Abstract

Selecting a location to open an outlet for any business sector or expanding your ventures is a major concern. There are certain factors which need to be kept in mind before making decisions. The majors factor which is given the utmost importance is the interest of the people in the type of products being sold or service being provided by the company. If the area with most number of people having interest can be located and an outlet is opened in that area. The probability of earning a large profit increases as since more people are interested in that type of products ,They tend to visit the outlet more as compared to the other location. We will be using the tweets to know which area hast the most number of people interested in a particular type of product. People usually tweet about the topics they are interested in, it can be a type of food or a brand of clothes . We will create a graph by mapping the keyword found in the tweets with the location from which the tweet was posted. After this by applying hits algorithm we will find the best city for expansion.

Along with the we then apply sentimental analysis on the tweets to differentiate between the positive and the negative tweets. At the end we will be selecting the city with the most number of tweets having majority of the tweets as positive tweets.

Contents

1 Introduction

A part from the just the textual aspects tweets also contain important location information. This information can be used to perform analysis on the trends or interest of a group of people from a particular location. The result obtained from the analysis can be used to target the location where the people are already interested in the products that are same or similar to the products produced by the company.

We will be working on two data-sets. In the first data-set we will find the best city to open a new Pizza outlet. In the second data-set we will provide the best city for 'Grofer's' to expand in India. We will using the location of the tweets containing some specific keywords to find the city which has the most number of people interested in 'Eating Pizza' and 'Grofer's'. This information can be used to select the optimum city as it is expected that the people tweeting about pizza or something related to pizza are expected to visit the the outlet more as compared to other people. This will increase the sale of the outlet subsequently increasing the profit of the business. Similarly we people tweeting about delivery and grocery are expected to use the Grofer's service more than the other people.

1.1 Graph

The graph formed will be a bipartite graph in which will have two type of nodes . The first type will be the keywords and the second type will be the location of the tweet in which on the keyword was present.
An directed edge will be formed from the key word to the location of the tweet indicating that there is a tweet from a particular location which contains the keyword from which the edge is originating.

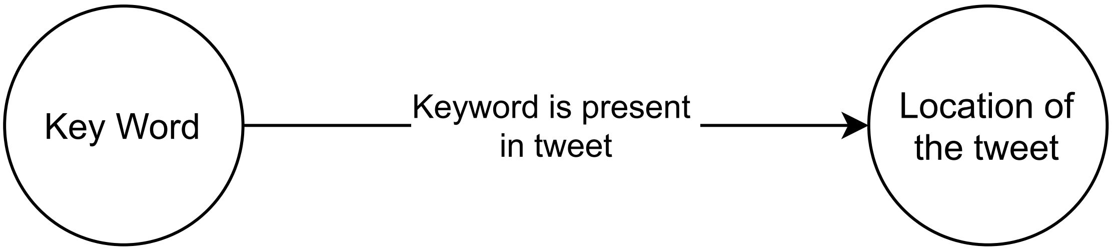


Figure 1.1: formation of edge in a graph

This way a graph containing all the keywords with their corresponding location will be formed. This will be a directed graph. The location which is common between multiple keywords will have incoming edges from all the related keywords.

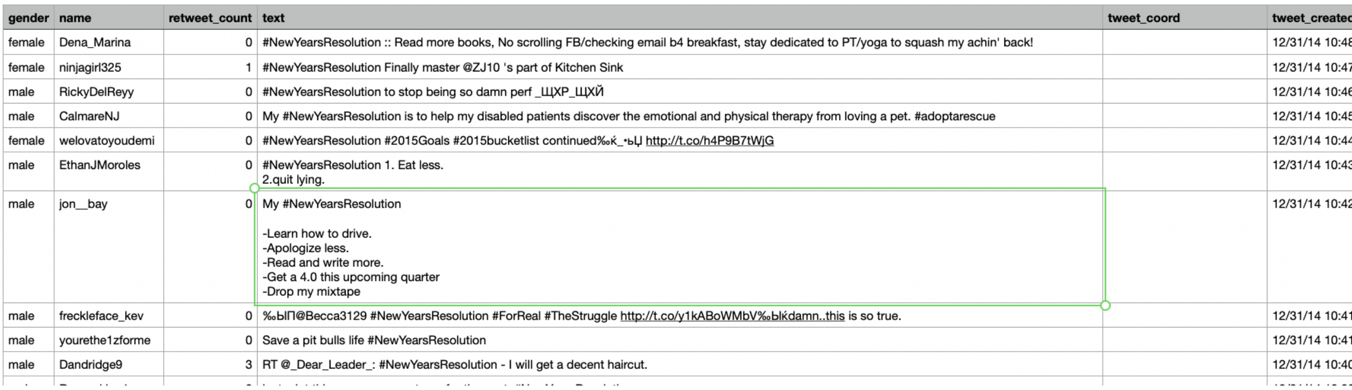
2 Literature Review

Earlier study related to Location Analysis for business purposes, Omar I. Aboulola et.al.[1] worked on location analysis for retail site location based on multiple factors and subfactors in predicting a location like Social Media data that influences the importance of a location economically, tweets, Yelp(Reviews, Locations, Ratings, etc.), Correlation markets and businesses that fits and fulfills the business and they used model of Multi Criteria Decision Making (MCDM), Neural network. Zulazeze Sahri et. al.[2] targets marketing by location from e-commerce website using web Analytics data, proposed a conceptual model on how e-commerce website's visitor data can helps in selecting targeted sales location using Digital Analytics tools. In order to allow Google Analytics to capture and analyze website user's data and behavior, they setup and install analytics tracking code into the sample website, used Geo-Location(Number of visitors, Average session duration). Anant Gupta et. al.[3] proposed a location based personalized Restaurant Recommendation system for Moblie Environments, they used google's geo-location api and map to track the user's location to get the useful insights.They have divided their system architecture in two sections: one in online activity when user is in motion and other is which processes data offline.CHRIS H.Q. DING† , HONGYUAN ZHA‡ , XIAOFENG HE† , PARRY HUSBANDS† , AND HORST D. SIMON† worked on applying Hits algorithm on bipartite graph. They concluded that the Authority value is very closely related to the in-degree of node.

3 Data-sets

3.1 First Data-set

Two data-sets were used in the project. First data-set was downloaded from a GitHub repository on dataworld.com . This data-set consisted of the new year resolutions of the people from different US cities. The format of the data-set is shown in figure 3.1.



gender	name	retweet_count	text	tweet_coord	tweet_create
female	Dena_Marina	0	#NewYearsResolution :: Read more books, No scrolling FB/checking email b4 breakfast, stay dedicated to PT/yoga to squash my achin' back!		12/31/14 10:48
female	ninjagirl325	1	#NewYearsResolution Finally master @ZJ10 's part of Kitchen Sink		12/31/14 10:47
male	RickyDelReyy	0	#NewYearsResolution to stop being so damn perf _ЩХР_ЩХЙ		12/31/14 10:46
male	CalmareNJ	0	My #NewYearsResolution is to help my disabled patients discover the emotional and physical therapy from loving a pet. #adoptarescue		12/31/14 10:45
female	welovatoyoudemi	0	#NewYearsResolution #2015Goals #2015bucketlist continued‰ќ_•ьЏ http://t.co/h4P9B7tWjG		12/31/14 10:44
male	EthanJMoroles	0	#NewYearsResolution 1. Eat less. 2.quit lying.		12/31/14 10:43
male	jon__bay	0	My #NewYearsResolution -Learn how to drive. -Apologize less. -Read and write more. -Get a 4.0 this upcoming quarter -Drop my mixtape		12/31/14 10:42
male	freckleface_kev	0	‰ЫП@Becca3129 #NewYearsResolution #ForReal #TheStruggle http://t.co/y1kABoWMbV‰ЫЌdamn..this is so true.		12/31/14 10:41
male	yourethe1zforme	0	Save a pit bulls life #NewYearsResolution		12/31/14 10:41
male	Dandridge9	3	RT @_Dear_Leader_: #NewYearsResolution - I will get a decent haircut.		12/31/14 10:40

Figure 3.1: Initial data-set

3.2 Second Data-set

Second data-set was formed by getting live tweets directly from twitter using tweepy. We Created Twitter developer account which got accepted only after we provided the proper reason for the requirement of the data and brief description of our research work. We Created twitter App via the twitter app dashboard page. Then, we generated access tokens on the "Keys and Tokens" tab in an app's "Details" section within the Twitter app dashboard. Then, We Connected with twitter API and extracted the live data from twitter.

3.2.1 Using the access tokens

First we create twitter developer account, after getting the twitter developer account, we create a dummy app on twitter dashboard then from that app we got the necessary keys and tokens which we need to connect the twitter API.

Code to Connect with Twitter API

```
#Get from developers.twitter.com/App—>Setting—>keys&tokens
#Just assign the credentials

consumer_key = ”YCyZzlGlH4bVcgu9DEtXGGCtb”
consumer_secret = ”zBH1Yus7hpCE7R1P9oRTqu7fHXzSKRzWNfGQRLQ3qA3HsuH3V5”
access_token = ”1406151650465648640–HNSctNOHvpkYIiwNaUnIwCL6uhpycN”
access_token_secret = ”fuR5lHVZKOdNB1Qn9A2awjzSuqtptFLwHu5muErvzH5gI”

#Use the above credentials to authenticate the API.

auth = tweepy.OAuthHandler( consumer_key , consumer_secret )
auth.access_token_secret
auth.set_access_token( access_token , access_token_secret )
api = tweepy.API(auth)
```

3.2.2 Extracting tweets using Tweepy and storing in data frame

Using tweepy.Cursor() we will fetch the tweets for the Topic we selected, Every tweet will come with information such as username of the person who tweeted it, Likes/Retweets for that tweet, Location of the User and so on.
We can just use whichever attributes that interest us and store it in a DataFrame so we can further process it.

Code for fetching tweets and Storing them in dataframe

```
 df = pd.DataFrame(columns=[”Date”,”User”,”IsVerified”,”Tweet”,”Likes”,”RT”,
 ”User_location”])


 def get_tweets(Topic,Count):
      i=0
 for tweet in tweepy.Cursor(api.search,q=Topic,count=100,lang=”en”,
 exclude='retweets').items():
           print(i, end='\r')
           df.loc[i,”Date”] = tweet.created_at
           df.loc[i,”User”] = tweet.user.name
           df.loc[i,”IsVerified”] = tweet.user.verified
           df.loc[i,”Tweet”] = tweet.text
           df.loc[i,”Likes”] = tweet.favorite_count
           df.loc[i,”RT”] = tweet.retweet_count
           df.loc[i,”User_location”] = tweet.user.location
           #df.to_csv(”TweetDataset.csv”,index=False)
           df.to_excel('{}.xlsx'.format(”TweetDataset”),index=False)
## Save as Excel
           i=i+1
           if i>Count:
                break
           else:
                pass

 Topic=[”Grofers”]
 get_tweets(Topic , Count=1000)
```

3.2.3 Pre-processing the Tweets

We have cleaned the tweets using the **re** library of regular expression. In this, we have removed any tags like “@,#,$,%,” . We will also remove Special characters and any links.

Code for Processing the data

```
import re
def clean_tweet(tweet):
return ' '.join(re.sub('(@[A-Za-z0-9]+)|([^0-9A-Za-z \t])|(\w+:\/\/\S+)|
([RT])', ' ', str(tweet).lower()).split())

# Call function to get Clean tweets

df['clean_tweet'] = df['Tweet'].apply(lambda x : clean_tweet(x))
df.head(1000)
```

4 Sentiment Analysis

Sentiment analysis was performed on the text of tweets using **Textblob**. This was done to differentiate the positive tweets from the other tweets. This allowed to focus on the cities where the people are reacting positively to the service already present there or they are interested in a particular product. The tweets were divided into three categories.

- Positive Sentiments : People are satisfied with the service or product.
- Neutral Sentiments: People are neither satisfied nor unsatisfied by the service or product.
- Negative Sentiments: People are not satisfied with the service or product.

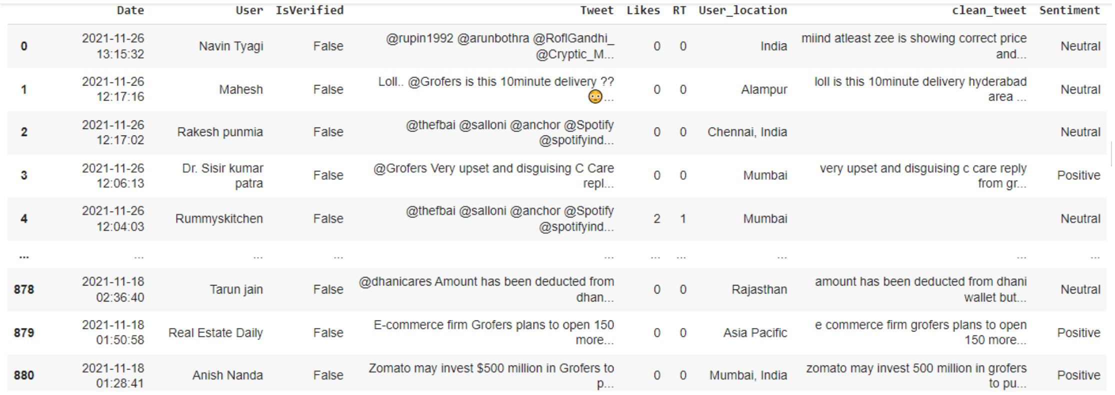
4.1 Code for Sentiment Analysis

```
# Funciton to analyze Sentiment

from textblob import TextBlob
def analyze_sentiment(tweet):
    analysis = TextBlob(tweet)
    if analysis.sentiment.polarity > 0:
        return 'Positive'
    elif analysis.sentiment.polarity == 0:
        return 'Neutral'
    else:
        return 'Negative'

# Call function to get the Sentiments

df[”Sentiment”] = df[”Tweet”].apply(lambda x : analyze_sentiment(x))
df.head(1000)
```

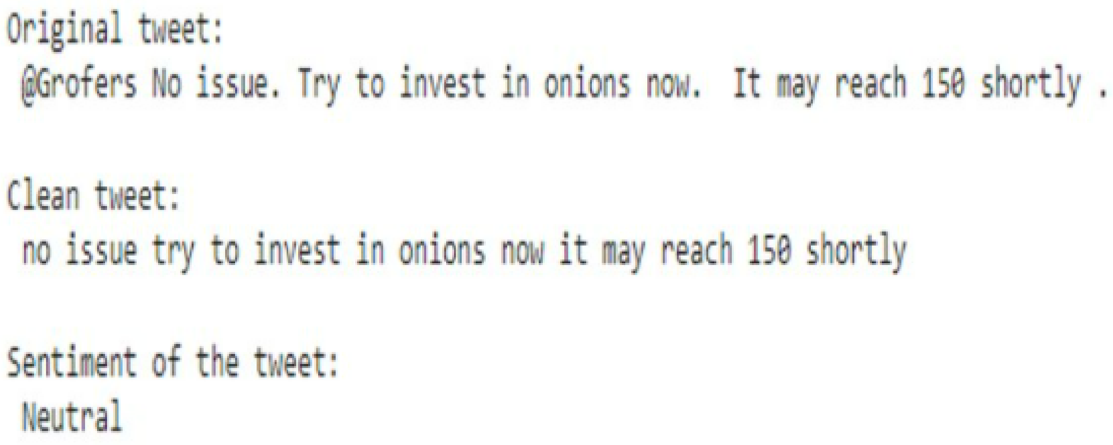


	Date	User	IsVerified	Tweet	Likes	RT	User_location	clean_tweet	Sentiment
0	2021-11-26 13:15:32	Navin Tyagi	False	@rupin1992 @arunbothra @RoflGandhi_ @Cryptic_M...	0	0	India	miind atleast zee is showing correct price and...	Neutral
1	2021-11-26 12:17:16	Mahesh	False	Loll.. @Grofers is this 10minute delivery ?? 😳...	0	0	Alampur	loll is this 10minute delivery hyderabad area ...	Neutral
2	2021-11-26 12:17:02	Rakesh punmia	False	@thefbai @salloni @anchor @Spotify @spotifyind...	0	0	Chennai, India		Neutral
3	2021-11-26 12:06:13	Dr. Sisir kumar patra	False	@Grofers Very upset and disguising C Care repl...	0	0	Mumbai	very upset and disguising c care reply from gr...	Positive
4	2021-11-26 12:04:03	Rummyskitchen	False	@thefbai @salloni @anchor @Spotify @spotifyind...	2	1	Mumbai		Neutral
...	...	...	...	...	...	...	...	...	...
878	2021-11-18 02:36:40	Tarun jain	False	@dhanicares Amount has been deducted from dhan...	0	0	Rajasthan	amount has been deducted from dhani wallet but...	Neutral
879	2021-11-18 01:50:58	Real Estate Daily	False	E-commerce firm Grofers plans to open 150 more...	0	0	Asia Pacific	e commerce firm grofers plans to open 150 more...	Positive
880	2021-11-18 01:28:41	Anish Nanda	False	Zomato may invest $500 million in Grofers to p...	0	0	Mumbai, India	zomato may invest 500 million in grofers to pu...	Positive

Figure 4.1: output of sentiments

4.2 Example of each type of tweet

Neutral Tweets



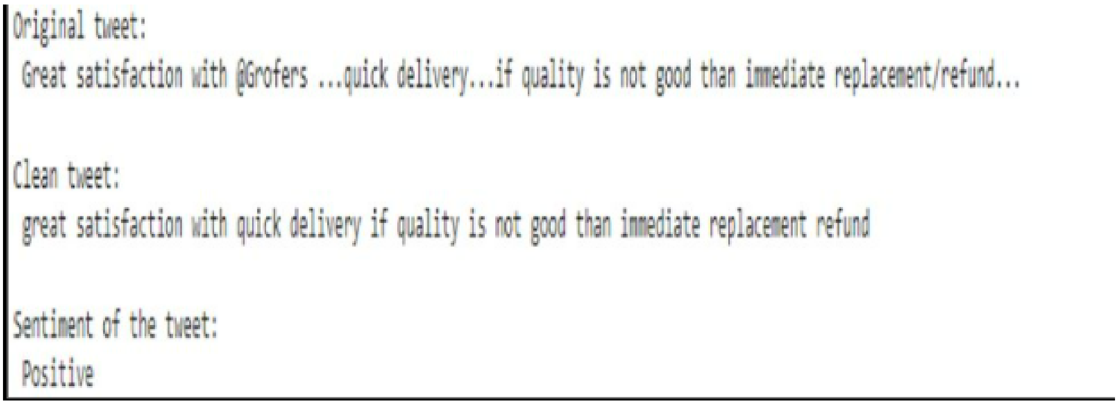
```
Original tweet:
 @Grofers No issue. Try to invest in onions now.  It may reach 150 shortly .

Clean tweet:
 no issue try to invest in onions now it may reach 150 shortly

Sentiment of the tweet:
 Neutral
```

Figure 4.2: Neutral Sentiments of tweets

Positive Tweets



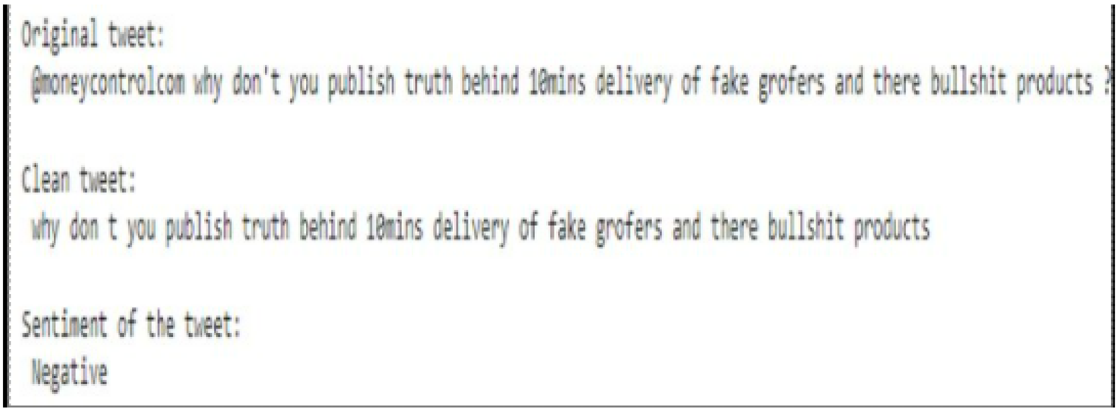
```
Original tweet:
 Great satisfaction with @Grofers ...quick delivery...if quality is not good than immediate replacement/refund...


Clean tweet:
 great satisfaction with quick delivery if quality is not good than immediate replacement refund


Sentiment of the tweet:
 Positive
```

Figure 4.3: Positive Sentiments of tweets

Negative Tweets



```
Original tweet:
 @moneycontrolcom why don't you publish truth behind 10mins delivery of fake grofers and there bullshit products ?


Clean tweet:
 why don t you publish truth behind 10mins delivery of fake grofers and there bullshit products


Sentiment of the tweet:
 Negative
```

Figure 4.4: Negative Sentiments of tweets

4.3 Graphical representation of Sentiments

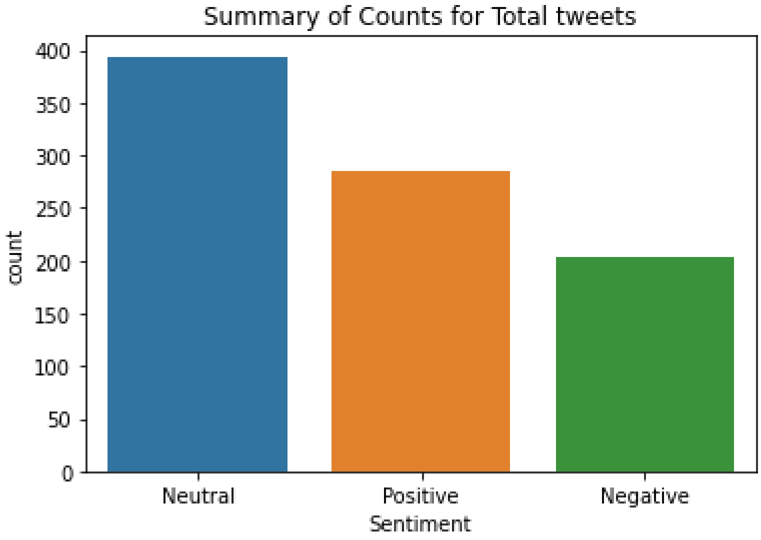


Figure 4.5: Plot to show summary of total counts

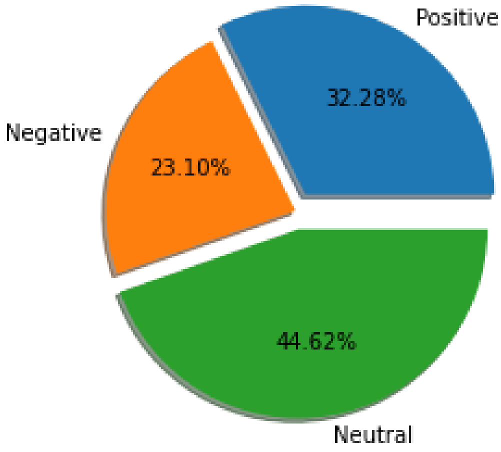


Figure 4.6: Pie-chart to show the percentage of each types of tweets

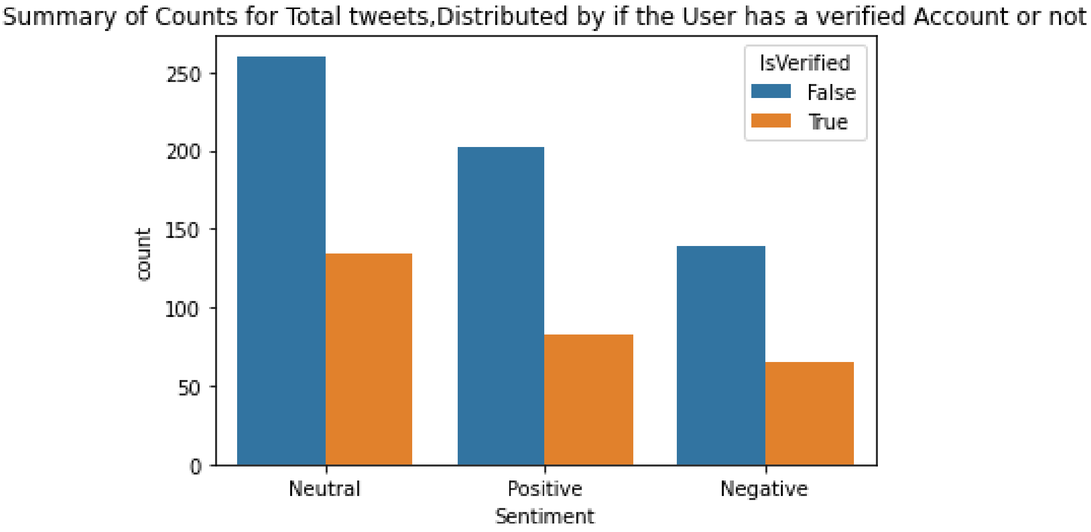


Figure 4.7: Plot to show the summary of verified users

Status of verification of an user was also found out. To avoid the spam tweets in the data-set. It allowed us to focus on only the real tweets ad avoid the spams by fake twitter ids.

5 World Cloud

World cloud is great way of visualising the data in a unique manner. Using this we can know the most frequent or the least frequent word in a data. Along with this we can also visualise the different words present in the data at the same time.
A word cloud was formed using the data. In the word cloud the size of the word was varied according to the frequency of occurrence of word in the data-set. Word with the most frequency was given the largest size and size kept on decreasing with the frequency. Here the words were the location of positive tweets. Largest size indicates that the location has the most number of tweets containing the keyword ans small size means that location do not have much tweets containing the keyword.

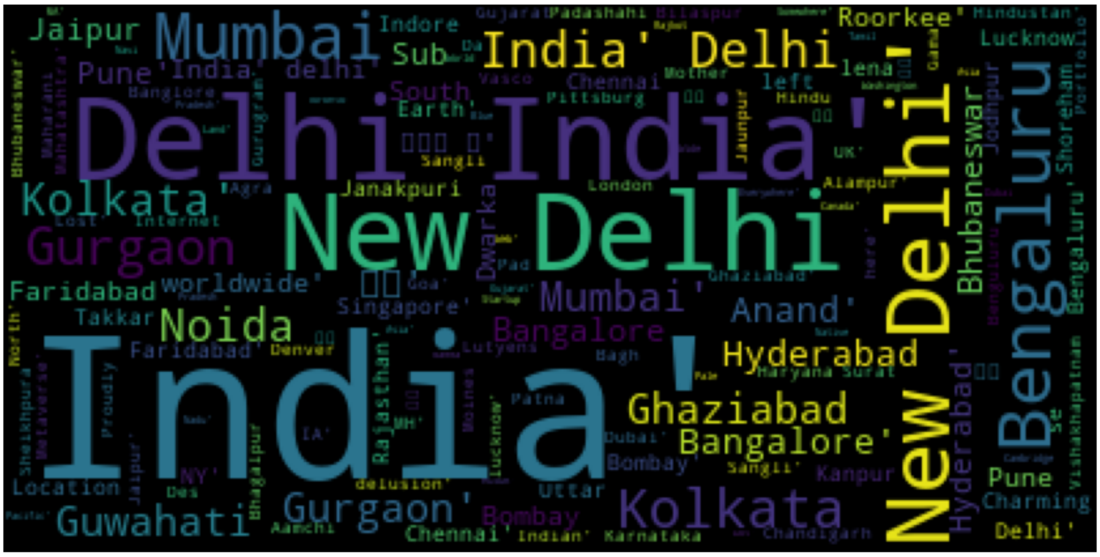


Figure 5.1: output of sentiments

6 Graph Plotting

6.1 Individual keywords Graphs

Initially graphs with individual keywords were plotted which contained location and a single keyword as nodes. An edge was formed from location to the keyword. A example graph is shown in figure 6.1. Here the keyword is Domino's and the location of the tweets containing this keyword are connected to the node.

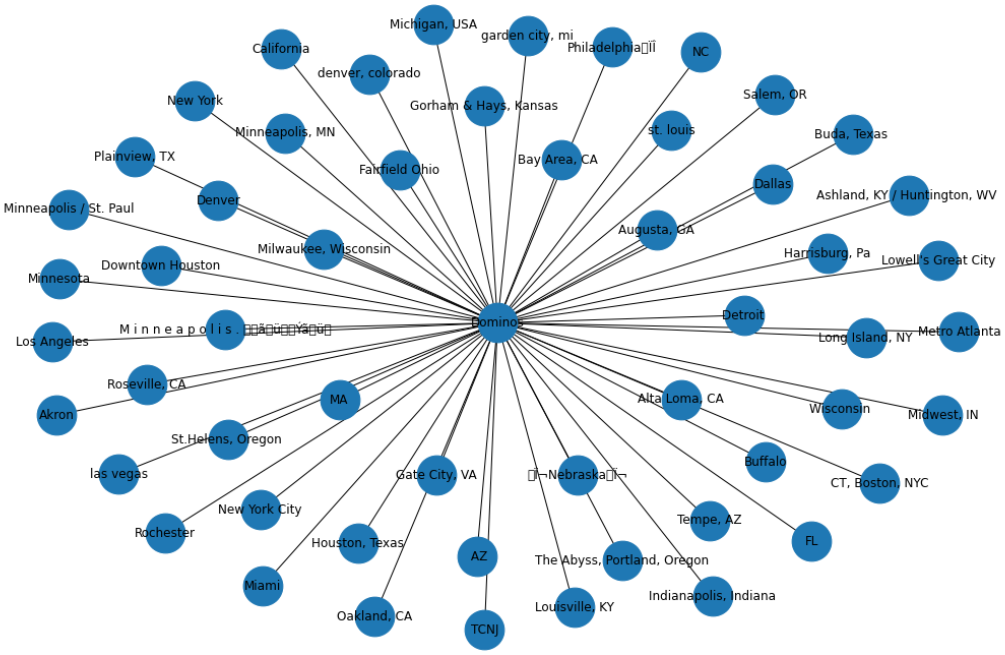


Figure 6.1: graph of Domino's keyword

6.2 Combined graph

A common graph was created in which all the keywords along with their location were mapped together. The locations in which the tweets contained multiple keywords were connected to multiple nodes with an edge.
Gephi was used to plot this graph. To improve the observations from the graph Size and colour of the nodes were varied based on the in-degree. Such that the node with the highest in-degree was biggest. No overlap layout was used in Gephi while creating the graph. Observation was made by looking at the size that is, node with the biggest size contains the greatest number of positive tweets containing the keywords.

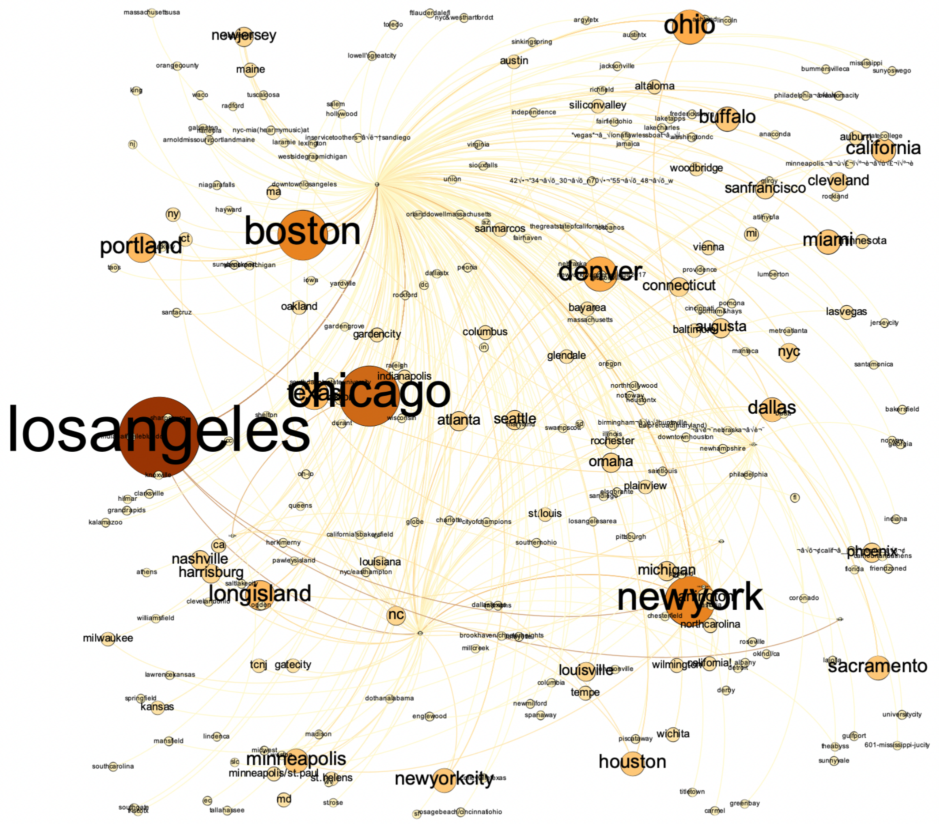


Figure 6.2: Graph for first Data-set

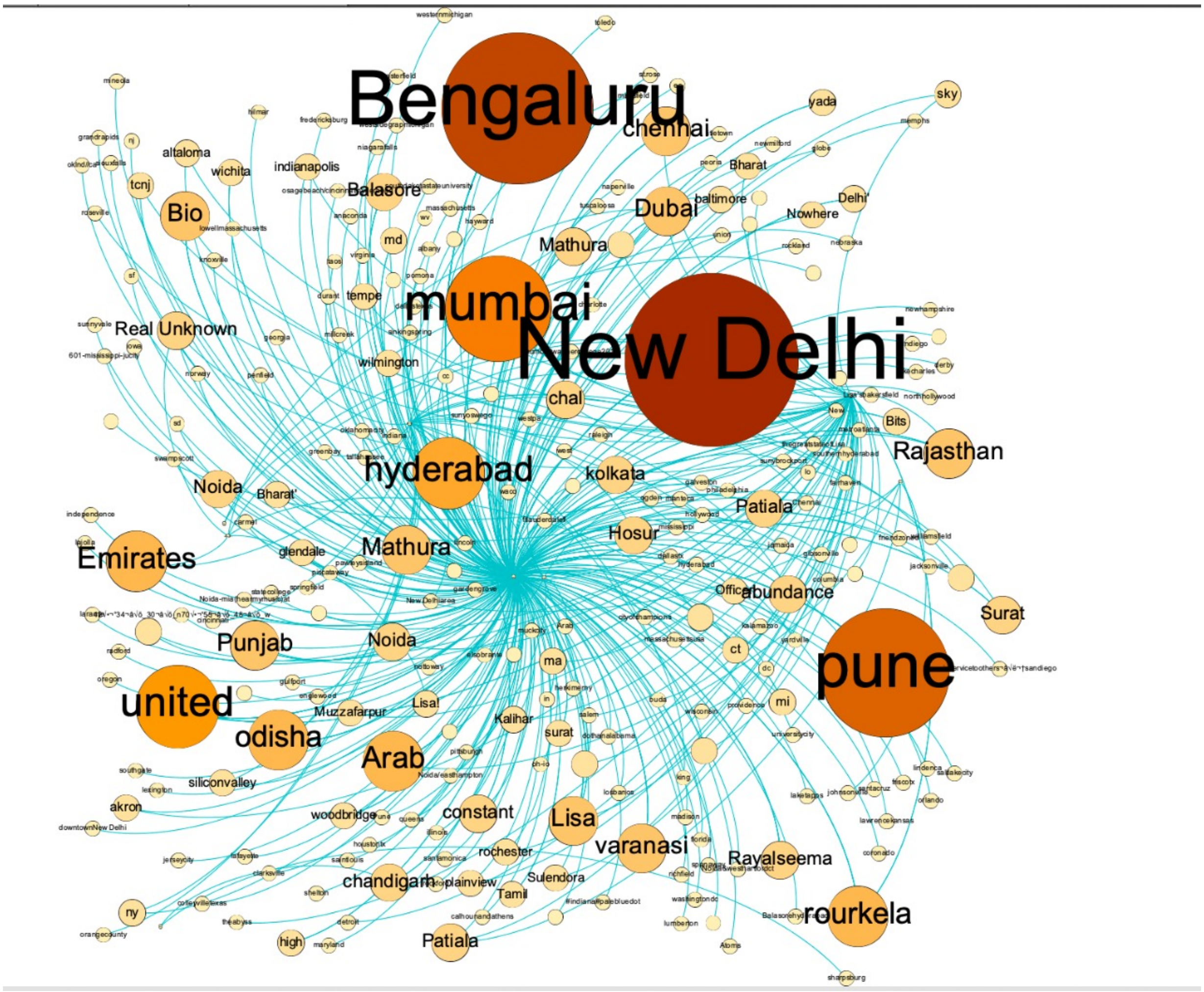
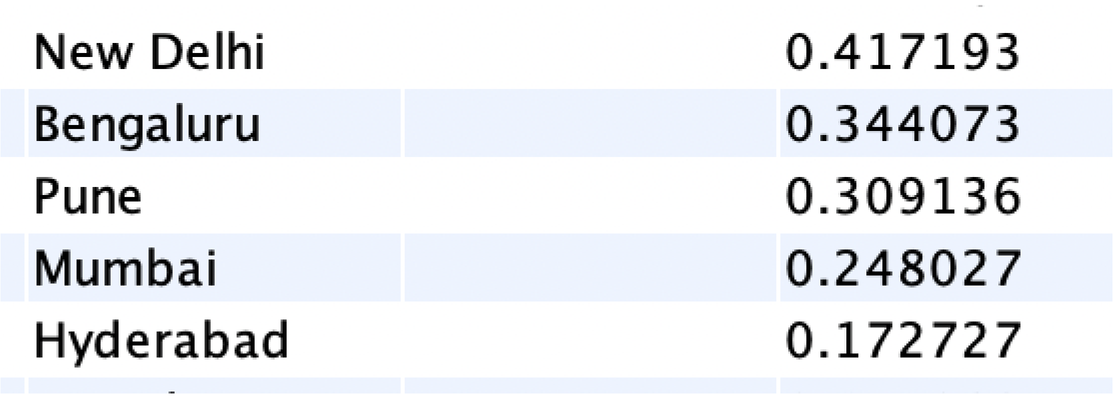


Figure 6.3: Graph for the Second Data-set

If two node had parallel edges between them then they were merged together to improve the visualisation of the graph. These edges were treated separately while running the statistics algorithms on the graph. The observation can easily made by looking at the size of nodes in the graph. The size of the labels were also changes according to the in-degree.

7 HITS Algorithm

Hits algorithm was applied on the graph to get the most optimum location. Since the Authority value of the node is very closely related to the the in-degree of the node. So the nodes were ranked only on the basis of their authority value. As the more authority value will imply that location has the more tweets containing the keywords.
Hits algorithm was applied using the statistics feature from Gephi on the graph formed using the data. The top 5 Authority values obtained from the results for data-set 2 are mentioned in the figure 7.1.



New Delhi	0.417193
Bengaluru	0.344073
Pune	0.309136
Mumbai	0.248027
Hyderabad	0.172727

Figure 7.1: Top 5 Authority values in the graph

8 Conclusion

Conclusion was made on the basis of observations made from the word cloud,graphs and the Authority values of the nodes.
The conclusion is -

- For the first data-set Los Angeles is the best city to open a new pizza outlet.
- For the Second data-set New Delhi is the best city to expand Grofer's.

The conclusion is a rough estimate based on the data we obtained. Some scenarios also need to be considered but this can help in the initial shortlisting of cities.

9 Future Work

Some future work can be done to improve this model get more accurate results. Some the ideas that can be used to achieve this are -

- Same model can be used on a larger dataset of tweets to get more precise and accurate results.
- Same model can be used to get the tweets of the competitor organisation. The location with the most negative reviews of the competitor can be targeted for expansion or for the imporving the already present services.
- The location of the tweets can be made more precise with in the city to get the best locality within the city to open a new outlet.
- Along with the location of tweets the number of similar business can also be taken into account. Then the best location will be the one with the greatest number of interested people and least competition.

10 References

- A Literature Review of Spatial Location Analysis for Retail Site Selection, Omar I. Aboulola, University of Jeddah.
- Location Based Personalized Restaurant Recommendation System for Mobile Environments by kuldeep sing and Anant Gupta, 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI)
- Targeted Marketing by Location from e-Commerce Website Using Web Analytics Data,Zulazeze Sahri* , Roslan Sadjirin, Roger Canda, Faculty of Computer Science and Mathematics, Universiti Teknologi MARA, Cawangan Pahang, Kampus Raub, Malaysia
- Link Analysis: Hubs And Authorities On The World Wide Web,CHRIS H.Q. DING† , HONGYUAN ZHA‡ , XIAOFENG HE† , PARRY HUSBANDS† , AND HORST D. SIMON†
- Dataset-1:https:(//data.world/crowdflower/2015-new-years-resolutions)