

INFLUENCE OF TWEET FEATURES ON EMOTICON
USAGE

By

YOSHITHA ALAHARI

Bachelor of Science in Information Technology

Chaitanya Bharathi Institute of Technology

Hyderabad, Telangana

2017

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF SCIENCE
December 2019

INFLUENCE OF TWEET FEATURES ON EMOTICON USAGE

Thesis Approved:

Dr. K.M George

Thesis Adviser

Dr. Johnson P Thomas

Dr. Esra Akbas

ACKNOWLEDGEMENTS

I would first like to thank my thesis advisor Dr. George at Oklahoma state university for his advice in the right direction in every step of this research. The door is always open for me whenever I ran into an issue regarding my research or writing.

I would also like to thank my committee members Dr. Johnson P Thomas and Dr. Esra Akbas for their validation of this research. Without their inputs, the research wouldn't have been conducted successfully. Acknowledgements reflect the views of the author and are not endorsed by committee members or Oklahoma State University.

Name: YOSHITHA ALAHARI

Date of Degree: DECEMBER 2019

Title of Study: INFLUENCE OF TWEET FEATURES ON EMOTICON USAGE

Major Field: COMPUTER SCIENCE

Abstract:

Twitter, social media created 13 years ago is a platform that provides users to express their views regarding numerous issues. Reportedly, Twitter currently has 321 million users from all around the world and produces about 200 million tweets in a day. There is a large collection of data available that enables the performance of analysis to solve real-world issues. The usage of emoticons is rapidly growing as they provide higher impressibility and feasibility. In late 90's, emoticons were simple representations, like {":)" ":(""|" " : *"} etc. After Unicode consortium standardization, all the leading social media and mobile developers came up with improved versions of emoticons. This research involves analyzing the emoticon usage patterns in topics to identify the influence of multiple features such as age and location on the utilization of emoticons. Correlation of tweets and emoticon count is evaluated.

TABLE OF CONTENTS

| Chapter | Page |
|--|------|
| I. INTRODUCTION..... | 1 |
| II. REVIEW OF LITERATURE | 3 |
| III. METHODOLOGY | 8 |
| 3.1 Data Collection..... | 8 |
| 3.1.1 Apache Flume | 8 |
| 3.1.2 Twitter API..... | 9 |
| 3.1.3 Data Sets | 9 |
| 3.1.4 Tweet Structure | 11 |
| 3.1.5 Data Storage..... | 12 |
| 3.1.6 Data Preprocessing..... | 13 |
| 3.1.6.1 Tokenization..... | 13 |
| 3.1.6.2 Elimination of Tweets without Emoticons | 13 |
| 3.1.6.3 Noise Elimination | 14 |
| 3.1.7 Classification of Emoticons..... | 14 |
| 3.1.8 Categorization of Features..... | 14 |
| 3.1.8.1 Location..... | 15 |
| 3.1.8.2 Age..... | 15 |
| IV. FINDINGS | 16 |
| 4.1 Analysis performed..... | 16 |
| 4.1.1 Comparison of Emoticon Classes | 16 |
| 4.1.2 Impact of Week Days on Emoticon Usage | 17 |
| 4.1.3 Use of emoticon classes by Age group | 19 |
| 4.1.4 Analysis based on Topic and Location Features | 22 |
| 4.1.5 Trend Analysis | 25 |
| 4.1.5.1 Trend of emoticons count | 28 |
| 4.1.6 Correlation between tweet count and emoticon count | 32 |
| V. CONCLUSION..... | 34 |
| REFERENCES..... | 36 |
| EXTERNAL LINKS..... | 38 |
| APPENDICES..... | 39 |

LIST OF TABLES

| Table | Page |
|---|------|
| 3.1 List of Datasets..... | 11 |
| 3.2 Classification of Emoticons..... | 14 |
| 4.1 R squared values of tweets count and emoticon count for each dataset | 31 |
| 4.2 Trend values for each dataset | 32 |
| 4.3 Correlation coefficient values of tweets and emoticons count for each dataset.. | 32 |

LIST OF FIGURES

| Figure | Page |
|--|------|
| 3.1 Keywords of Food Dataset..... | 10 |
| 3.2 Structure of a Tweet | 11 |
| 3.3 Location Feature extracted from a Tweet | 14 |
| 4.1 Total count of Emoticons in each dataset | 16 |
| 4.2 Total count of Happy Emoticons | 17 |
| 4.3 Total count of Sad Emoticons | 18 |
| 4.4 Total count of Angry Emoticons..... | 18 |
| 4.5 Total count of Celebrations Emoticons..... | 19 |
| 4.6 Total count of Emoticons based on Age | 20 |
| 4.7 Distribution of emoticon classes for age group 18-25 | 20 |
| 4.8 Distribution of emoticon classes for age group 25-35 | 21 |
| 4.9 Distribution of emoticon classes for age group 35-60 | 21 |
| 4.10 Distribution of emoticon classes for age group 60+ | 22 |
| 4.11 Distribution of emoticon classes in Trump dataset | 22 |
| 4.12 Distribution of emoticon classes in Immigration dataset | 23 |
| 4.13 Distribution of emoticon classes in Food dataset | 23 |
| 4.14 Distribution of emoticon classes in Bitcoin dataset | 24 |
| 4.15 Distribution of emoticon classes in Pulwama dataset | 24 |
| 4.16 Distribution of emoticon classes based on Location | 25 |
| 4.17 Day based-Time series of emoticon and tweet count of Trump dataset | 26 |
| 4.18 Day based-Time series of emoticon and tweet count of Immigration dataset .. | 26 |
| 4.19 Day based-Time series of emoticon and tweet count of Food dataset | 27 |
| 4.20 Day based-Time series of emoticon and tweet count of Bitcoin dataset | 27 |
| 4.21 Day based-Time series of emoticon and tweet count of Pulwama dataset | 28 |
| 4.22 Trend of emoticons count of Trump data. | 29 |
| 4.23 Trend of emoticons count of Immigration data | 29 |
| 4.24 Trend of emoticons count of Food data | 30 |
| 4.25 Trend of emoticons count of Bitcoin data | 30 |
| 4.26 Trend of emoticons count of Pulwama data | 31 |
| 5.1 Office of National statistics results for happiest age group | 35 |

CHAPTER I

INTRODUCTION

Back in history when people tried to communicate through writings, they never used alphabets but used pictorial representations of what they were trying to express. Each pictorial symbol represents a word. Emoticons are millennial's pictorial symbols used to express their emotions [3]. The utilization of emoticons adds sentiment value to the text. They either emphasize the emotion expressed by the text or add a positive or negative sentiment. The usage of emoticons is rapidly growing as they are more expressive and faster [2]. In late 90's, when emojis came into picture they were very simple representations like {":)" ":("";" ": *"} etc. After Unicode consortium standardization, all the leading social media and mobile developers came up with their improved versions of emoticons. Currently, emojis are known as emoticons which express not only emotions but also represent places, items, animals, etc. [3]

Social media is one of the widely used platforms for people to share their opinions on political issues, product reviews, movie reviews, etc. People speak their mind irrespective of consequences. This platform has its negatives and positives. The platform is also widely used for feedbacks. Businesses collect and analyze social media data to improve their performance. In-text messaging, there is a restriction placed upon the number of characters which led to the popularity of emoticons usage. Ninety-five percent of the users end up using emoticons at some point in life. Around 10 billion emoticons are used in a day on different platforms. Most of the emoticons used on Twitter are positive (75 percent positive and 25 percent negative). 2.5 quintillion bytes of data are generated every day in the world from various platforms available to users like Twitter, [6] Facebook,

Instagram, Blogs, etc. Twitter is one of the rapidly growing social media platforms for users to post their opinions. With a character limit of 140 per tweet, emoticons can be used to express a user's point of view more strongly. The usage of emoticons in a tweet emphasizes their sentiment value i.e. they can either make a text positive or make it negative. When the sentiment of the text is neutral with the help of emoticons, we can determine the emotion of the tweet.

E.g. It's raining outside ☺

It's raining outside "

In both these tweets, the text is the same but with an addition of emoticon the entire sentiment of the tweet changes. Emoticons are more useful and faster in sentimental analysis, as by computing the sentiment of emoticons we can ignore the text or keywords. But there are some cases where just by looking at an emoticon we correctly cannot determine the sentiment. The best approach for sentimental analysis would be to analyze text with emoticons.

CHAPTER II

REVIEW OF LITERATURE

In this section, we discuss the work conducted which aligns with our emoticon analysis. Sentimental analysis on twitter data is widely used which has started in the 2000s where priority was initially given to text in the tweets to understand the tweet context. Later work progressed more on considering other factors like emoticons which are influencing the tweet context. More work was conducted on this part where analysis is done using emoticons and text + emoticons.

Emoticons have come into play in the late '90s but have recently grown in terms of the number of emoticons as well as usage. Human behavior can be predicted with the analysis of emoticon usage [ref.4]. We can apply various methodologies used in different papers alongside emoticons to study the impact of various factors.

Hao Wang and Jorge A. Castanon [1] have developed an algorithm that determines in which context the emoticon is used appropriately. They performed four different types of analyses to analyze the context in which the emoticons are used and to determine the correlation between sentiment polarity and emoticons. They classified their analysis into 4 techniques i.e. in the first one they grouped certain users and acquired the most frequent emoticons used by them and calculated their sentiment polarity, In their second analysis they performed clustering of words and emoticons to understand the emotion expressed by the emoticons, In their third analysis, they performed comparison of text with and without emoticons and their final analysis they used two training models to analyze the sentiment of the text along with emoticons and without emoticons.

Georgios S. Solakidis, Konstantinos N. Vavliakis and Pericles A. Mitkas [2] have adopted a semi-supervised technique for emoticon detection. They have classified their approach using two feature vectors and 4 different sentiments. The two feature vectors used by them are 1) Classification of emoticons as positive, negative and undefined 2) 30 different fervent positive and negative words. The four different sentiments they focused upon are Love, Anger, Joy, and Sadness. They further categorized the feature vectors into unigrams, bigrams, unigrams with bigrams, trigrams, etc. The data used for their analysis are mostly Greek documents. The results obtained by them are mean accuracy of 90 percent for subjectivity classification and 93 percent for polarity classification. During their analysis, they calculated the sentiment score for text and emoticons separately.

Xuan Lu, Wei Ai, Xuanzhe Liu, Qian Li, Ning Wang, Gang Huang and Qiaozhu Mei [3] stated that factors like location and culture have an effect on emoticon usage. To perform analysis, they used data worth a month from 212 different countries from 3.88 million active users. They used the Kika Emoticon keyboard and analyzed all the regions where it is used by users. They further classified the type of emoticon users and emoticons used by them. They focused on factors like gender, age, religion and relationship status. Future research is being done on emoticons based on the context in different regions.

Peijun Zhao, Jia Jia, Yongsheng An, Jie Liang, Lexing Xie and Jiebo Luo [4] performed analysis of twitter data using multiple features like tweet content, tweet structure, and user demographics. They propose "a mmGRU model for predicting emoji categories and positions ". motivated by the observations. It performs an in-depth analysis of classes of emoticons and their positions in the text.

With the help of this model, there is a rise in the accuracy of prediction of classes of emoticons by +9% and their positions by +4%.

Marco Vicente, Joao P. Carvalho, Fernando Batista [5] focused on the factors like frequency, sentiment, and location of emoticons over the data to analyze the emoticons. They have a dataset of 1.6 million tweets which consists of 13 different European languages. They have built a sentiment lexicon map which gives the emoticons a sentiment score of -1 which means negative, 0 neutral and +1 positive. Along with this, to improve the accuracy, entropy and frequency are the considered features where initial screening is done for the users who have a very high ratio of tweets. They have also come up with the new analysis apart from the traditional analysis were analyzing the use of emoticons from a group of people to how an individual is interpreting and understanding the emoticon.

Marco Vicente, Joao P. Carvalho and Fernando Batista [5] emphasized the twitter data analysis using emoticons and on how emoticons can build the model of trust, which helps in knowing any fraudulent or scam activities. They have used user demographics, age, and gender factors to analyze the anonymity of users. For example, a user profile with a male name when posts anything using lipstick emoticons can be considered suspicious. Users will try to be more cautious about fake or spam accounts. Factors like age and gender will help us to understand user credibility. Among age groups, the usage of the maximum emoticon falls in the range of 20 -30. They have included feature extraction by using vector representation for each user by providing a trusted user list to acquire accuracy.

Itisha Gupta and Nisheeth Joshi [6] used various features that can be extracted from the user profile to detect the gender of the user. Two datasets were used to perform analysis i.e. English and

Portuguese. They implemented supervised and unsupervised approaches (Naïve Bayes, Logistic Regression, Support Vector Machines, Fuzzy c-Means clustering, and k-means) to determine the gender. They have acquired high accuracy results of 97.9 percentile for supervised and for the clustering techniques 96.4 percentile.

Brandon Lwowski, Paul Rad and Kim-Kwang Rayong Choo [7] described various preprocessing techniques used to clean data. Twitter data consist of a lot of noise in the form of misspelled words, URLs, hashtags, etc. Due to the presence of such words, there is a lot of noise in the results. To get a more accurate output we need to eliminate them. The authors used two elimination methods namely removal of stop words, punctuations, etc. and normalization of words. During the preprocessing of tweets, normalization is performed on abbreviations i.e. they are converted to their full forms. A comparison of manually preprocessed tweets to automatic preprocessed tweets was performed which resulted in an accuracy of 87.6 percent for proposed preprocessed.

Sharath Chandra Guntuku, Ming yang Li, Louis Tay, and Lyle H. Ungar [8] performed analysis based on the geographical location of users. The authors collected tweets based on location using twitter API and detected events based on geospatial emotion vector. They used a combination of graph theory, statistics and machine learning semantics.

Comparison of different factors like frequency, context and topic association across multiple countries on emoticon usage is observed [9]. Data are collected from two major parts of the world i.e. East (China & Japan) and West (US, UK & Canada). They performed a three-step analysis, first differences and similarities are analyzed based on different factors like people, food, drink, etc. In the second step mapping of emoticons was done based on previous features used to identify cultural

differences. In the third step, Psycholinguistic classification using Ekman's emotions is implemented.

Andrea Trevino [10] aimed at the analysis of twitter data based on emoticons in their sentimental analysis. So, in the process of interpreting emoticons, they have come up with the typology in addition to the semantics. In this approach, the author used keyword spotting for emoticons where this helps using methods like sentiment expression, sentiment enhancement, and sentiment modification. In this method sentiment, polarity and emoticon polarity were detected, and comparison was made on emoticon polarity and sentiment polarity. This approach of using labels will help to train the model to recognize the sentence usage automatically.

Earlier research found in literature has been related to the sentiment analysis of an emoticon or popularity of emoticon or frequency of an emoticon. To our knowledge, research has never been performed with regards to the features which can affect the usage of an emoticon. This research focused on the effect of multiple features like Age, Location, and type of topic on the emoticon usage on Twitter.

The chapters that follow outline the methodology followed in this research, the analysis performed, and the findings are described.

CHAPTER III

METHODOLOGY

The following section primarily focuses on the various methods implemented for this research. This section is further divided into various subsections like Data Collection, Datasets, Emoticon, Features, etc.

3.1 Data Collection

Volume, Velocity, and Variety are characteristics of Big Data. Thus, Twitter data is classified as Big Data. One of the main issues with such a huge amount of data is stored (of structured & unstructured data). The processing of such data is also a concern. Analysis of Big Data will lead to better insights on the data which resolves real-world problems. Apache Flume is a tool that is used to collect such large amounts of data.

3.1.1 Apache Flume

Apache Flume acts as a bridge between data generators like Facebook, Twitter, Google, etc. and centralized data storage systems. Apache Flume is used to collect streaming data. Qualities like fault tolerance and failure recovery make Apache Flume a reliable tool for the collection of streaming data.

The architecture of Flume consists of 3 main components like channel, source, and sink. The source will collect the data from twitter and stores the data in the channel. Channel acts as a buffer to store the data for a short period to avoid data loss. The sink is used to write the data into HDFS.

Twitter API is used for data collection. To collect data from twitter, a twitter developer account is needed. Tweets are collected based upon the keywords specified in the configuration file which also consist of properties related to the source, sink, and channel. Tweets extracted from twitter are stored in Hadoop in JSON format.

3.1.2 Twitter API

It is used to provide users with programmatic access to Twitter. Each user is provided with a set of unique parameters like consumer key, consumer secret, access token and access token secret which need to be specified in the configuration file. Twitter API can access any information that is made public by the user.

3.1.3 Datasets

This research makes use of 5 different datasets on five different topics, namely Trump, Immigration, Bitcoin, Food, and Pulwama Attack. These datasets are collected using related keywords. Trump dataset consists of all the tweets and retweets which contain the keyword "Donald Trump". This dataset is for the duration of 31 days (i.e. July 2019). The size of the dataset is 45GB. Immigration dataset consists of all the tweets and retweets which contains the keywords "Immigration", "Child Separation", "Parent" and "Illegal Immigration". This dataset is for the duration of February 2019 and March 2019. The size of the dataset is 33GB for each month. Food dataset consists of all the tweets and retweets which contains the keywords (Figure 3.1) "Diarrhea", "Abdominal pain", "Vomiting", "Puke" and "Fever". This dataset is from March through April 2019. The size of the dataset is 17GB. Bitcoin dataset consists of all the tweets and retweets which contain the keyword "Bitcoin". This dataset is from April through May 2019. The size of this

dataset is 7.7GB and 18GB. Pulwama was a terrorist attack on the Indian army in the region Pulwama on February 14, 2019. The keywords used are "Pulwama", "Pulwama Attack", "Terrorist", and "Indian Army". The size of the dataset is 50 GB. Table 3.1 shows the datasets, keywords used, period of data collection and data size.

```
TwitterAgent.sources.Twitter.accessToken = 75396490895947592-h88dcX9qQwIB01964yck0y1601
TwitterAgent.sources.Twitter.accessTokenSecret = psYQPQFaYIubys8aIWrxHo4D4FktVIClrvutk
TwitterAgent.sources.Twitter.keywords = Diarrhea, abdominal pain, vomiting, puke, fever
TwitterAgent.sinks.HDFS.channel = MemChannel
```

Figure 3.1 Keywords of Food Dataset

| Dataset Topics | Keywords | Data collection Period | Data Size |
|--------------------|--|------------------------------|---------------------|
| Trump | "Donald Trump" | July 2019 – August 2019 | 45GB |
| Immigration | "Immigration", "Child Separation", "Parent" and "Illegal Immigration". | February 2019 and March 2019 | 33GB for each month |
| Food | "Diarrhea", "Abdominal pain", "Vomiting", "Puke" and "Fever". | March 2019 - April 2019 | 17GB |
| Bitcoin | "Bitcoin". | April 2019 -May 2019 | 7.7GB and 18GB |

| | | | |
|-----------------------|---|--------------------------|-------|
| Pulwama Attack | “Pulwama”, Pulwama Attack”, “Terrorist “and “Indian Army”. | Feb 2019 – March 2019 | 50 GB |
|-----------------------|---|--------------------------|-------|

Table 3.1 List of Datasets

3.1.4 Tweet Structure

During the extraction of tweets from twitter apart from the text of a tweet some additional features are also extracted. Some of the features which are extracted along with a tweet are retweet count, geo, favourite_count, reply_count, followers count, location, etc. The following figure illustrates the same. Figure 3.2 shows a tweet sample.

```
{
  "extended_tweet": {
    "entities": {
      "urls": [],
      "hashtags": [],
      "user_mentions": []
    },
    "symbols": [],
    "full_text": "Hal sederhana yang bikin cewe bahagia:\nSelalu kasih kabar\nDilembutin\nNgga ganien sama cewe lain\nSelalu cerita kalau ada sesuatu\nPerhatian\nDipedulin\n\nSimple)",
    "display_text_range": [0, 165],
    "in_reply_to_status_id_str": null,
    "in_reply_to_status_id": null,
    "created_at": "Mon Jul 08 14:49:51 +0000 2019",
    "in_reply_to_user_id_str": null,
    "source": "<a href='\"http://twitter.com/download/android\" rel='\"nofollow\">Twitter for Android</a>",
    "retweet_count": 0,
    "retweeted": false,
    "geo": null,
    "filter_level": "low",
    "in_reply_to_screen_name": null,
    "is_quote_status": false,
    "id_str": "1148242800183349253",
    "in_reply_to_user_id": null,
    "favorite_count": 0,
    "id": "1148242800183349253",
    "text": "Hal sederhana yang bikin cewe bahagia:\nSelalu kasih kabar\nDilembutin\nNgga ganien sama cewe lain\nSelalu cerita k\u2026",
    "https://t.co/48NIDkZnFK",
    "place": null,
    "lang": "in",
    "quote_count": 0,
    "favorited": false,
    "coordinates": null,
    "truncated": true,
    "timestamp_ms": "1562597391540",
    "reply_count": 0,
    "entities": {
      "urls": [
        {
          "display_url": "twitter.com/i/web/status/1148242800183349253",
          "indices": [117, 140],
          "expanded_url": "https://twitter.com/i/web/status/1148242800183349253",
          "url": "https://t.co/48NIDkZnFK"
        }
      ],
      "hashtags": [],
      "user_mentions": [],
      "symbols": []
    },
    "contributors": null,
    "user": {
      "utc_offset": null,
      "friends_count": 7094,
      "profile_image_url_https": "https://pbs.twimg.com/profile_images/1103027261651836928/710d17Bc_normal.jpg",
      "listed_count": 8,
      "profile_background_image_url": "http://abs.twimg.com/images/themes/theme1/bg.png",
      "default_profile_image": false,
      "favorites_count": 1057,
      "description": "Tempat BerCerita Tentang Kegelisahan Hati dan Fikiran",
      "created_at": "Sat Sep 21 12:27:57 +0000 2013",
      "is_translator": false,
      "profile_background_image_url_https": "https://abs.twimg.com/images/themes/theme1/bg.png",
      "protected": false,
      "screen_name": "reinmujaddid",
      "id_str": "1890062485",
      "profile_link_color": "00FFFF",
      "translator_type": "regular",
      "id": "1890062485",
      "geo_enabled": false,
      "profile_background_color": "05E7F7",
      "lang": null,
      "profile_sidebar_border_color": "FFFFFF",
      "profile_text_color": "333333",
      "verified": false,
      "profile_image_url": "http://pbs.twimg.com/profile_images/1103027261651836928/710d17Bc_normal.jpg",
      "time_zone": null,
      "url": null,
      "contributors_enabled": false,
      "profile_background_tile": true,
      "profile_banner_url": "https://pbs.twimg.com/profile_banners/1890062485/1558562503",
      "statuses_count": 68401,
      "follow_request_sent": null,
      "followers_count": 25174,
      "profile_use_background_image": true,
      "default_profile": false,
      "following": null,
      "name": "Malaikat Izrail",
      "location": "Yogyakarta, Indonesia",
      "profile_sidebar_fill_color": "DDEEF6",
      "notifications": null
    }
  }
}
```

Figure 3.2 Structure of a Tweet

3.1.5 Data Storage

The sink from Apache Flume will write the data permanently into HDFS (i.e. Hadoop Distributed File System). HDFS uses clusters to store its data which can be accessed in parallel. It has two main functions write and read. In HDFS write is performed only once whereas read can be performed n number of times. The data structure used to implement HDFS is a tree, where there exist a single name node and multiple data nodes. The HDFS system used for this research is a multimode Hadoop cluster where there exists one name node and 23 data nodes. Data when stored in the HDFS cluster it is in a raw format that needs to be processed.

3.1.6 Data Preprocessing

Since the data is present in a raw format (i.e. unstructured), data needs to be processed to eliminate noise. To perform preprocessing the data needs to be in the CSV format. To convert the data into a CSV file, data is transferred into a hive table which is an application that runs on the Hadoop framework. Hive converts the unstructured data into structured data. Tweets are present in flume files on HDFS which are loaded into the hive table by running the following command. This command will copy all the tweets from flume files into a hive table.

```
"hive> LOAD DATA LOCAL INPATH '/autohome/yalahar/Immigration_data/Feb_2019/02/10' INTO  
table Immigration_data_Feb_2019_02;"
```

Data can be downloaded in a CSV format using the following command.

```
"hive -e 'select * from tweets ' | sed 's/[\t]//,/g' >  
/autohome/yalahar/file.csv;"
```

3.1.6.1 Tokenization

Tweets are divided into tokens by using space as a delimiter. For example, if there is a tweet like "It's raining today J" then our tokens would be "It's", "raining", "today", "J". Now to eliminate non-English tweets, tokens are compared with an English dictionary. Tokens that don't match are eliminated.

3.1.6.2 Elimination of tweets without emoticons

The second step to preprocess our data is to eliminate the tweets which don't contain any emoticons.

3.1.6.3 Noise Elimination

To increase the accuracy, all the stop words, punctuations, hashtags, URLs, and special characters are eliminated from the tweets.

3.1.7 Classification of Emoticons

The following table (Table 3.2) gives a list of emoticons grouped into 4 different classes (i.e. Happy, Sad, Anger, and Celebrations).




| Happy Emoticons | Sad Emoticons | Angry Emoticons | Celebrations Emoticons |
|---|------------------------------------|--|--|
|  | <p>!"#\$ &'()* + ,</p> |  |  <p>-.012 34♀6789</p> |

Table 3.2 Classification of Emoticons

3.1.8 Categorization of Features

Some of the tweet features considered in this research for analysis are location and age.

3.1.8.1 Location

The location of a user is extracted from the flume file. The location has two values i.e. city and country. From this only, the country value of the user is extracted if he/she belongs to one of the 4 countries considered for analysis. The countries considered for this research are the United States of America, the United kingdom, Indian and Japan. Figure 3.3 illustrates the location of a tweet.

```

follow_request_sent":null,"followers_count":25174,"profile_use_background_image":true,"default_profile":false,"following":null,"name":"Malaikat Izrail","location":"Yogyakarta, Indonesia","profile_sidebar_fill_color":"DDEF6","notifications":null))

```

Figure 3.3 Location of the Tweet

3.1.8.2 Age

Some of the users display their birthday on twitter on their profile page. This information can be extracted using their IDs. All the user's ids are collected, and their birthdays are extracted from their profile page.

CHAPTER IV

ANALYSIS AND FINDINGS

Analyses performed in this research and the results observed are described in this chapter.

4.1 Analysis performed

As stated previously, we used five different datasets for our analysis. These datasets represent different domains such as politics, terrorism, and food. Available datasets were collected at different periods. Therefore, causality or similar relations among the datasets are not explored. The information extracted and their explanations are given in the following subsections.

4.1.1 Comparison of Emoticon Classes

As described in the previous section, all emoticons are grouped into four classes named 'happy', 'sad', 'angry', and 'celebration'. Figure 4.1 shows the distribution of these four classes in each dataset

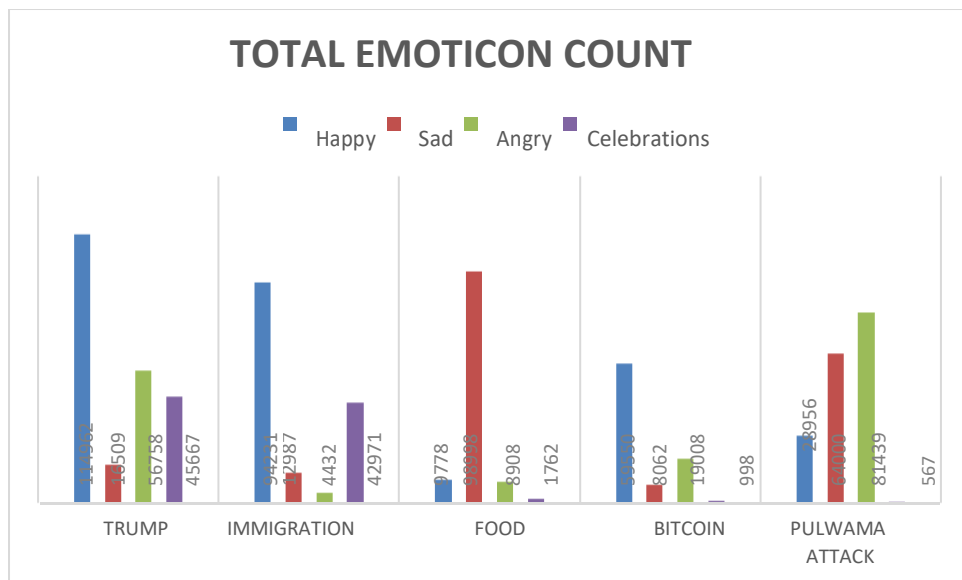


Figure 4.1 Total count of Emoticons in each dataset.

It can be observed from the figure that the distribution of the four classes varies from dataset to dataset. It is reasonable to assume that the classes of emoticons are dependent on the tweet topics. For example, followers of Trump are supportive of him and use happy emoticons, users who tweet on food poisoning topics are most likely affected by food poisoning and not happy, and users who tweet on terror-related topics are angry and also have no reason to celebrate.

4.1.2 Impact of Week Days on Emoticon Usage

This section explores the impact of weekdays on the usage of emoticon classes. Figures 4.2 – 4.5 show emoticon class usage on the different days of the week.

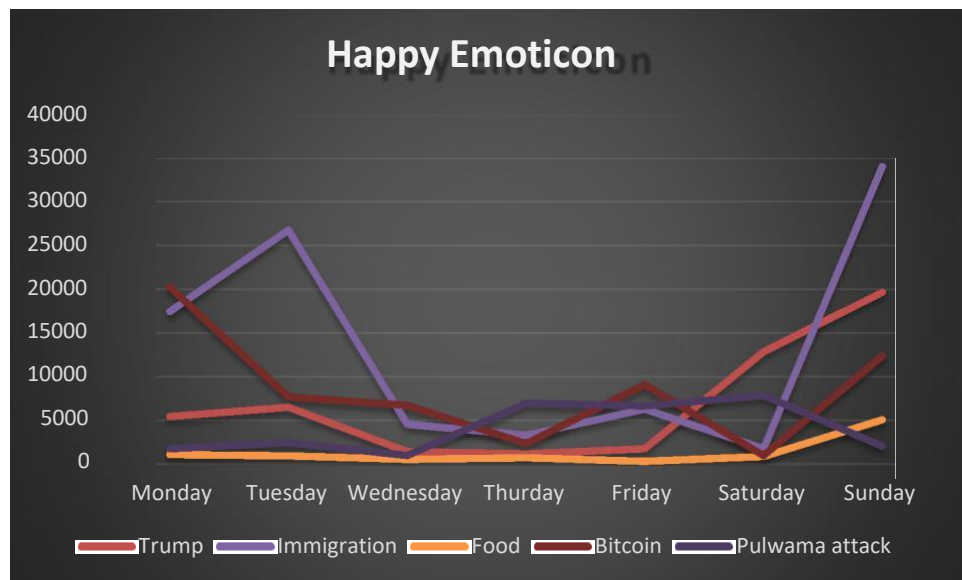


Figure 4.2 Total count of Happy Emoticons.

Figure 4.2 shows 'happy' emoticon counts by day for the different datasets. Emoticon counts are low in all datasets on Wednesday, Thursday and Friday.

Figures 4.3, 4.4 and 4.5 display similar graphs for the other three classes of emoticons. Based on these graphs, we can conclude that unless there are specific news events, the count of different

classes of emoticons does not change by the day. The spike on Thursday in the 'sad' and 'angry' groups can be explained by the occurrence of the Pulwama terror attack on a Thursday.

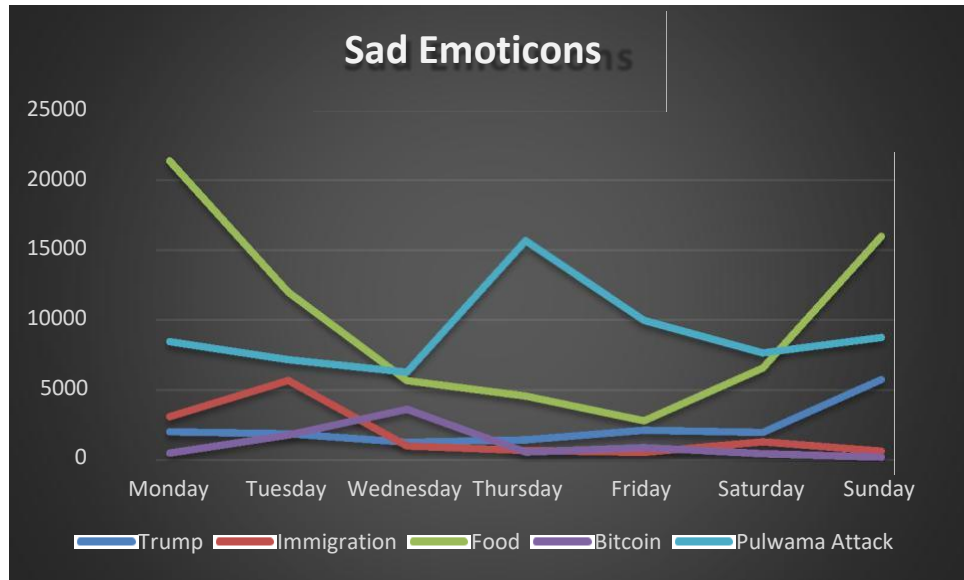


Figure 4.3 Total count of Sad Emoticons.

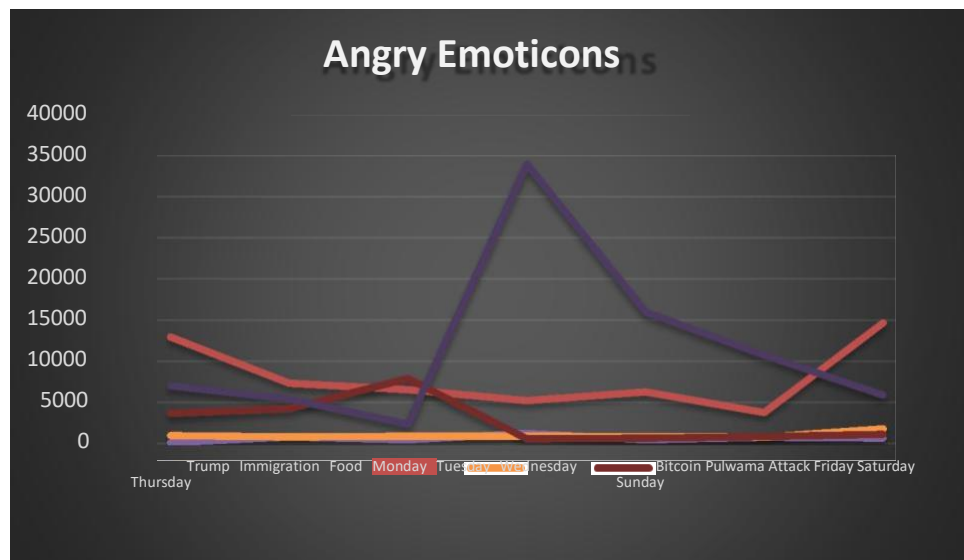


Figure 4.4 Total count of Angry Emoticons.

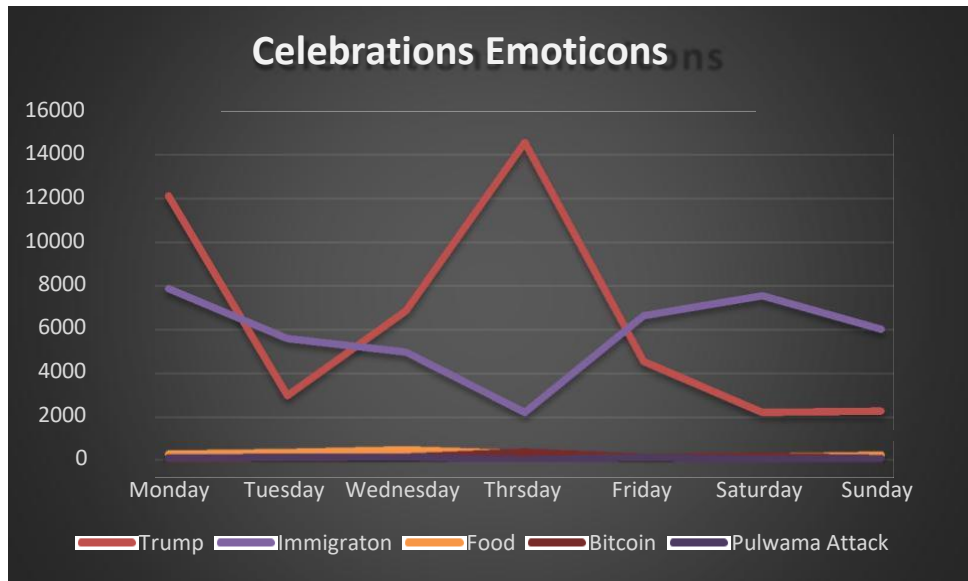


Figure 4.5 Total count of Celebrations Emoticons.

4.1.3 Use of emoticon classes by age groups

This section focuses on the usage of emoticons by age groups. Age is classified into 4 groups i.e. 18-25, 25-35, 35-45, and 60+. Users must meet the minimum age requirement of 18+ to have an account on Twitter.

Figure 4.6 displays the age group distribution of emoticon usage by different age groups. Emoticon usage increases by age. 44% of emoticon usage is by users in the age group 18-25, while only 14% is associated with users over 60 years old. Figures 4.7 – 4.10 show the distribution of different categories of emoticons used by the different age groups. Younger age groups (18-25) seem to use more 'happy' and 'sad' emoticons. The 25-35 and 60+ age groups use more 'happy' emoticons. The 35-60 age group uses more 'sad' and 'angry' emoticons. The 25-35 age group use a similar number of 'sad', 'angry' and 'celebration' emoticons.

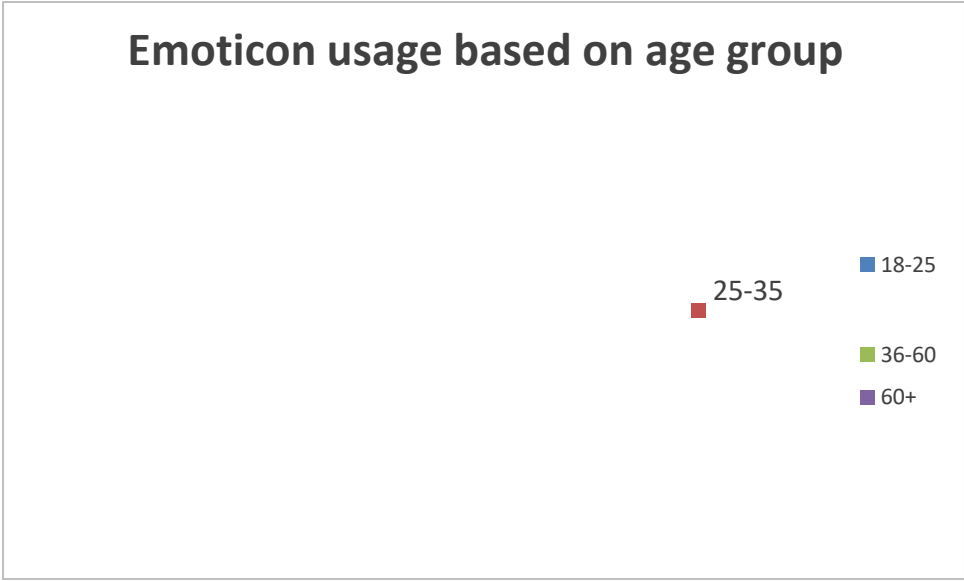


Figure 4.6 Total count of Emoticons based on Age.



Figure 4.7 Distribution of emoticon classes for age group 18-25.

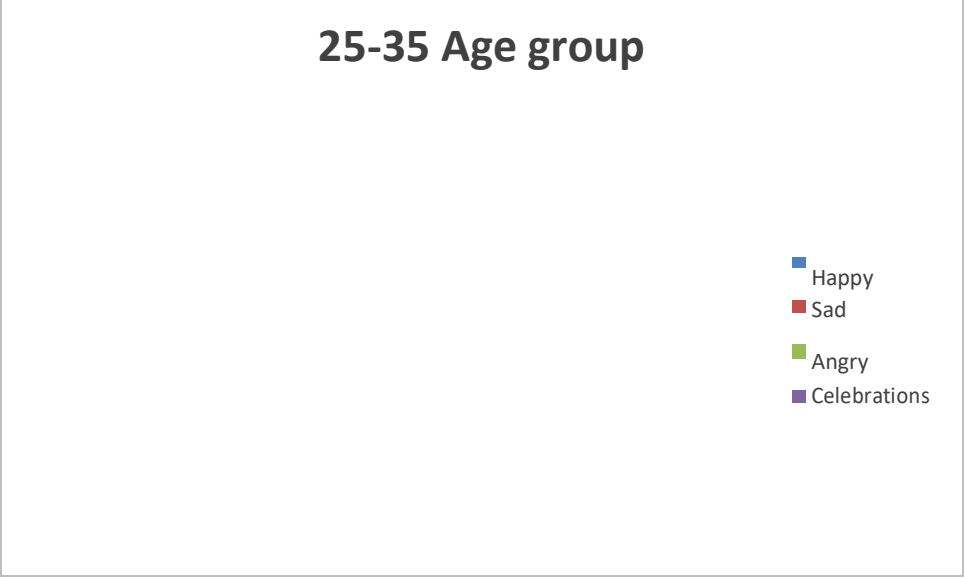


Figure 4.8 Distribution of emoticon classes for age group 25-35.



Figure 4.9 Distribution of emoticon classes for age group 35-60.

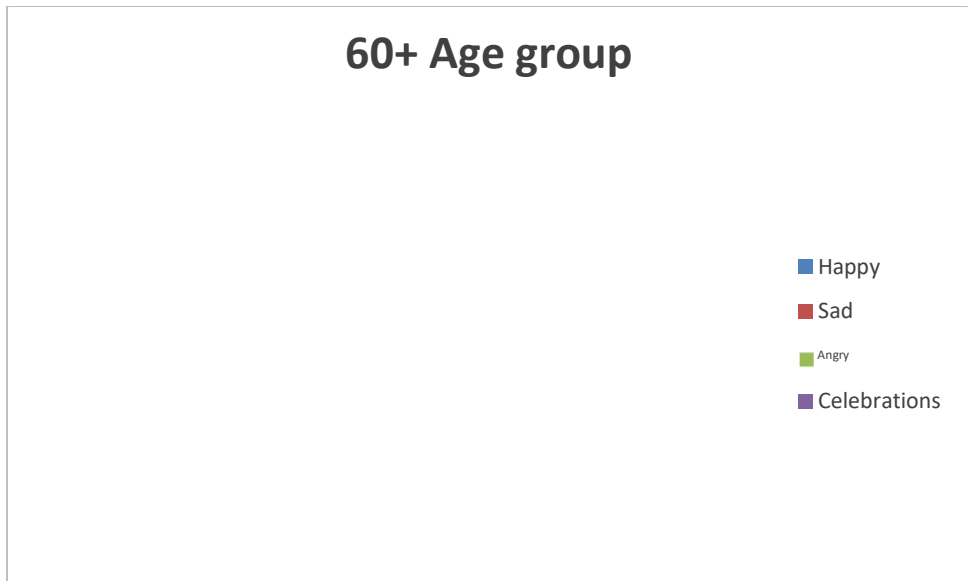


Figure 4.10 Distribution of emoticon classes for age group 60+.

4.1.4 Analysis based on Topic and Location

Four countries, USA, UK, Japan, and India are used to select tweets for the topic and location-based analysis. Figure 4.11 shows the histograms of the emoji classes for the Trump dataset. This dataset shows a country bias. Figures 4.12 – 4.15 show the distributions for the other datasets. Among the first four datasets, the USA has more emoticons than the others. In the Pulwama dataset, India has more emoticons.

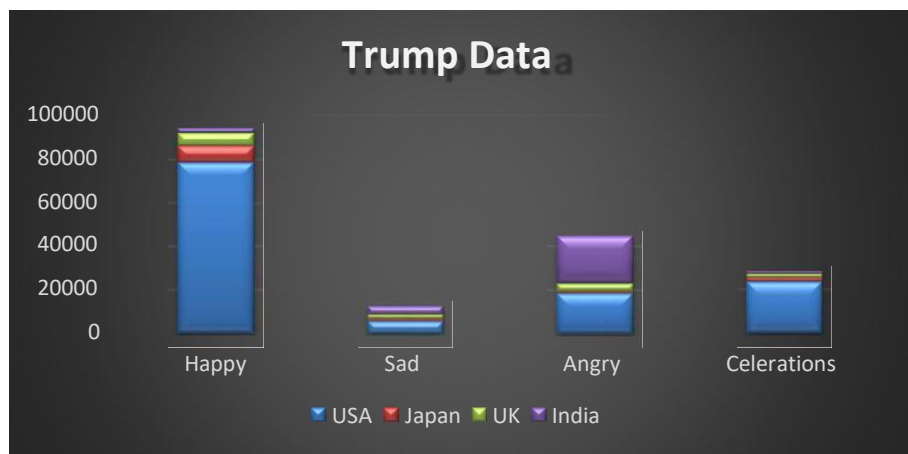


Figure 4.11 Distribution of emoticon classes in Trump dataset.

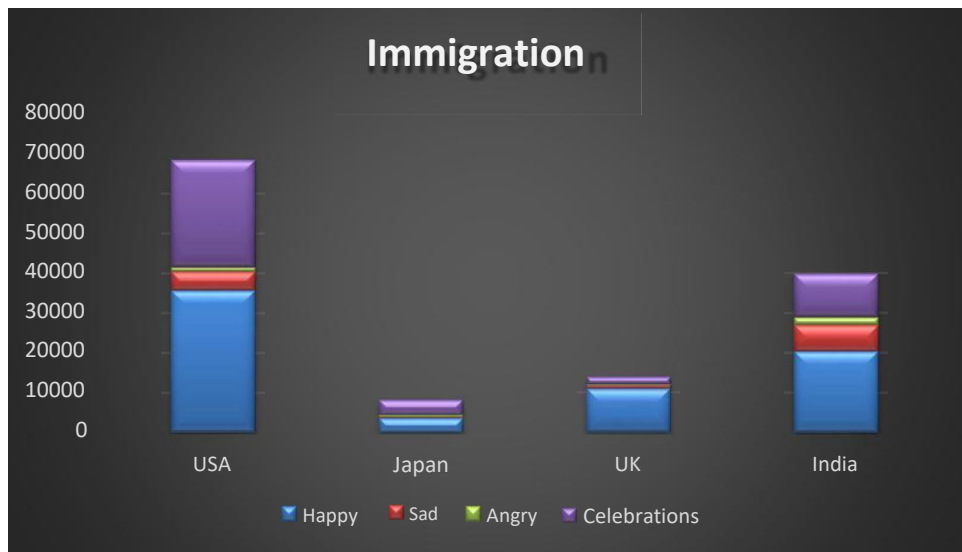


Figure 4.12 Distribution of emoticon classes in Immigration dataset.

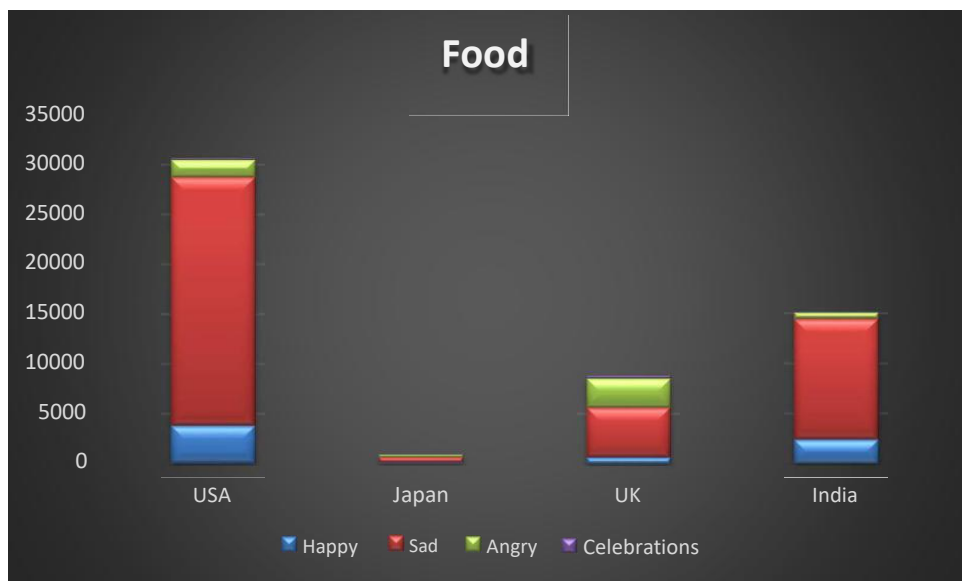


Figure 4.13 Distribution of emoticon classes in Food dataset.

In Figure 4.13 Sadness is the emotion expressed highly with the usage of emoticons by most of the countries. In figure 4.14 Happiness the emotion expressed highly with the usage of emoticons by most of the countries.

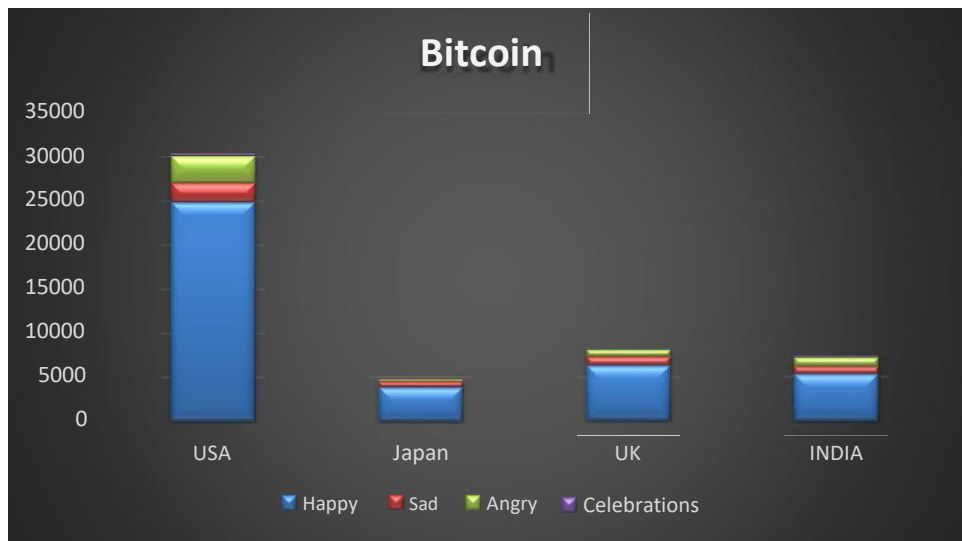


Figure 4.14 Distribution of emoticon classes in Bitcoin dataset.

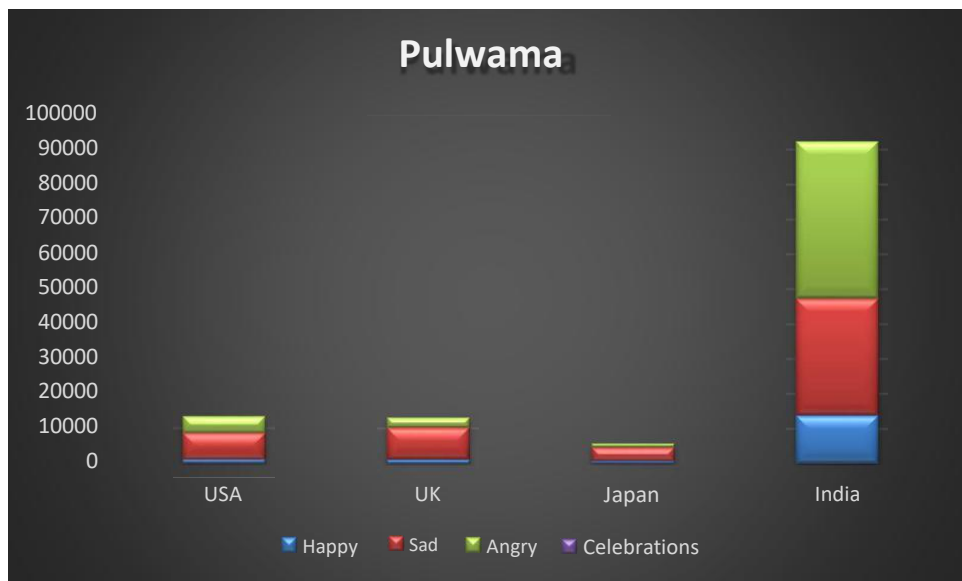


Figure 4.15 Distribution of emoticon classes in Pulwama dataset.

In figure 4.15 Pulwama is a terrorist attack that happened in India on Feb 14, 2019. Most of the emoticons used are from the users belonging to India. Sadness and Anger are expressed highly with the help of emoticons by users from India. Sadness is the emotion that is most expressed by users related to this topic. This dataset also shows a country bias.

Figure 4.16 is a summary of all the emoticon class distribution based on location. In the USA happy emoticon class is used maximum. In Japan and the United Kingdom, happy class emoticons are used slightly more when compared to the rest. India has used angry class emoticons more when compared to the rest.

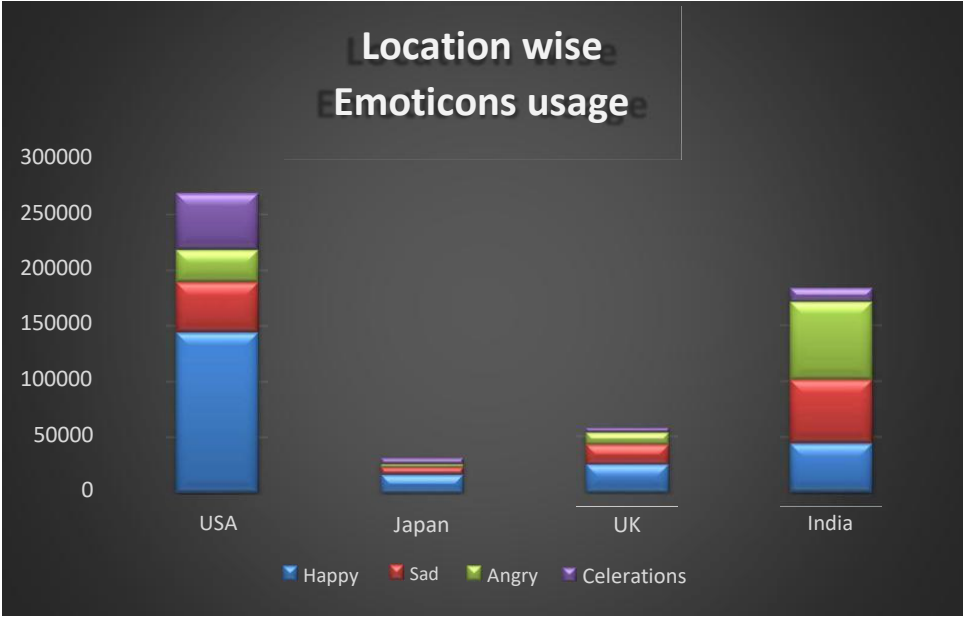


Figure 4.16 Distribution of emoticon classes based on Location.

4.1.5 Trend Analysis

In this section, we explore the correlation between all tweets and those contain emoticons. Daily counts of the two categories of tweets are shown as time series, Trends of emoticon usage is also displayed. Figures 4.17 – 4.21 show a comparison of total tweets vs tweets with emoticons for the different datasets.

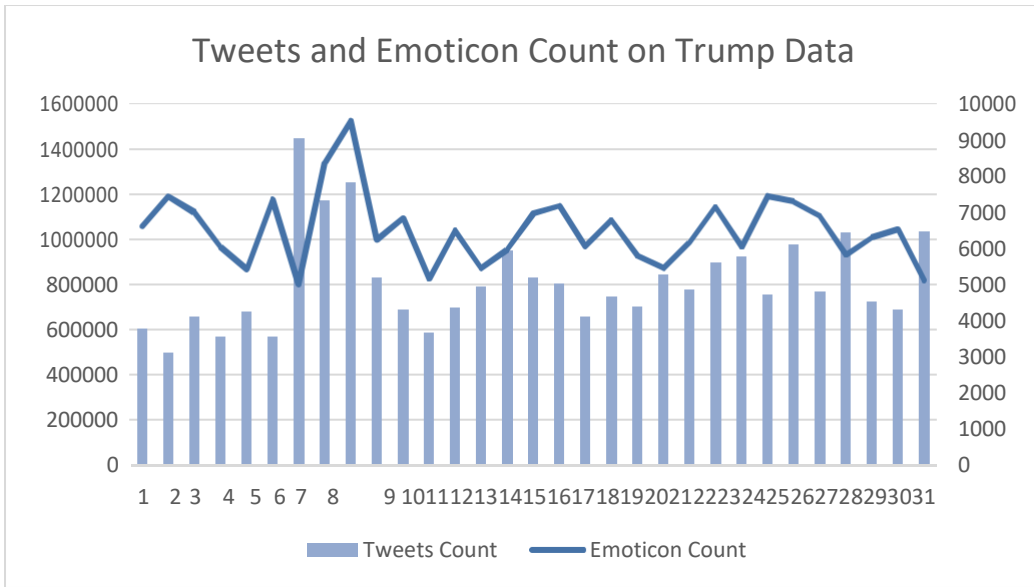


Figure 4.17 Day based -Time series of emoticon and tweet count of Trump data.

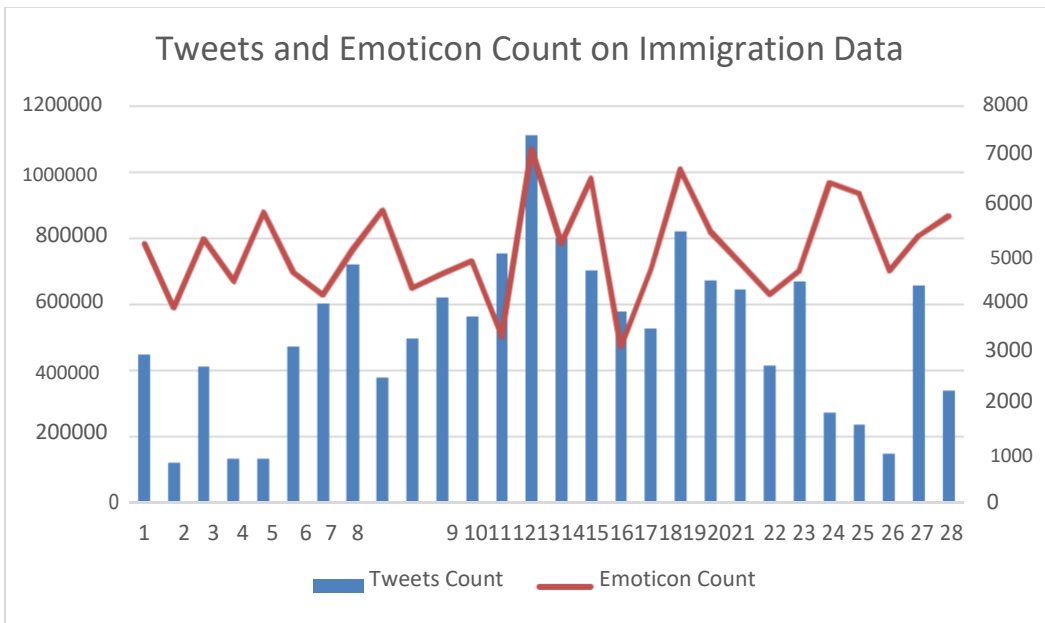


Figure 4.18 Day based -Time series of emoticon and tweet count of Immigration data.

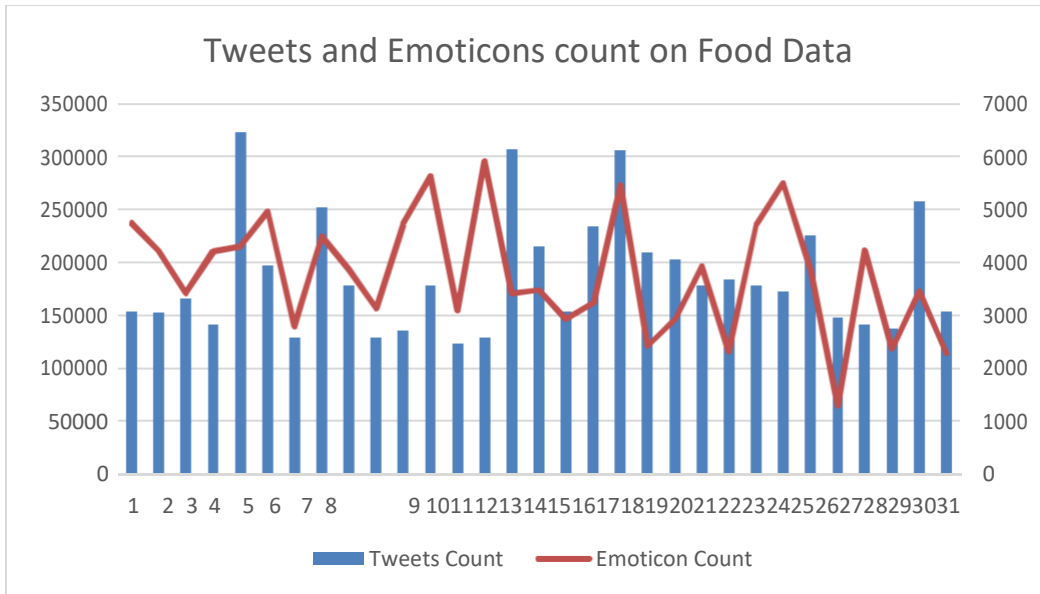


Figure 4.19 Day based -Time series of emoticon and tweet count of Food data.

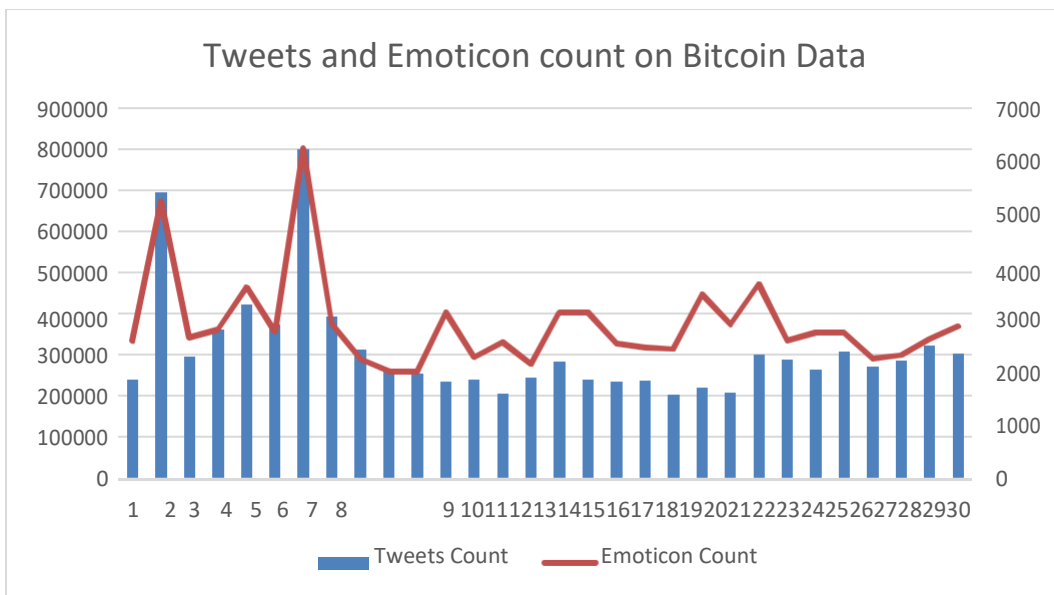


Figure 4.20 Day based -Time series of emoticon and tweet count of Bitcoin data.

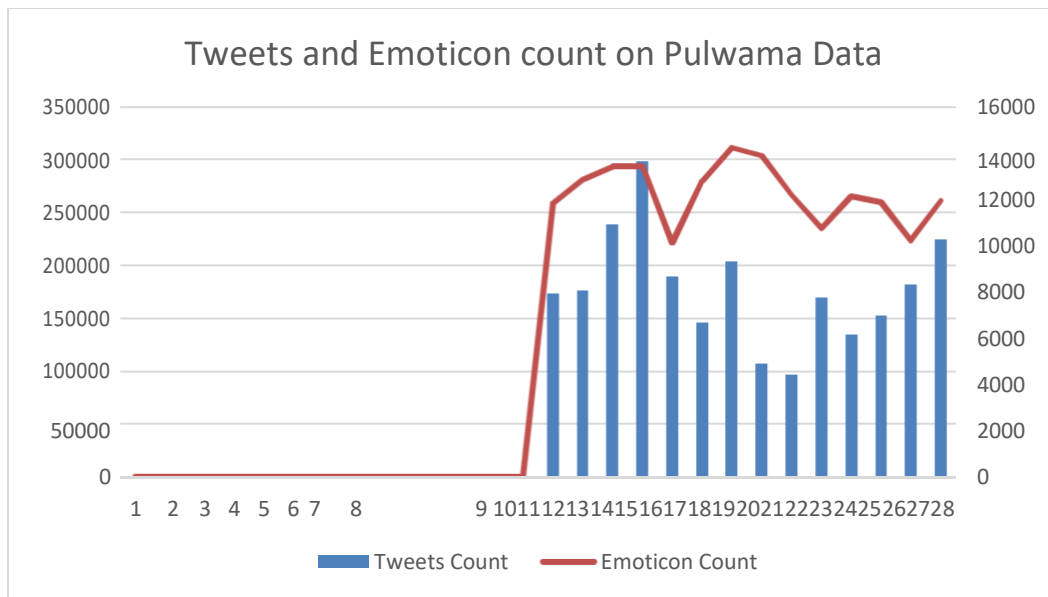


Figure 4.21 Day based -Time series of emoticon and tweet count of Pulwama data.

Figures 4.17-4.21 visualize the variations of tweets count and emoticon count over a period of a month. On further observation, we can say that as tweets count increases then there is growth in the emoticon count.

4.1.5.1 Trend of emoticons count

Figures 4.22- 4.27 illustrate the trends of emoticons count over a period of a month in each dataset. For datasets corresponding to Trump, Food and Bitcoin the trend indicates a downward trendline over the days. On the contrary, it shows an upward trendline concerning Pulwama and Immigration.

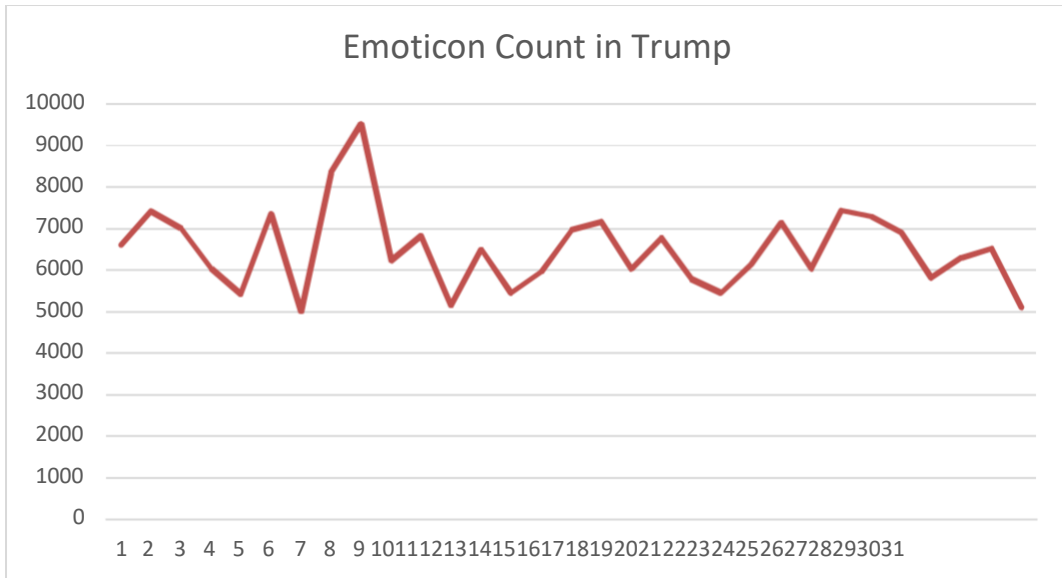


Figure 4.22 Trend of emoticons count of Trump data.

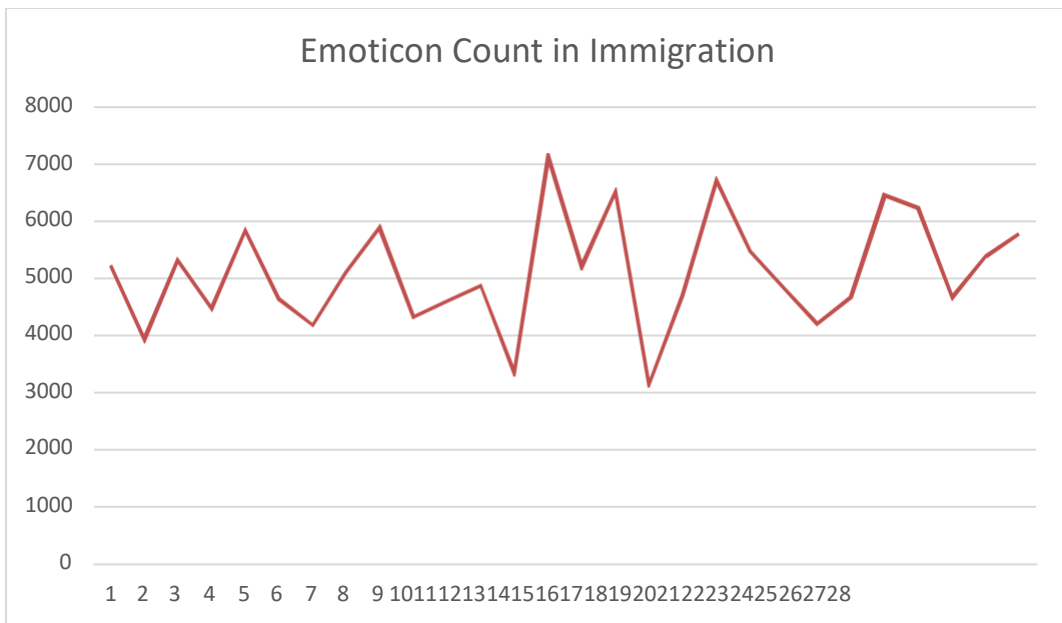


Figure 4.23 Trend of emoticons count of Immigration data.

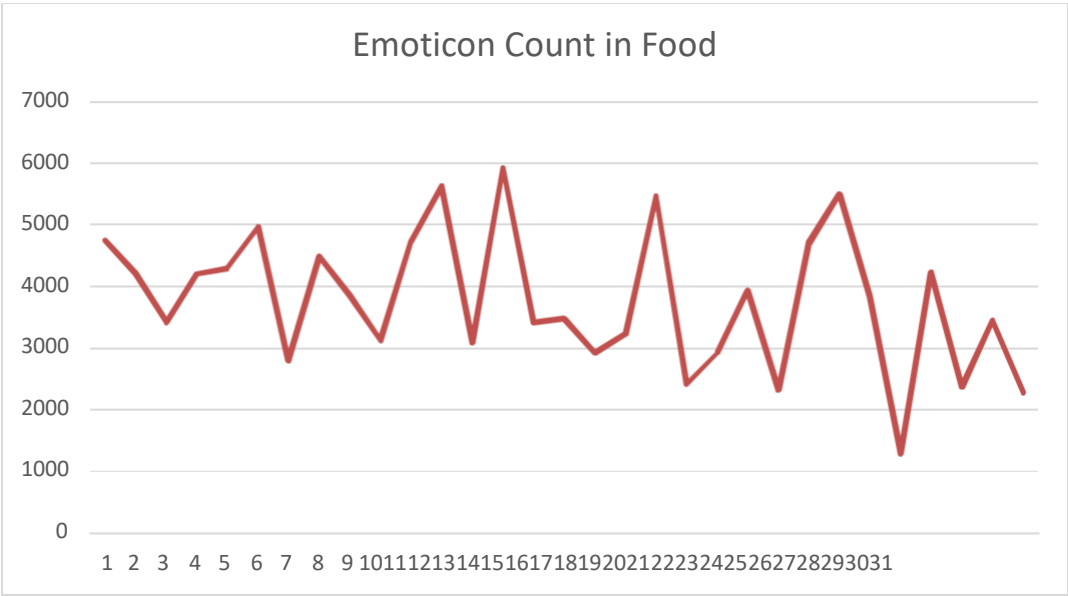


Figure 4.24 Trend of emoticons count of Food data.

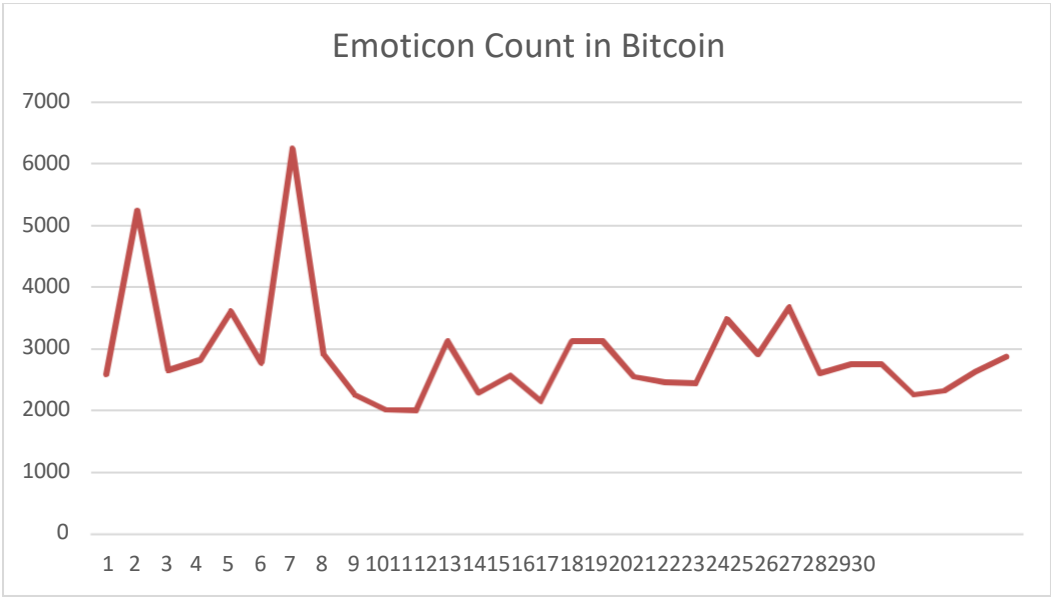


Figure 4.25 Trend of emoticons count of Bitcoin data.

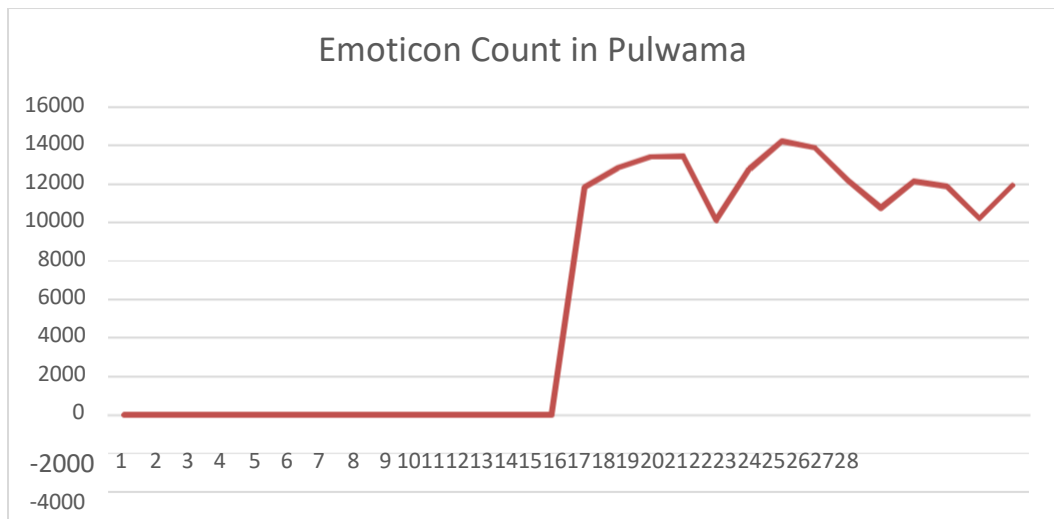


Figure 4.26 Trend of emoticons count of Pulwama data.

The above results indicate that the emoticon count depends on the tweet count. However, the trend in the usage of emoticons varies from dataset to dataset and depends on the period of analysis.

| Datasets | R square value for tweets count | R square value for emoticons count |
|-------------|---------------------------------|------------------------------------|
| Trump | 0.0329 | 0.0299 |
| Immigration | 0.0167 | 0.0519 |
| Food | 0.0019 | 0.1195 |
| Bitcoin | 0.1636 | 0.0754 |
| Pulwama | 0.5899 | 0.7058 |

Table 4.1 R squared values of tweets count and emoticon count for each dataset.

| Datasets | Trend values with respect to emoticon count and tweet counts |
|----------|--|
| Trump | 818638.2 |

| | |
|-------------|-----------|
| Immigration | 509684.6 |
| Food | 171176 |
| Bitcoin | 309476.03 |
| Pulwama | 3146.481 |

Table 4.2 Trend values for each dataset.

4.1.6 Correlation between Tweet count and Emoticon count

Table 4.3 illustrates the coefficient values obtained for each dataset.

| Data Set | Correlation Coefficient Value for Tweet count and Emoticon count |
|------------------|---|
| Trump Data | 0.1032 |
| Immigration Data | 0.1744 |
| Food data | 0.1455 |
| Pulwama Data | 0.1328 |
| Bitcoin Data | 0.8572 |

Table 4.3 Correlation Coefficient values of tweets count and emoticons count for each Dataset.

Table 4.3 depicts the positive correlation between tweets count and emoticon count in each dataset.

Positive correlation showcases the dependency of emoticon count on tweets count. Time series also demonstrates a homogeneous relationship between emoticon count and tweets count. In Bitcoin

data, there exists a fairly strong positive correlation between tweets count and emoticon count, whereas in the rest of the datasets there exists a moderate positive correlation between the two variables.

CHAPTER V

CONCLUSION

Twitter was created 13 years ago, it is a social media platform that provides users to express their views regarding various issues. Reportedly Twitter currently has 321 million users from all around the world who produce about 200 million tweets in a single day. There is plenty of data that can be analyzed to solve real-world issues. Analysis of such data can not only be used to solve real-world issues, but it can also be used to develop a prediction model and to boom the market value of products. Analysis of Twitter data has found its use in every industry possible. Analysis of this data can be used to predict the mood of users. Since there is a character limit of 140 for a tweet in twitter users tend to use emoticons to emphasize their emotion that can be expressed more conveniently than using a whole bunch of words. In this research different analysis was conducted on 5 datasets like Trump, Food, Immigration, Bitcoin and Pulwama attack. Observations of how the emoticon usage varied in these datasets when other features came into the picture. Analysis can be broken down into 4 main sections. In section 1 analysis was performed over the total emoticon count over all the five datasets. Emoticons are classified into 4 classes namely Happy, Sad, Anger and Celebrations. In section 2 analysis was performed to see how each class of emoticons behaved for the 5 datasets. In section 3 analysis was performed by taking into the feature of the age. Users were classified into 4 age groups depending upon their age and emoticon usage was observed for each age group. In section 4 analysis was performed based upon the location the users were since the users were from all around the world 4 locations were chosen to perform the analysis. From all the findings emoticon usage is affected by other features like location, topic, age, the day an event took place. These are some of the features that were observed in this research, but this can be extended to other features in the future. Emoticon usage depends upon location, topic and it is

incident driven. From our findings there exists a positive correlation between tweets count and emoticon count. This relationship is also depicted in Time series.

We made an effort to compare our results with published records. According to the article "Measuring National Well-being: At what age is Personal Well-being the highest?" by Office for National Statistics (UK - <https://www.ons.gov.uk/>), people aged between 65 to 74 had the highest average reported happiness – the lowest rating was amongst the group 45 to 59. A related graph is shown in Figure 5.1. This agrees with the results found in our study.

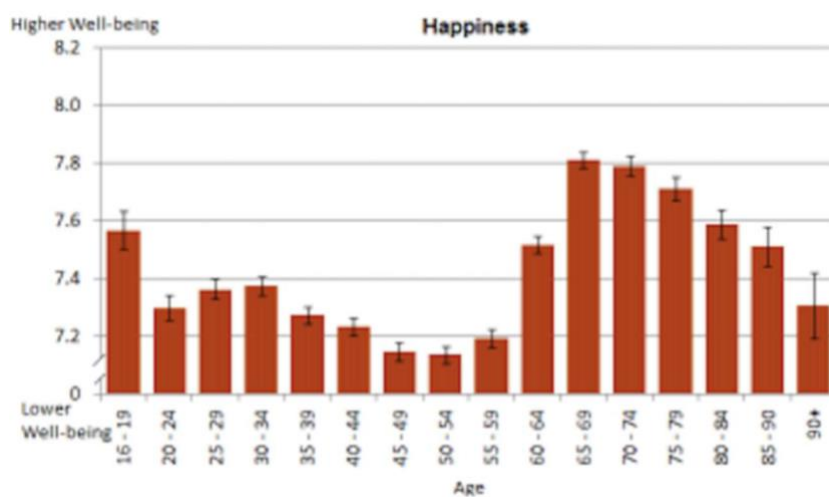


Figure 5.1 Office of National statistics results for happiest age group.

REFERENCES

- [1] Hao Wang, Jorge A. Castanon. Sentiment Expression via Emoticons on Social Media. Silicon Valley Lab. San Jose, USA.
- [2] Georgios S. Solakidis, Konstantinos N. Vavliakis, Pericles A. Mitkas. 2014. Multilingual sentiment analysis using emoticons and keywords. IEEE Computer Society Washington, DC, USA.
- [3] Xuan Lu, Wei Ai, Xuanzhe Liu, Qian Li, Ning Wang, Gang Huang, Qiaozhu Mei, 2016. Learning from the Ubiquitous Language: An Empirical Analysis of Emoticon Usage of Smartphone Users. UBIComp '16, HEIDELBERG, GERMANY.
- [4] Peijun Zhao, Jia Jia, Yongsheng An, Jie Liang, Lexing Xie, Jiebo Luo. 2018. Analyzing and Predicting Emoticon Usages in Social Media. Companion Proceedings of the Web Conference, Lyon, France.
- [5] Marco Vicente, Joao P. Carvalho, Fernando Batista. 2015. Using Unstructured Profile Information for Gender Classification of Portuguese and English Twitter Users. International Symposium on Languages, Applications, and Technologies.
- [6] Tweet normalization: A knowledge-based approach -Itisha Gupta, Nisheeth Joshi
- [7] Brandon Lwowski, Paul Rad, Kim-Kwang Raymong Choo. 2018. Geospatial Event Detection by Grouping Emotion Contagion in Social Media. IEEE Transactions on Big Data.
- [8] Sharath Chandra Guntuku, Mingyang Li, Louis Tay, and Lyle H. Ungar. 2019. Studying Cultural Differences in Emoticon Usage across the East and the West.

[9] 2016. Andrea Trevino Introduction to K-means Clustering from <https://blogs.oracle.com/datascience/introduction-to-k-means-clustering>

EXTERNAL LINKS

- [1] Marco Bonzanini. 2015. Mining Twitter Data from <https://marcobonzanini.com/2015/05/17/mining-twitter-data-with-python-part-6-sentiment-analysis-basics/>
- [2] <https://pinngle.me/blog/why-do-we-use-emoticons-and-emojis-science-based-facts/>
- [3] <https://www.wired.com/story/guide-emoji/>
- [4] Helen J. Wall, Linda K. Kaye, Stephanie A. Malone.2016. An exploration of psychological factors on emoticon usage and implications for judgement accuracy
- [5] <https://statisticsbyjim.com/basics/correlations/>
- [6] <http://www.abensourandpartners.com/sentiment-analysis/>
- [7] <https://www.mirror.co.uk/lifestyle/health/least-happy-age-group-country-7293609>
- [8] <https://monkeylearn.com/sentiment-analysis/>

APPENDICES

The below program is used to extract emoticons from a tweet and classify them i.e. if it belongs to Happy, Sad, Angry, or Celebrations class. The final output is the count of emoticons in each class as well as the total count of emoticons.

```
import java.io.*;
public class Main {
    public static String specialCharacters=" \\!#$%&'()*+,-./:;<=>?@[^_`{|}";
    public static String happy=" 😄 😁 😂 😃 😅 😆 😇 😈 😉 😊 😋 😌 😍 😎 😏 😐 😑 😒 😓 😔 😕 😖 😗 😘 😙 😚 😛 😜 😝 😞 😟 😠 😡 😢 😣 😤 😥 😦 😧 😨 😩 😪 😫 😬 😭 😮 😯 😰 😱 😲 😳 😴 😵 😶 😷 😸 😹 😺 😻 😼 😽 😾 😿 😺 😻 😼 😽 😾 😿";
    public static String sad=" 😞 😟 😠 😡 😢 😣 😤 😥 😦 😧 😨 😩 😪 😫 😬 😭 😮 😯 😰 😱 😲 😳 😴 😵 😶 😷 😸 😹 😺 😻 😼 😽 😾 😿";
    public static String angry=" 😡 😢 😣 😤 😥 😦 😧 😨 😩 😪 😫 😬 😭 😮 😯 😰 😱 😲 😳 😴 😵 😶 😷 😸 😹 😺 😻 😼 😽 😾 😿";
    public static String celebrations=" 🎆 🎇 🎈 🎉 🎊 🎋 🎌 🎍 🎎 🎏 🎐 🎑 🎒 🎓 🎔 🎕 🎖 🎗 🎘 🎙 🎚 🎛 🎜 🎝 🎞 🎟 🎠 🎡 🎢 🎣 🎤 🎥 🎦 🎧 🎨 🎩 🎪 🎫 🎬 🎭 🎮 🎯 🎰 🎱 🎲 🎳 🎴 🎵 🎶 🎷 🎸 🎹 🎺 🎻 🎼 🎽 🎾 🎿 🏀 🏁 🏂 🏃 🏄 🏅 🏆 🏇 🏈 🏉 🏊 🏋 🏌 🏍 🏎 🏏 🏐 🏑 🏒 🏓 🏔 🏕 🏖 🏗 🏘 🏙 🏚 🏛 🏜 🏝 🏞 🏟 🏠 🏡 🏢 🏣 🏤 🏥 🏦 🏧 🏨 🏩 🏪 🏫 🏬 🏭 🏮 🏯 🏰 🏱 🏲 🏳 🏴 🏵 🏶 🏷 🏸 🏹 🏺 🏻 🏼 🏽 🏾 🏿 🇧🇩 🇮🇳 🇵🇰 🇸🇪 🇸🇮 🇸🇯 🇹🇼 🇺🇸 🇻🇪 🇻🇮 🇻🇾 🇿🇦 🇿🇪";
    public static int counth=0;
    public static int counts=0;
    public static int counta=0;
    public static int countc=0;
    public static int count=0;
    public static void main(String[] args) throws IOException {
        File input = new File("/Users/yoshi/Desktop/input.txt");
        BufferedReader in = new BufferedReader(new FileReader(input));
        String emostring ;
        label:while((emostring=in.readLine()) != null) {
            for (int i=0; i < emostring.length( ); i++)
            {
                char c =emostring.charAt(i);
                if( (c >= 'a' && c <= 'z') || (c >= 'A' && c <= 'Z'))
                {
                    //eliminates all the ascii words
                }
                else if( (c >= '0' && c <= '9') )
                {
                    //eliminates numeric values
                }
                else if (specialCharacters.contains(Character.toString(emostring.charAt(i))))
                {

```

```

    {
        //eliminates special characters
    }
    else
    {
        if(happy.contains(Character.toString(emostring.charAt(i))))
        {
            counth++;
        }
        else if(sad.contains(Character.toString(emostring.charAt(i))))
        {
            counts++;
        }
        else if(angry.contains(Character.toString(emostring.charAt(i))))
        {
            counta++;
        }
        else if(celebrations.contains(Character.toString(emostring.charAt(i))))
        {
            countc++;
        }
        else
        {
            count++;
        }
    }
}

```

```

continue label;|

```

```
    }  
    count=count+counth+countc+counta+counts;  
    System.out.println("Total happy emoticons"+counth);  
    System.out.println("Total sad emoticons"+counts);  
    System.out.println("Total angry emoticons"+counta);  
    System.out.println("Total celebration emoticons"+countc);  
    System.out.println("Total count"+count);  
}  
}
```


VITA

Yoshitha Alahari

Candidate for the Degree of Master of Science

Thesis: INFLUENCE OF TWEET FEATURES ON EMOTICON USAGE

Major Field: Computer Science

Biographical:

Education:

Completed the requirements for the Master of Science in computer science at Oklahoma State University, Stillwater, Oklahoma in December, 2019.

Completed the requirements for the Bachelor of Science in Information Technology at Osmania University, Hyderabad, Telangana in 2017.