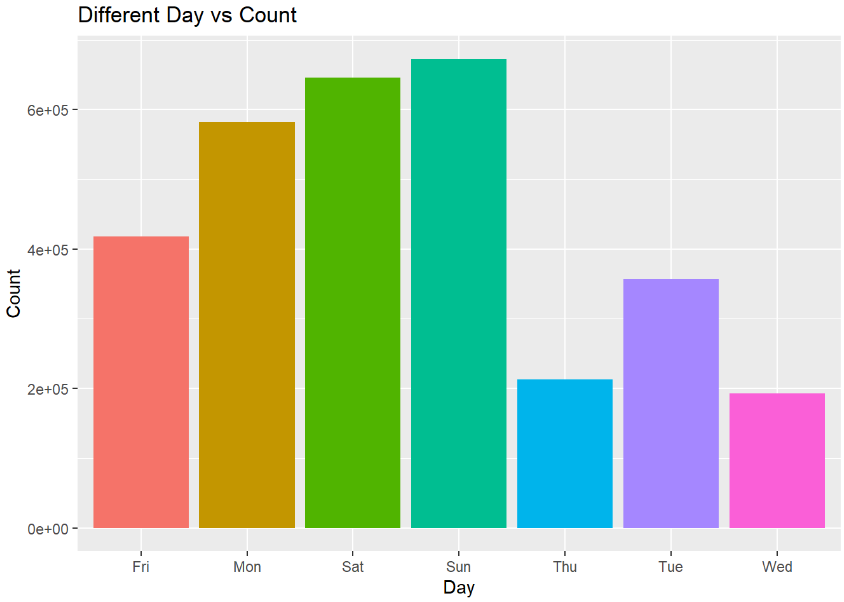




Finding: Across all the graphs provided, a notable trend is observed: the count of positive sentiments in tweets steadily increases until Sunday, followed by a sharp decline thereafter. Conversely, negative sentiments show a rising trend until Saturday, followed by a subsequent decrease. Similar patterns are observed across other sentiments depicted in the graphs, aligning with the observed behavior of positive sentiment trends.

2.4.1 Trend analysis looking at number of tweets per day of the

week



Finding: The top three days for tweeting are Saturday, Sunday, and Monday, aligning with the beginning of the weekend and the start of the workweek. Conversely, Wednesday and Thursday have the lowest number of tweets, likely due to their position in the middle of the week when individuals may be occupied with work or other responsibilities.

3. Summary

Following a meticulous analysis of 1.6 million pieces of Twitter data, we successfully decoded numerous emerging patterns and visualized them effectively. Through a combination of plots, text analysis/mining techniques, clustering methods, probability assessments, and time series data examination, we gained valuable insights into our business inquiries. This comprehensive approach allowed us to extract meaningful information and uncover actionable insights from the vast amount of Twitter data at our disposal.