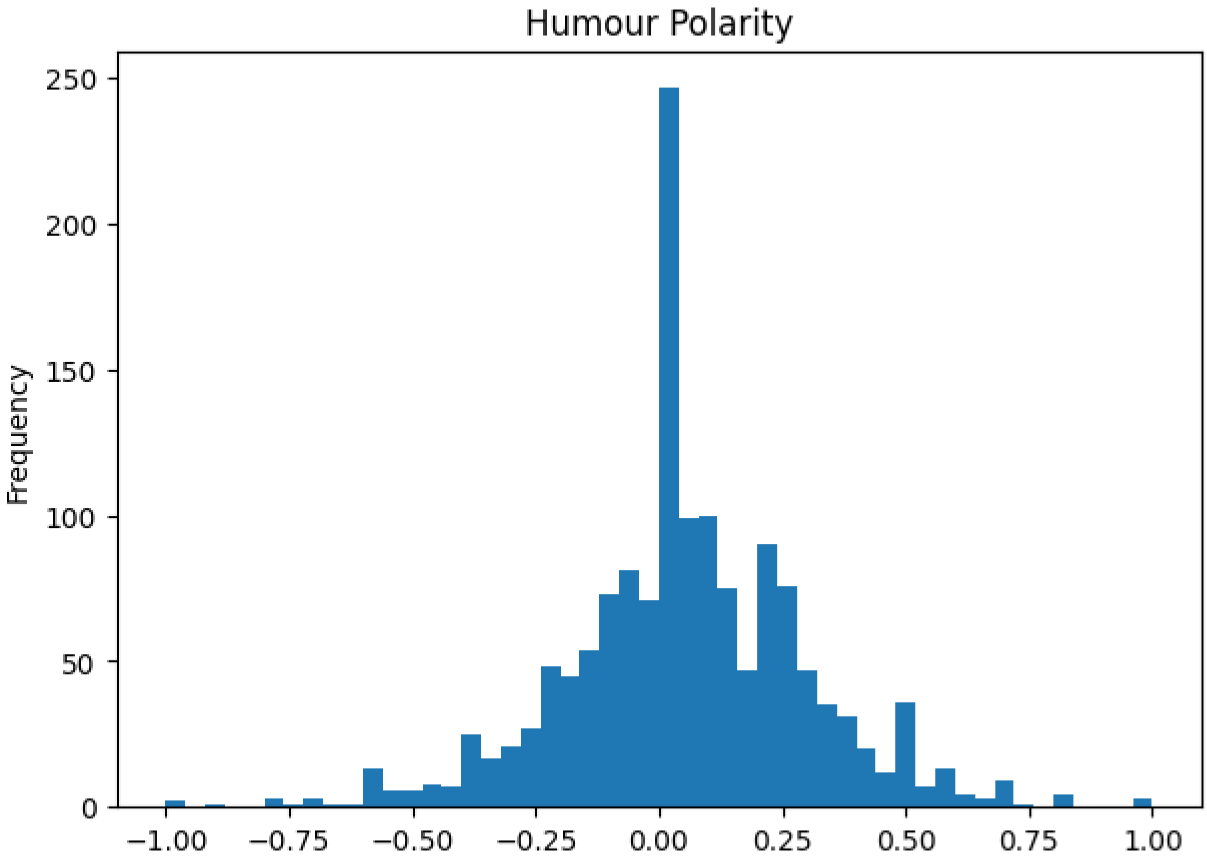
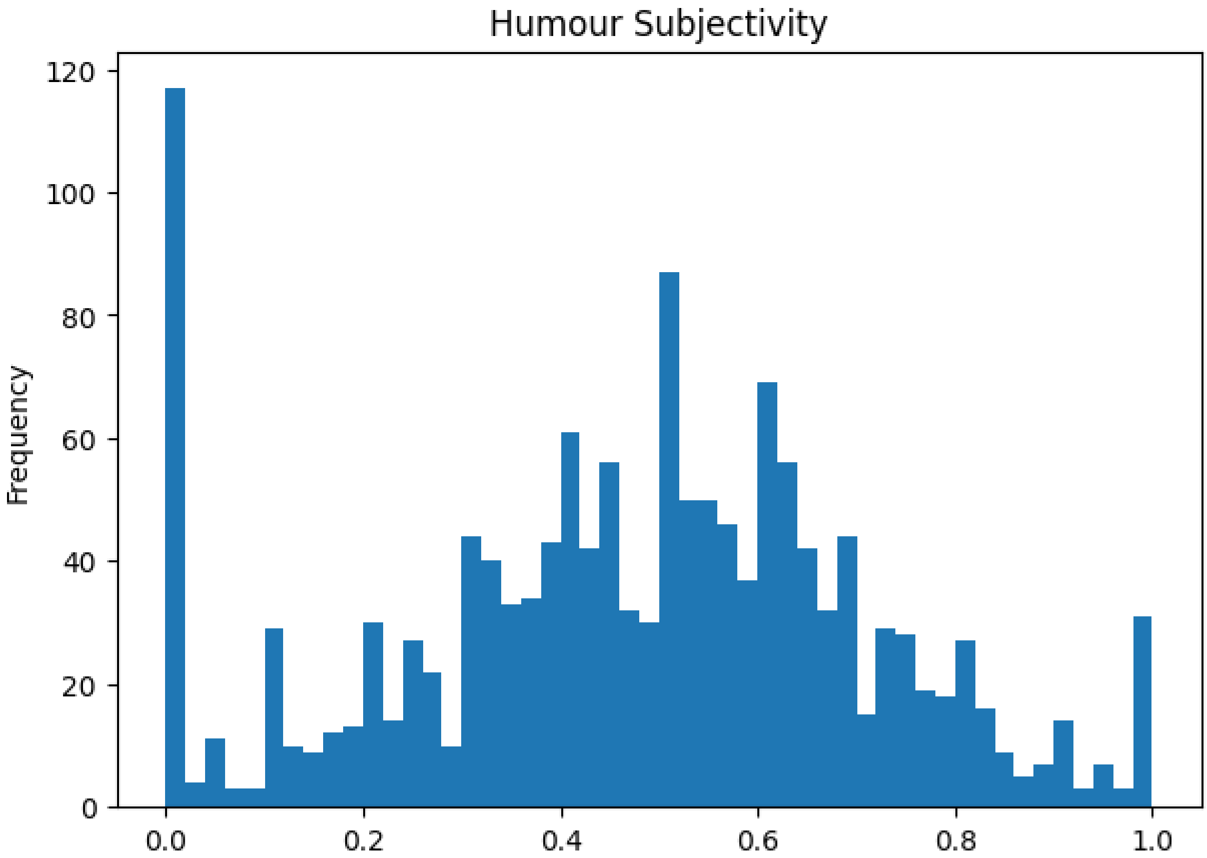
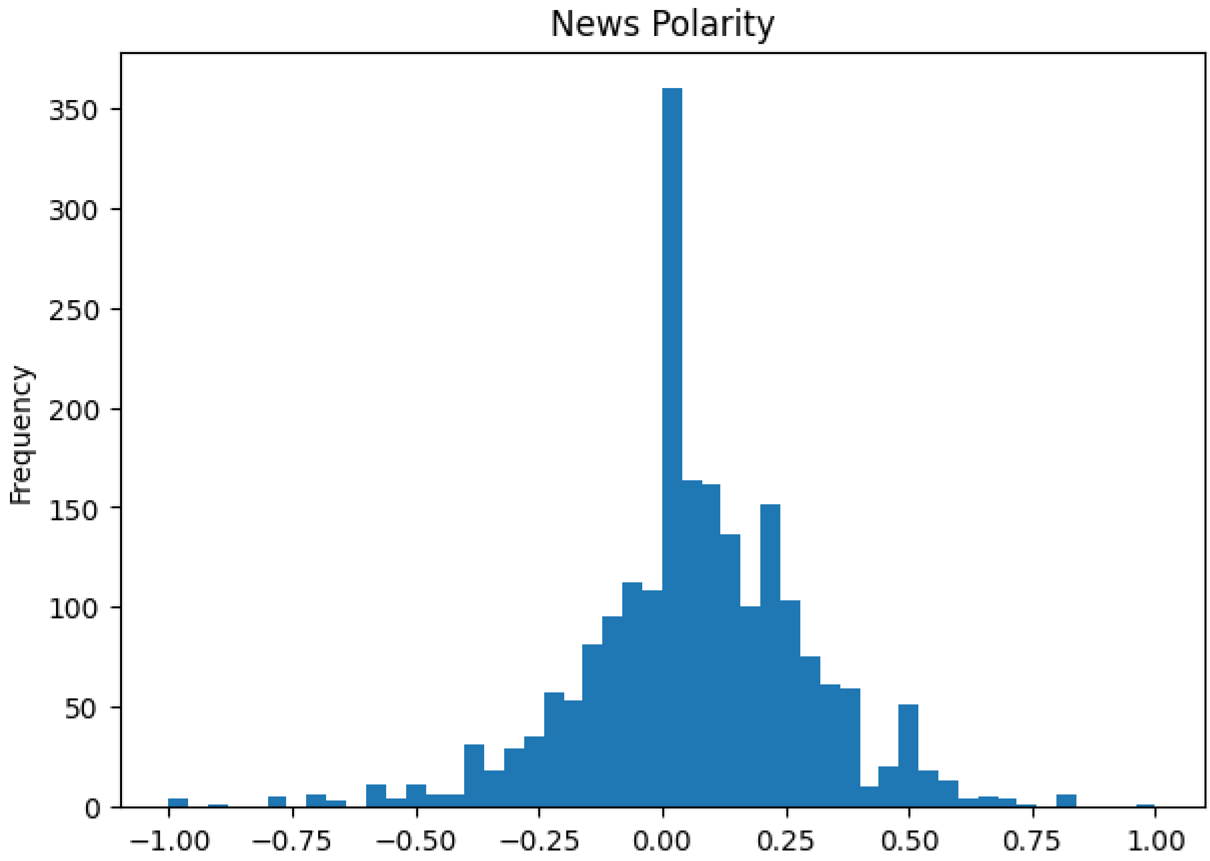
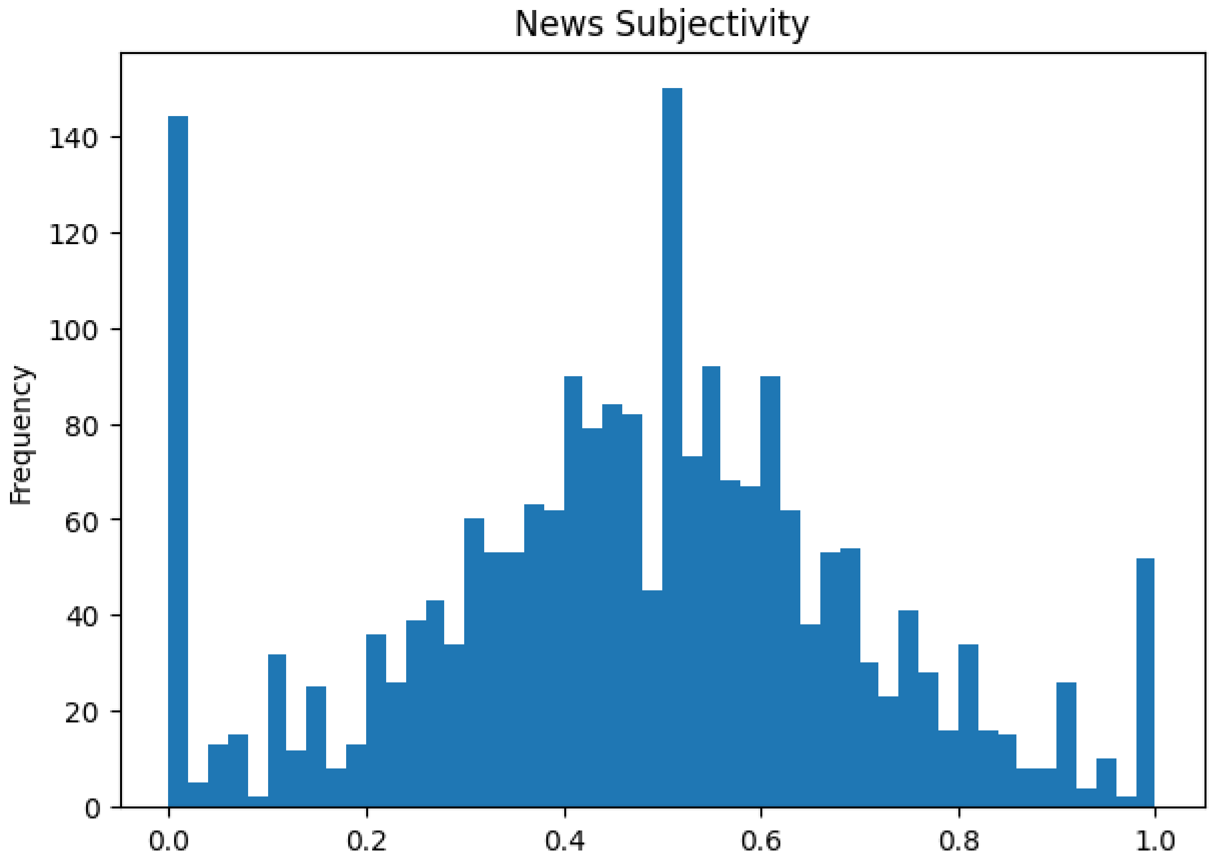
Reddit Data Analysis




Humour Subjectivity
Frequency
0
20
40
60
80
100
120
0.0
0.2
0.4
0.6
0.8
1.0


News Polarity
Frequency
0
50
100
150
200
250
300
350
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00



humour polarity

mean
0.05158728575619279
Standard deviation
0.24866391452327927

humour subjective

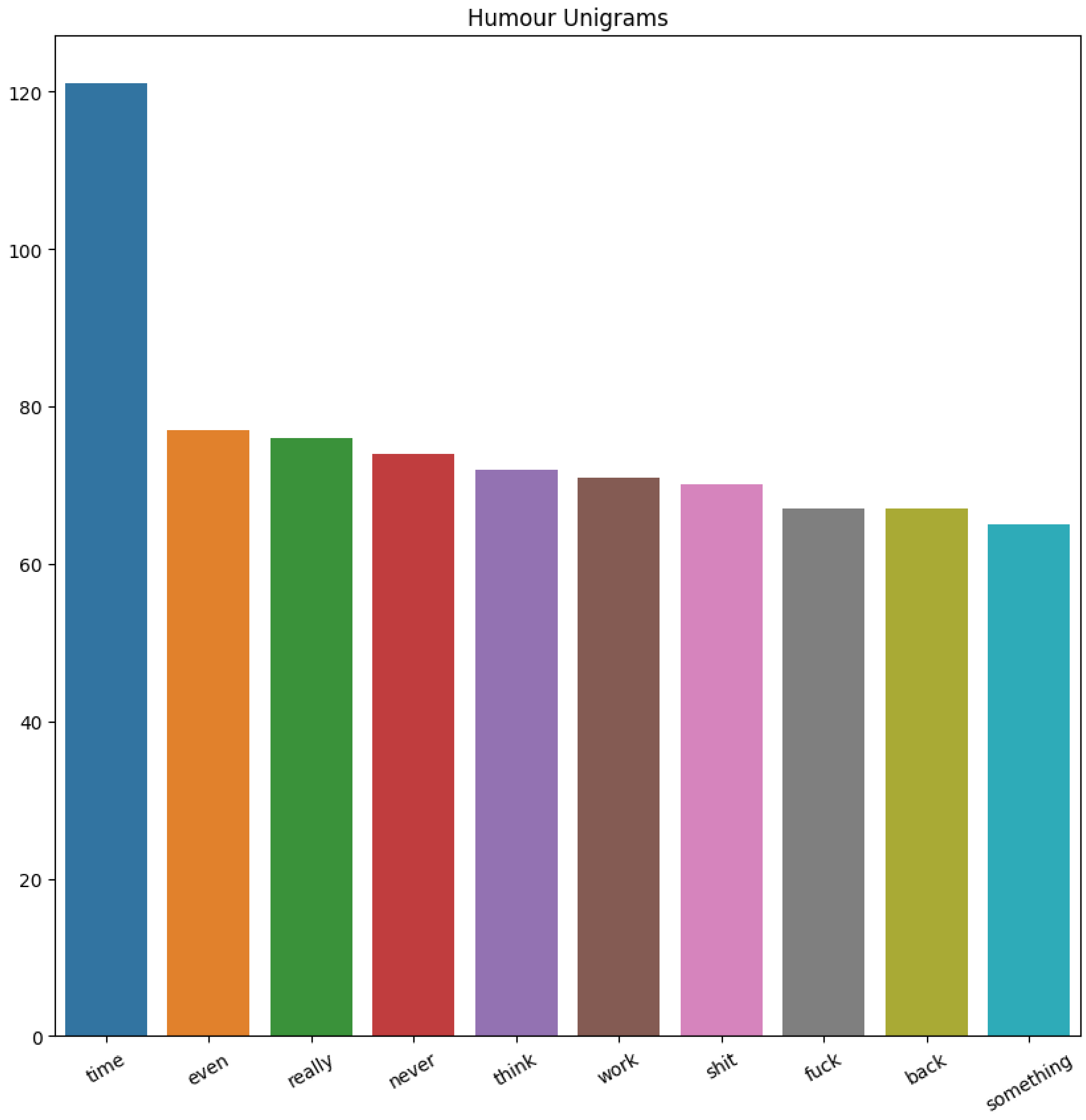
mean
0.47140332998753287
Standard deviation
0.24269324821277793

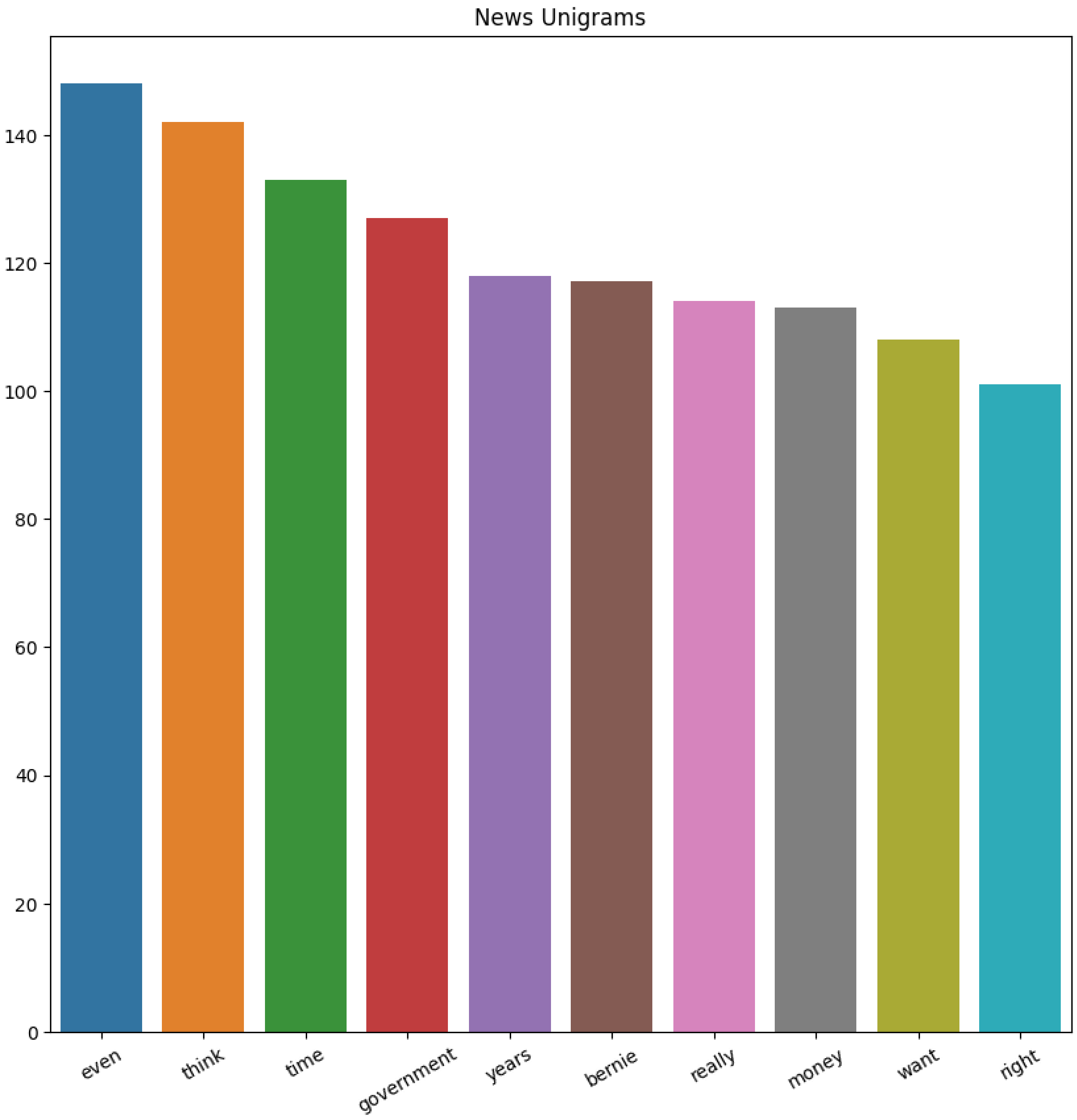
news polarity

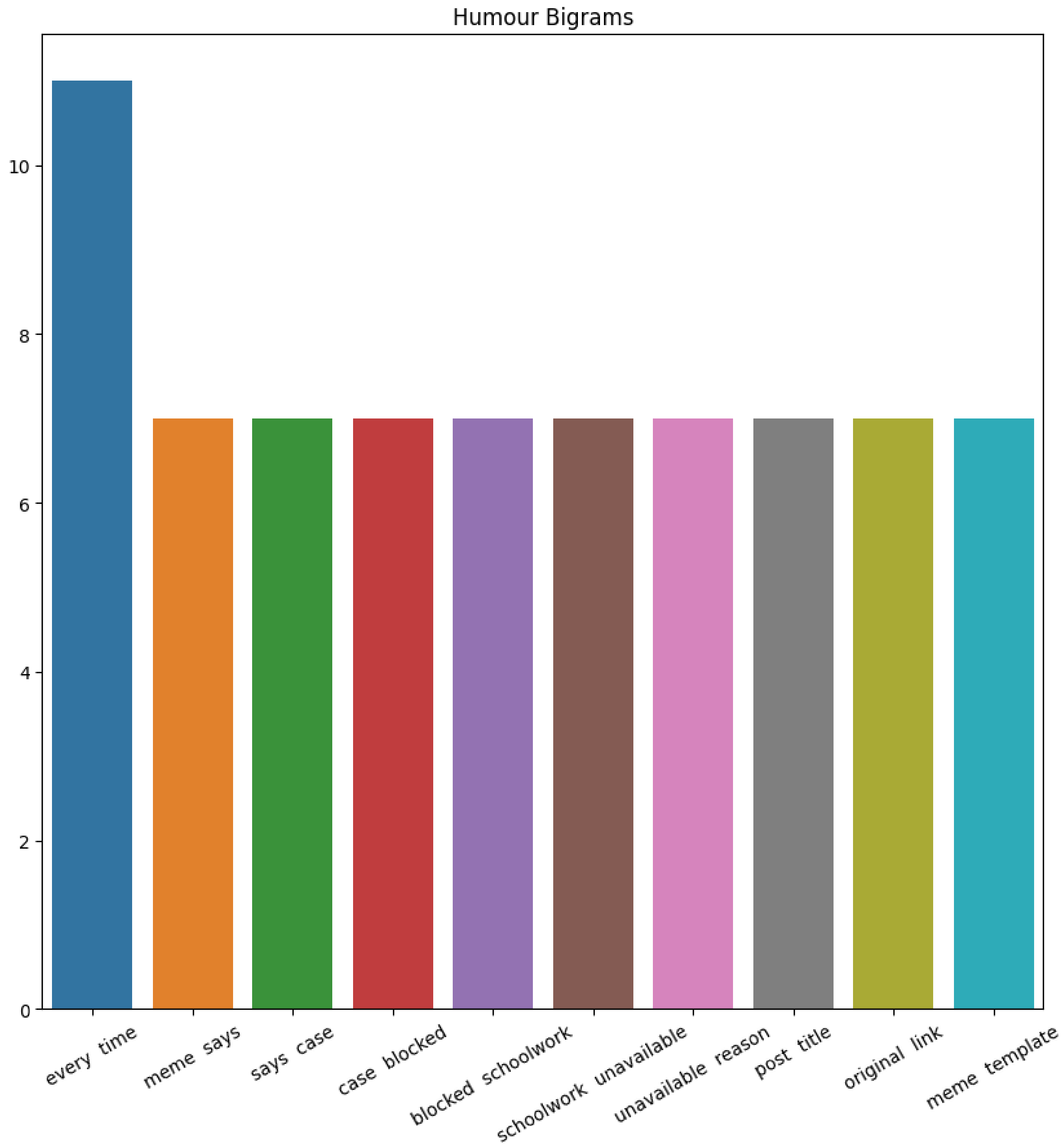
mean
0.064048534801197
Standard deviation
0.23041556585535447

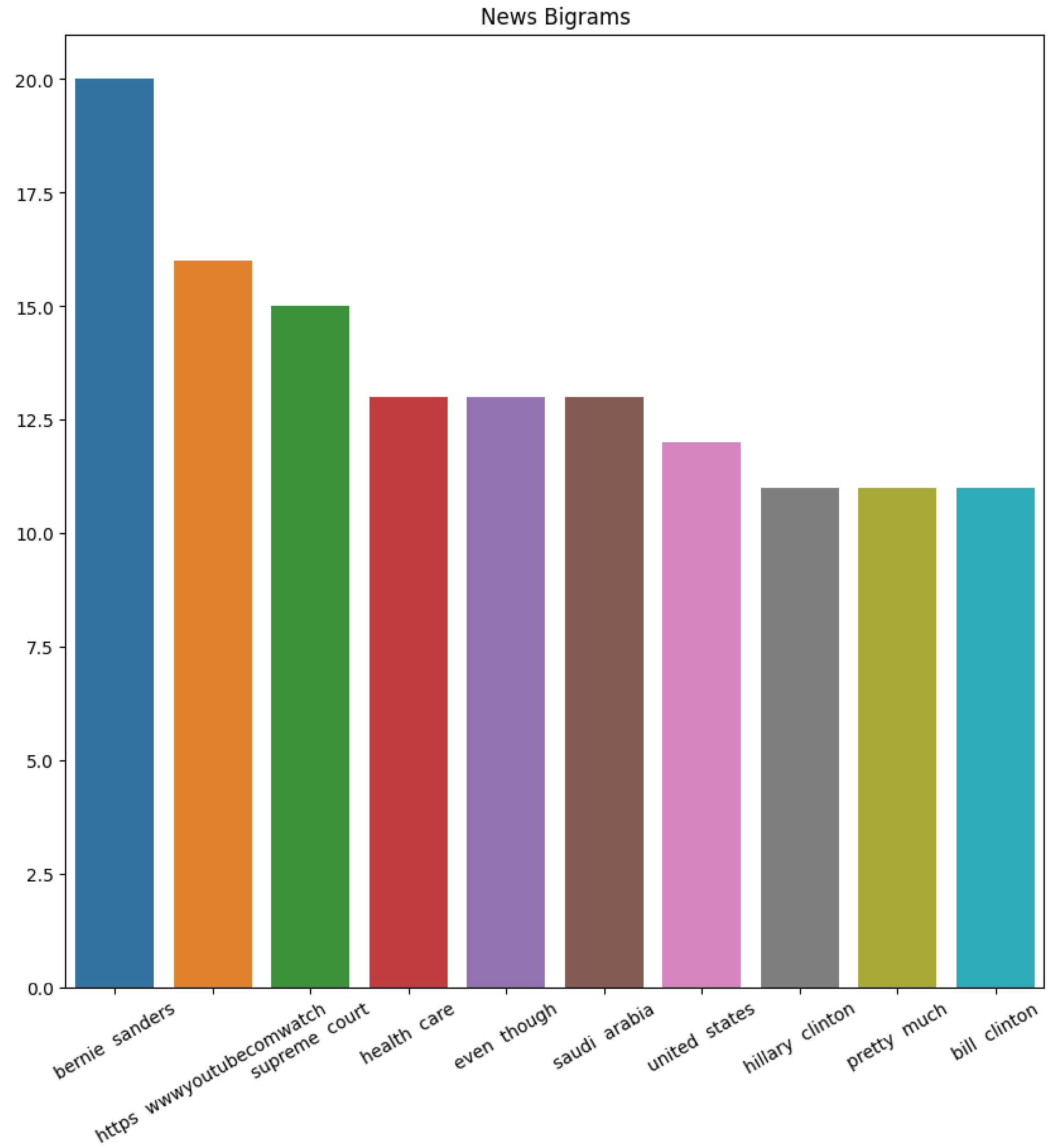
news subjective
mean
0.4704749753548112
Standard deviation
0.23274569397299044

We observe that polarity for both humour and news is around 0 meaning that most of the comments are neutral in nature. As the peak near 0 shows it.
Also the subjectivity of both are nearly equal telling us that news and humour comments both are equally subjective, somewhat less subjective .However news comment should have a higher subjectivity as people give there own opinion and views there not much of that happens in humour.


Humour Unigrams
0
20
40
60
80
100
120
time
even
really
never
think
work
shit
fuck
back
something


News Unigrams
0
20
40
60
80
100
120
140
even
think
time
government
years
bernie
really
money
want
right


Humour Bigrams
0
2
4
6
8
10
every time
meme says
says case
case blocked
blocked schoolwork
schoolwork unavailable
unavailable reason
post title
original link
meme template



Inferences-

In humour unigrams we can find words like shit , fuck etc. these type of words are generally used in jokes and things intended for humour. Words like these are generally used by adults so we can say that kids are commenting less in humour section.

In News Unigrams words like government , bernie , right , money etc. are used news are generally related to governments, money, human rights etc. involved so we can see those words here.

In humour bigrams we can see words like meme template , meme says which refers to memes and memes are a type of humour that is why we can see these bigrams here. Here people are talking about memes and it is obvious as the category is humour.

In news bigrams we can see words like bernie sanders , hilary clinton , bill clinton, united states this infers that news comments are related US elections and there is a lot of comments regarding political news.

Assumption- In bag if words removed words with length less than 3 and stopwords and also few common words like make , will ,people etc.

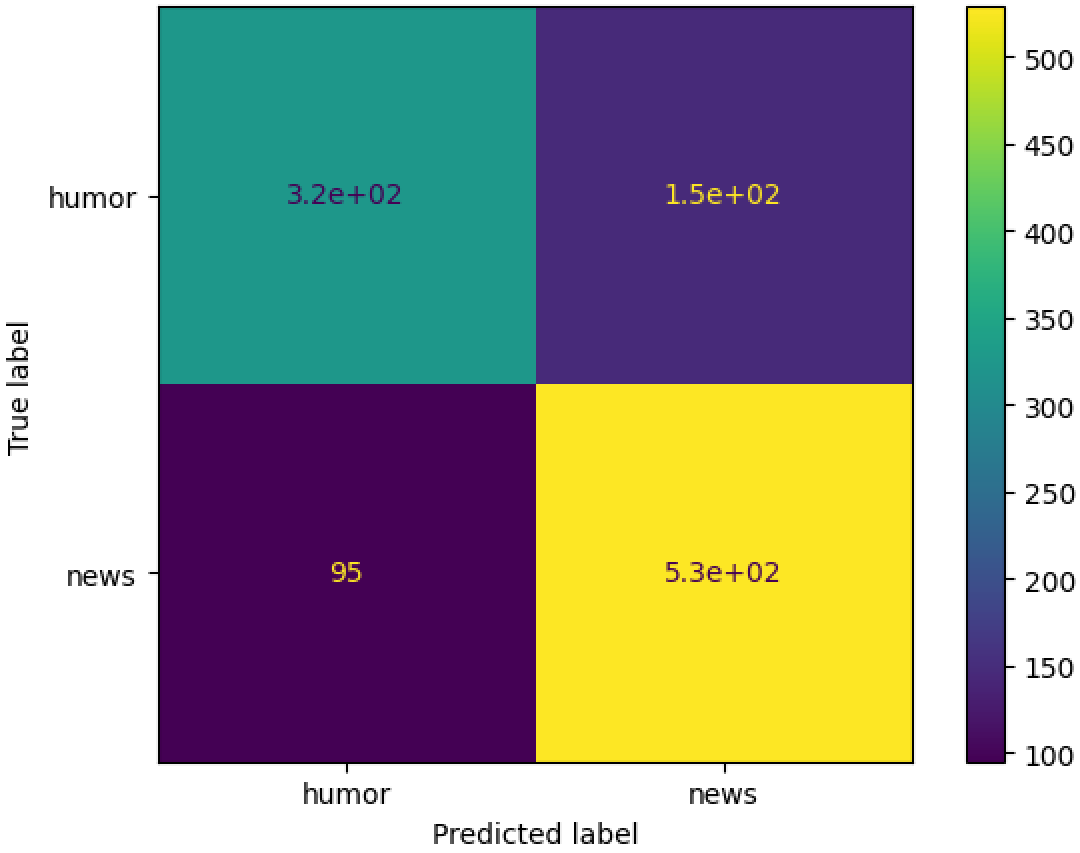
When we look at the data we find that most of the rows are duplicates so we need to delete all duplicate rows because if there are duplicate rows then test train split will be very bad and we will never get a true estimate about the model. Next we will go through the comments and remove all stop words as they are redundant and they don't help us in classification because they are very common. Next we also punctuations as they also don't help us in classification looking at punctuation can little bit tell you about emotion behind the comment but not about what comment is refereeing to.

Accuracy=0.7791970802919708=77.91%

Precision -0.7781769512538743

Recall- 0.7681578661451542

F1-score- 0.7712728527078304



Model Failed-

1) 'it s like my grandma always used to say watching deadpool is better than anal now hit this blunt you fucking pussy naturally she supports bernie sanders ' -missclassified as NEWS

Reason-

This is misclassified as news because of the word bernie sanders here it is used as a joke but bernie sanders is heavily associated with news.

2) 'd e a d p o o l e a d p o o l d a d p o o l d e d p o o l d e a p o o l d e a d o o l d e a d p o l d e a d p o l d e a d p o o d e a d p o o l s a n d e r s' -missclassified as NEWS

This is misclassified because every word has length 1 and there are no words so that using them model can classify the comment because comments don't really have attributes.

3) 'i think it is pretty hilarious that this is being referred to as a conservative seat on the scotus that s not actually how that works '- misclassified as Humour

This is misclassified because it has words pretty hilarious which is usually associated with humour

4) 'she ll be jailed only if there are other factions with enough sway that want to see her jailed it s looking like the intelligence faction wants that i m holding out hope jailing her wo nt really change much in and of itself but it s a nice dream ',- Misclassified as Humour

This is misclassified because this doesn't have any word that could be associated with news like something political it might be the case the she here is some political person that why comment in news section but model cant figure this out without having a proper name .

5) 'jesus christ i ca nt think of a better thumbnail than that for this post he s like i m just saying you know ' -Misclassified as Humour

This is misclassified because the comment might be on a post that is a news article or something related to news but simply having comment without any post associated with it makes model think it as humour since there are not enough words to classify it as news.

Model Passed-

1) 'holy fuck nuggets this david wolfe bastard again all the older women i work with share his asinine propaganda all over facebook it consumes me with a special kind of hatred '- Correctly classified as Humour
   We can see words like holy ,fuck, bastards and these are associated with humour hence correctly classified.

2) 'it s fun that the government has turned the words freedom and patriot in to code words for tyrannical and selfish ', - Correctly classified as News
   Words like government , tyrannical, patriot and freedom makes the model understand that this is a news comment as they are used in news comments mostly.

3) 'this is a much bigger deal than it would appear access to modern tech and resources could spark a boom in a country that is in desperate need of infrastructure that is known for the industrious nature of its workers and that has massively depressed labor costs it could be great for cuba and the north american economies '-
   Correctly classified as News
   Words like north american economies , country ,resources , industrious give us a hint that it is a political news related comment and these words in our bag of words are associated with news comment.

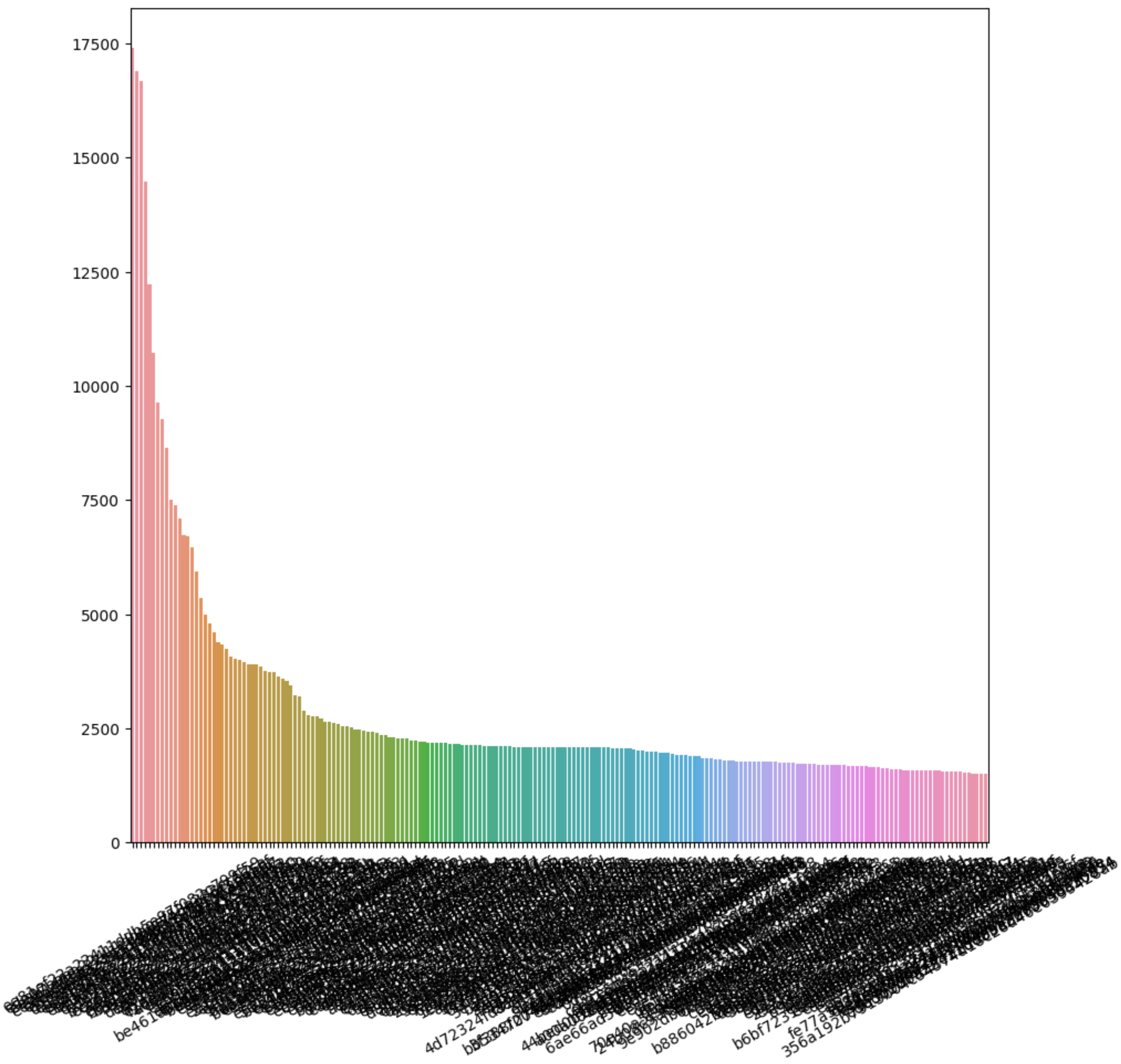
4) 'sounds like a poorly thought out idea do nt give control to random accounts either find more mods to help or keep it the way it is ',
   Correctly classified as News

There were no words like shit, fuck etc. that are associated with humour and accounts is something that can be associated with the news.

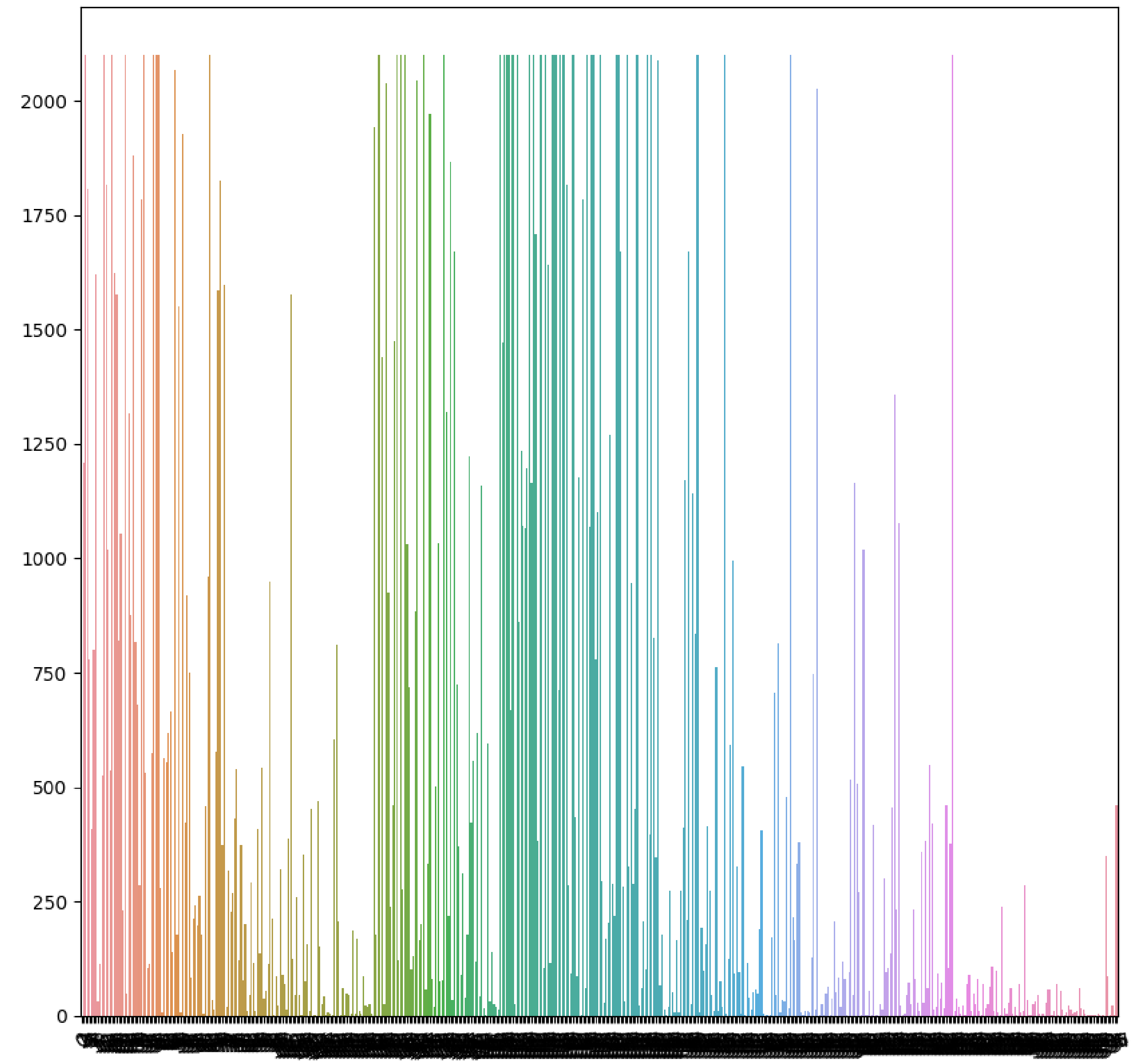
5) 'i m sorry to say but you re an idiot and this is actually your fault you did nt feel appreciated for months so you went way over the top on valentine s day that s on you reading about your valentine s day stunt made me feel nauseated in the future do one thing on valentine s day her favorite restaurant or handmade chocolates and if you write a letter burn it '
Correctly classified as News
Here idiot is giving an insight that why this is classified as humour.

BrightKite Data Analysis

Location Plot-

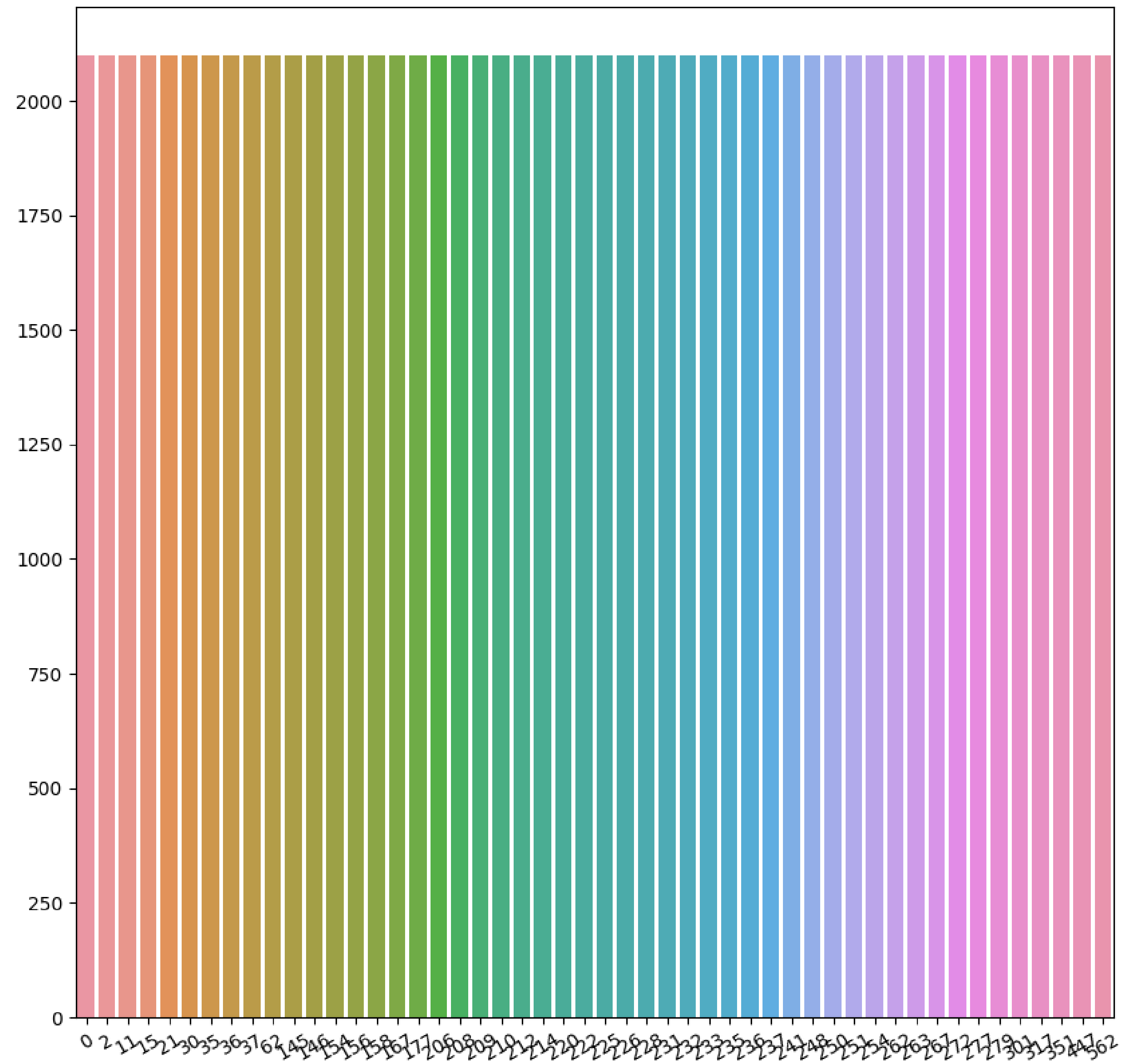

17500
15000
12500
10000
7500
5000
2500
0
be461
4d72324f
6ae66ad
b886042
b6bf723
fe77a
356a192b

User Plot-



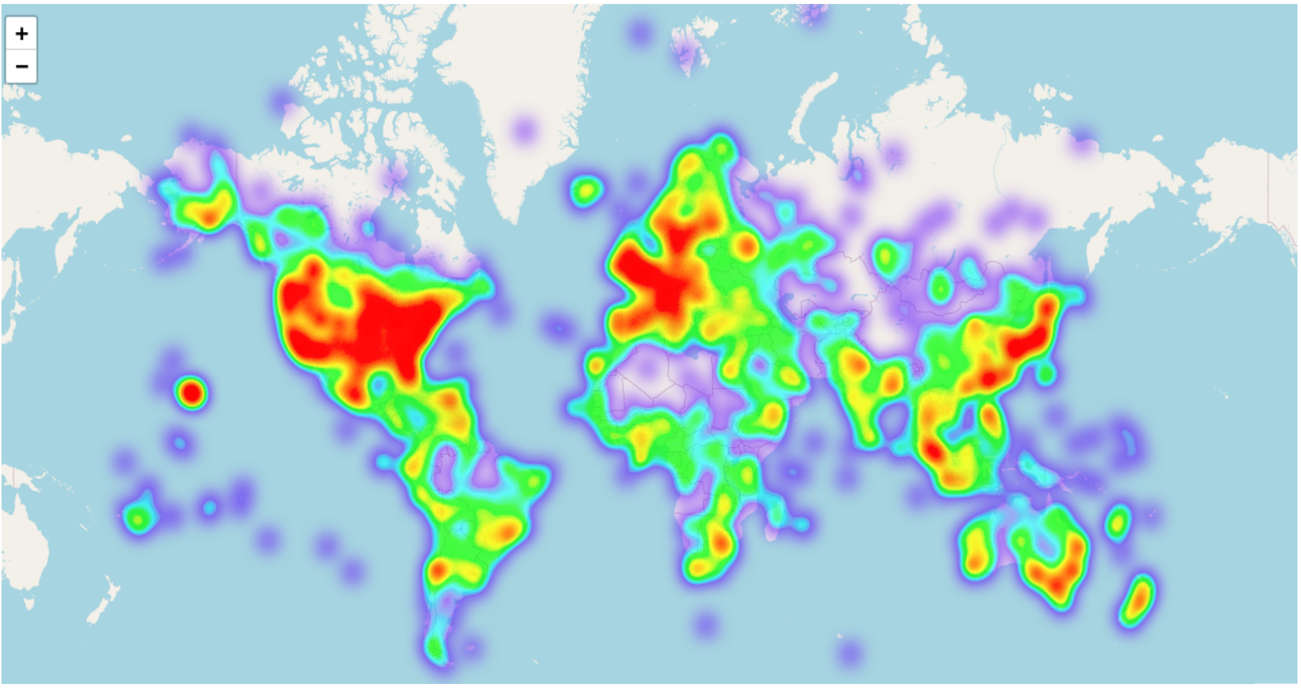
Assumption- Above graph is plotted for the first 500 users.

Top 200 users graph-



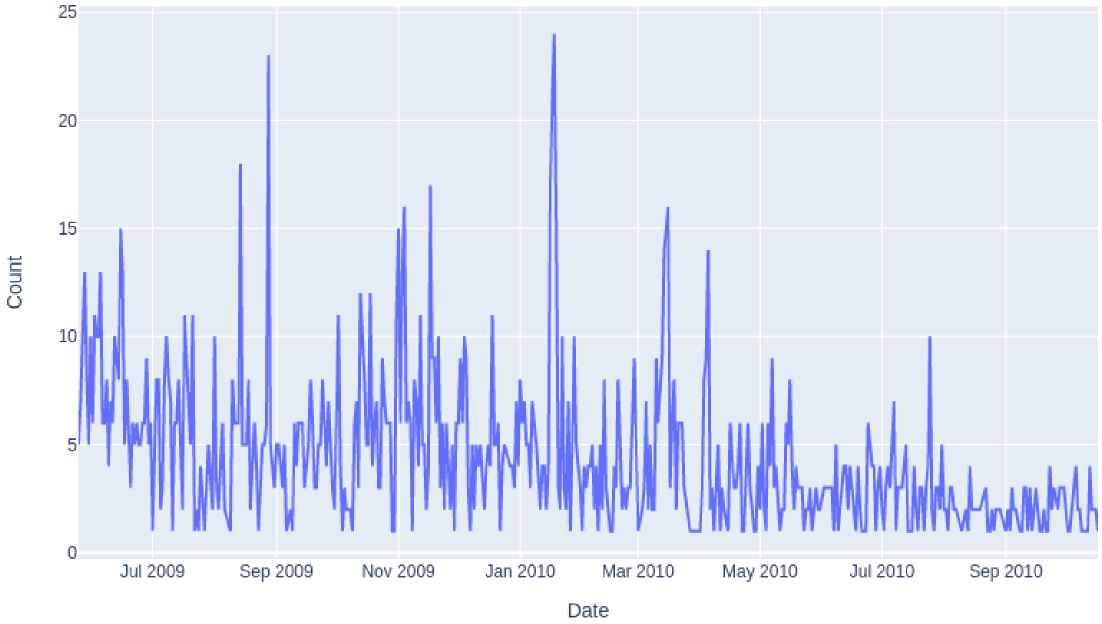
Inferences-
In location graph we can see that first few locations have the most numbers of checkins it means that app was quite famous in few locations however it was not very famous globally as if we look past approx top 10% of the locations the number of checkins were significantly lower.

If we look at the user graph we will find that many users have a value 2100 which might be the limit of check ins per user . Here also we can see that most significant part of check ins are contributed by the top users.



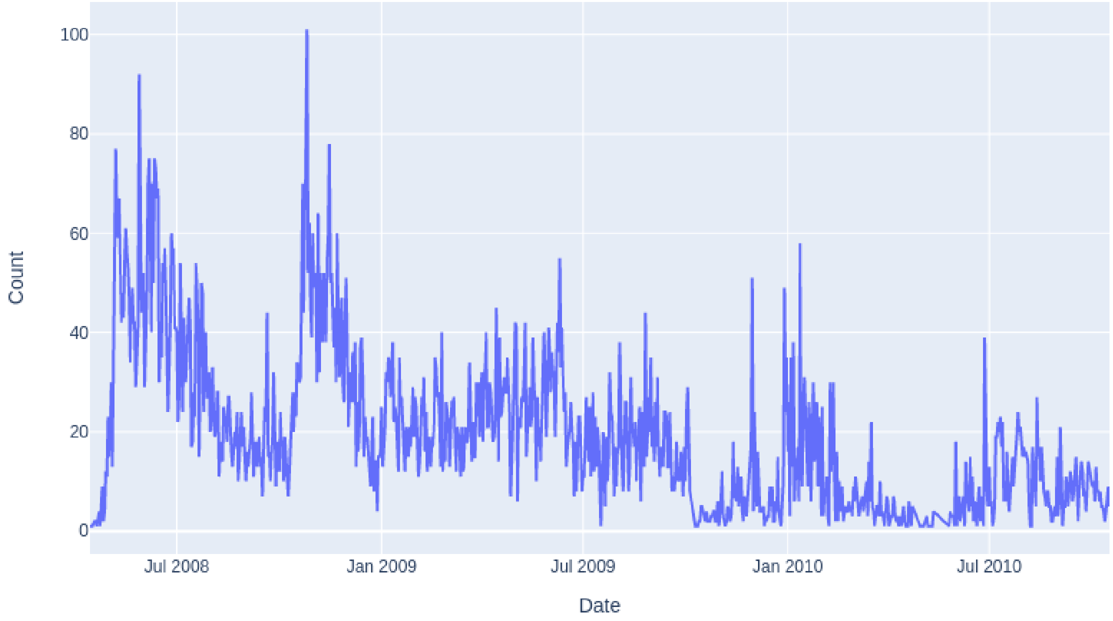
After looking at the heatmap we can get an idea of the global reach of the brightkite . We can see that Russia , China , Some parts of Africa have very little checkins this infers that there are less users in these countries and also brightkite users don't visit these countries often. These types of heatmap can be very helpful to look out the global expansion of service and also target specific location. Most heated part of the map is in USA we might infer from it that the app might be founded in US as there are large number of users in US.

For User-



We can see that for this user nearly every month there is a peak at least once. Which means that user has a increase in check ins every months that could be because he travels to new location every month. Also those peaks are decreasing with time meaning user is travelling lesser than he used to .

For Location-



In this graph of Location we can see that there are peak near Jul 2008 and Jan 2009
Which means that there were greater check ins at that location at that time of year this
Might be because this location becomes a tourist attraction at that time of the year similar pattern can be seen in 2009 and 2010 .
Temporal information can be used in many ways to target users like having once temporal information you can show them targeted and personalised ads one the basis of their location and also temporal information can be used to predict future movements and behaviour of the user which can pose a security and privacy threat.
Yes there is periodicity .

We can leverage the information in this dataset to target users on the platform as in-

1) We can build a recommender system that recommends users places nearby to visit based on the number of checkins in those places.
2) We can look at the heatmaps and many such graphs to understand global reach of the app and advertise our app more in the areas with less check ins.

3) We can also try to predict places where users will go in future like tourist attractions and using that predictions we can target those specific location for more sales.
4) We can look at the locations a user visit and according to that show him ads relevant to him and the location he is in like restaurant and hotels ads .
5) Some locations in the datasets were invalid which can be due to fake user profiles we can keep check on these user and verify them.

The challenges in terms of privacy and security in the context of location-based social networks like Brightkite are-

1) If the location data of a user is in wrong hands it can cause serious physical and mental threats to the user like kidnapping, stalking etc.
2) Companies can use this data to make users see ads based on the location they are in this can cause irrelevant ads like I’m at a relative's place to stay so a hotel ad would be irrelevant .
3) Companies like brightkite can sell your precious location data to third party apps without your consent.
4) Also with new machine learning techniques your location data can be used to predict your future movements and behaviour .
5) If someone steals your phone and visits to some illegal place you can be in trouble since it would be hard to identify that it wasn't really you.

Username Analysis of Users in Different Social Media Platforms

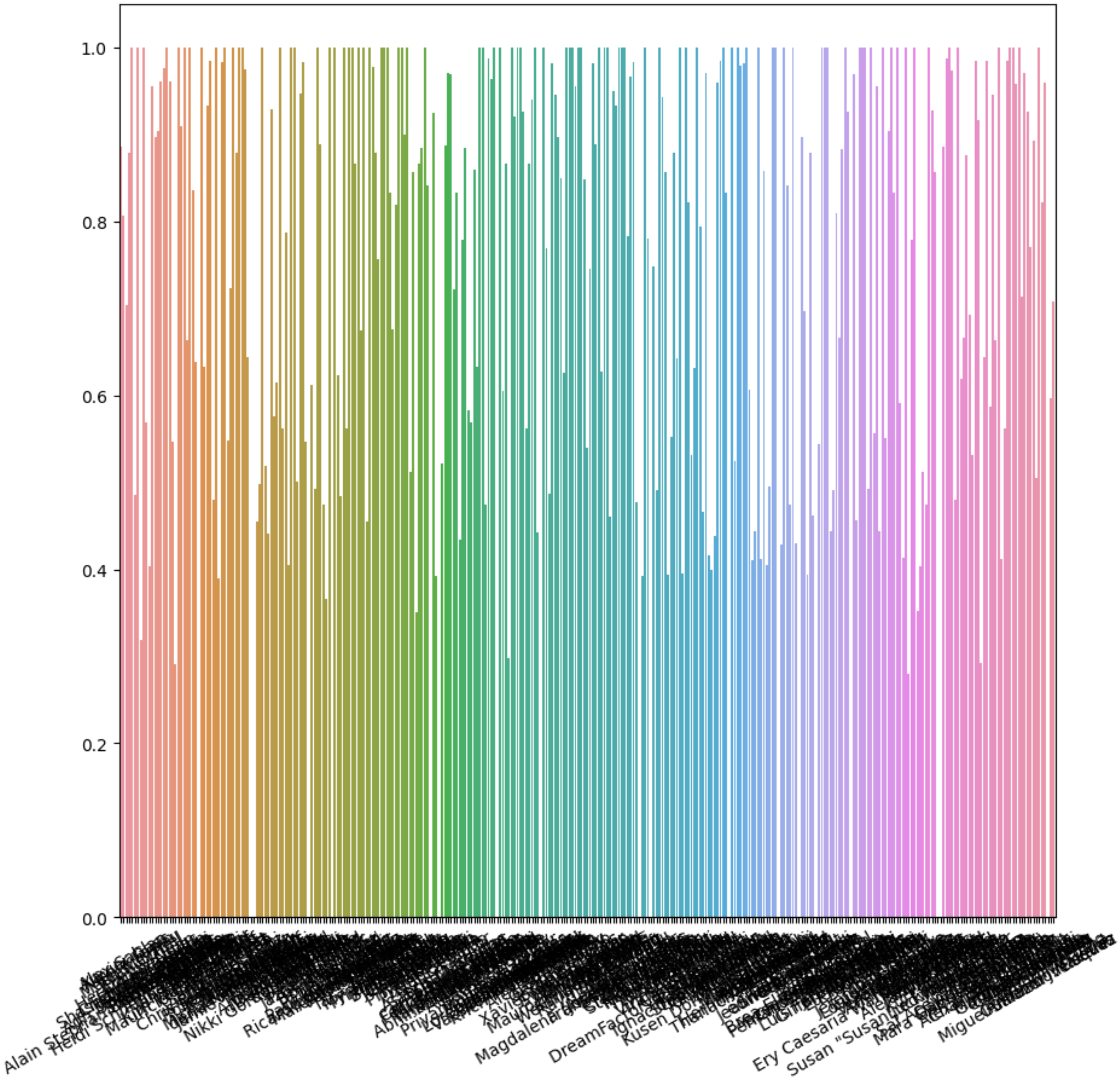
i) Twitter-Facebook Identity Resolution -Metric used Levensteins and Jaro Winkler best metric is Jaro Winkler . It is good because even when in username there were characters removed and jumbled it performed better than Levenshtein. For example - alainstephan","alainpato” levenshtein gave 0.5 and jaro winkler gave 0.886
This fact is further proven by higher mean of jaro winkler(0.74) array than levenshtein(0.58).

ii) Facebook-Instagram Identity Resolution-Metric used Levensteins and Jaro Winkler best metric is Jaro Winkler. It is good because even when in usernames there were special characters added and characters removed it performed better than Levenshtein. For example - "alexsablancom","a_sablan" levenshtein gave 0.53 and jaro winkler gave 0.73.
This fact is further proven by higher mean of jaro winkler(0.78) array than levenshtein(0.62)

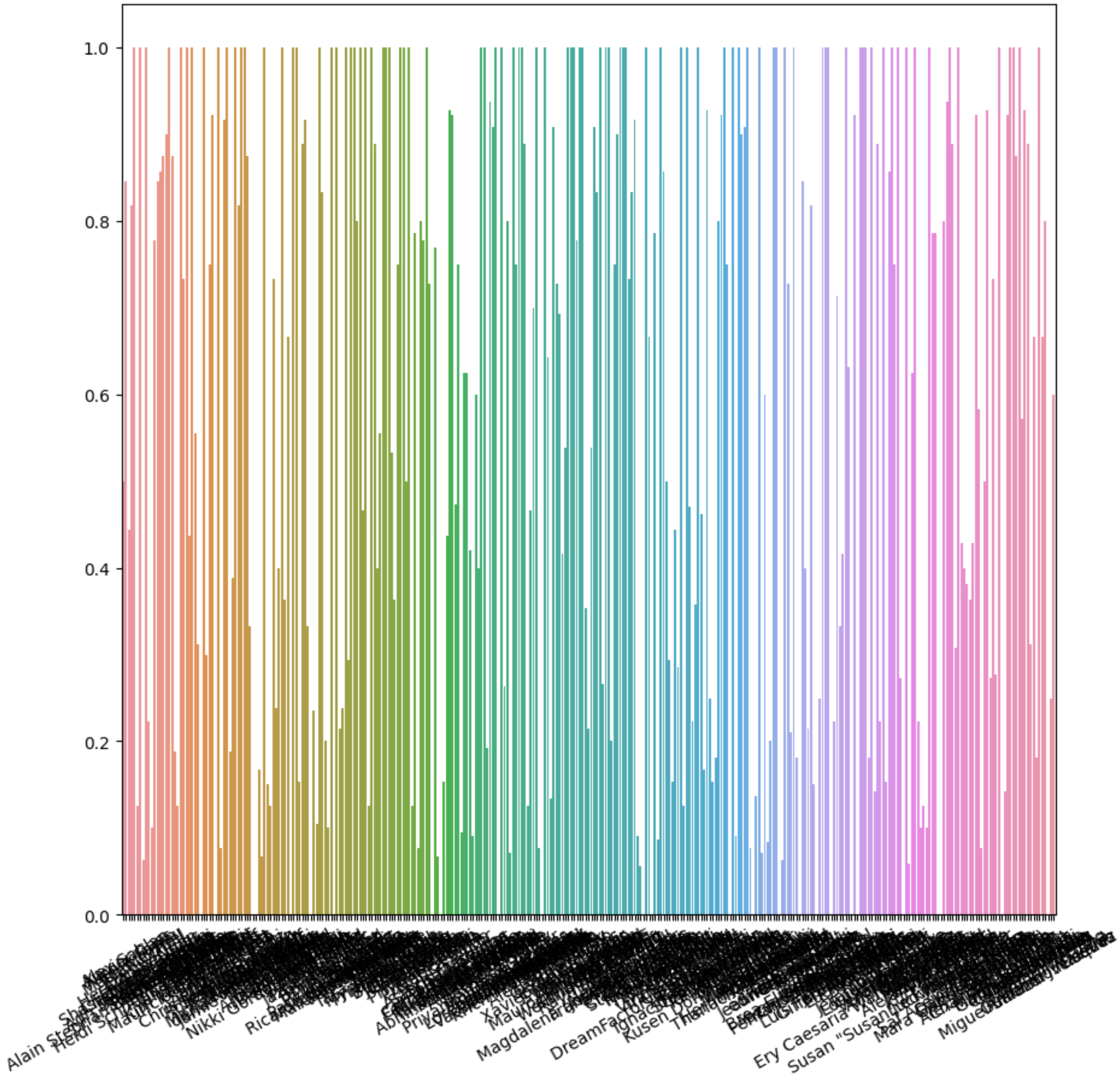
iii) Twitter-Instagram Identity Resolution-Metric used Levensteins and Jaro Winkler best metric is Jaro Winkler. It is good because even when in usernames there were difference in cases it performed better than Levenshtein. For example -"XaviGasso ","xavigasso"
levenshtein gave 0.7 and jaro winkler gave 0.82.
This fact is further proven by higher mean of jaro winkler(0.83) array than levenshtein(0.74)

Twitter-Facebook Identity Resolution Plot-

Jaro Winkler Metric plot-



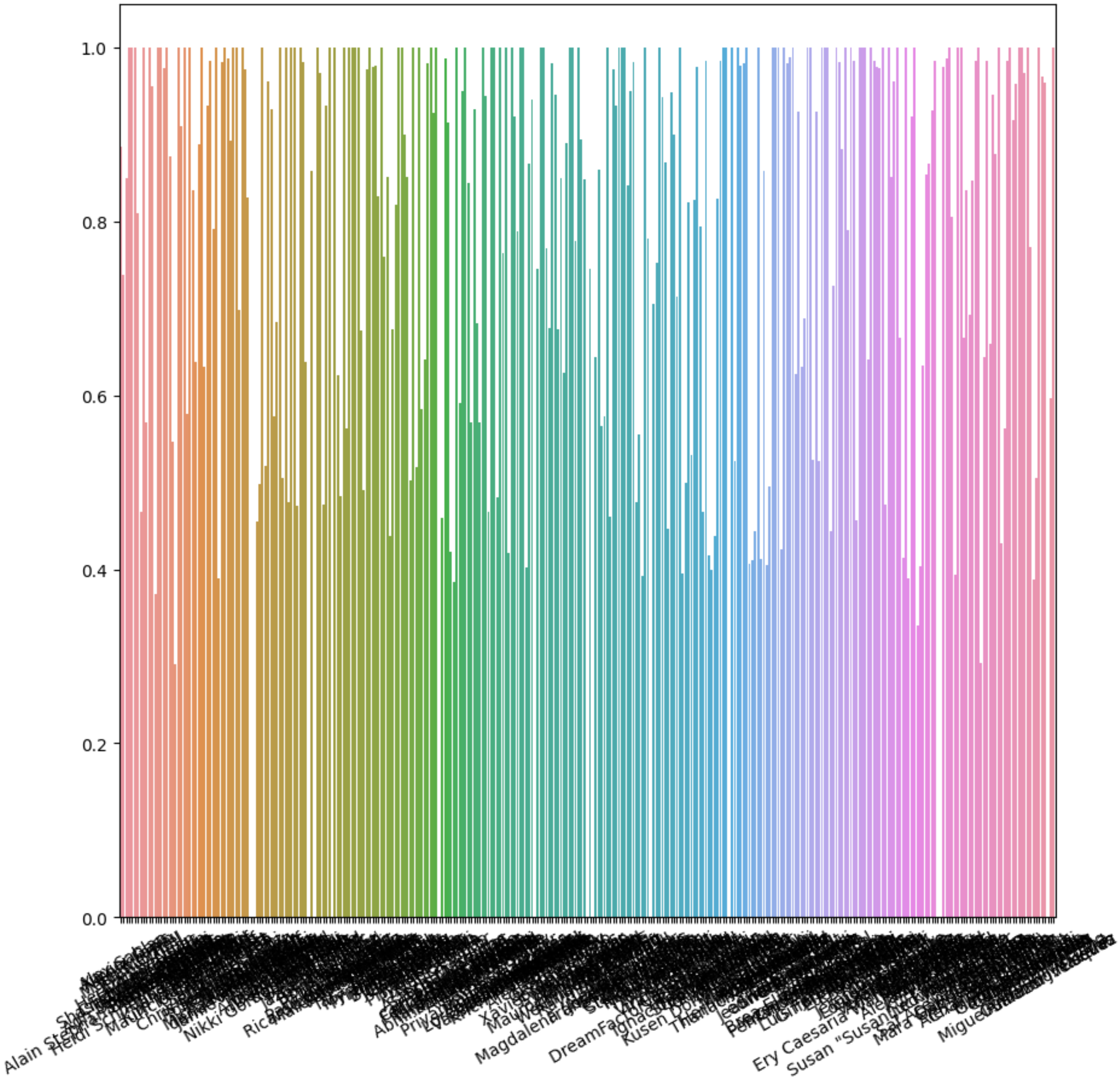
Levenshtein Metric Plot-



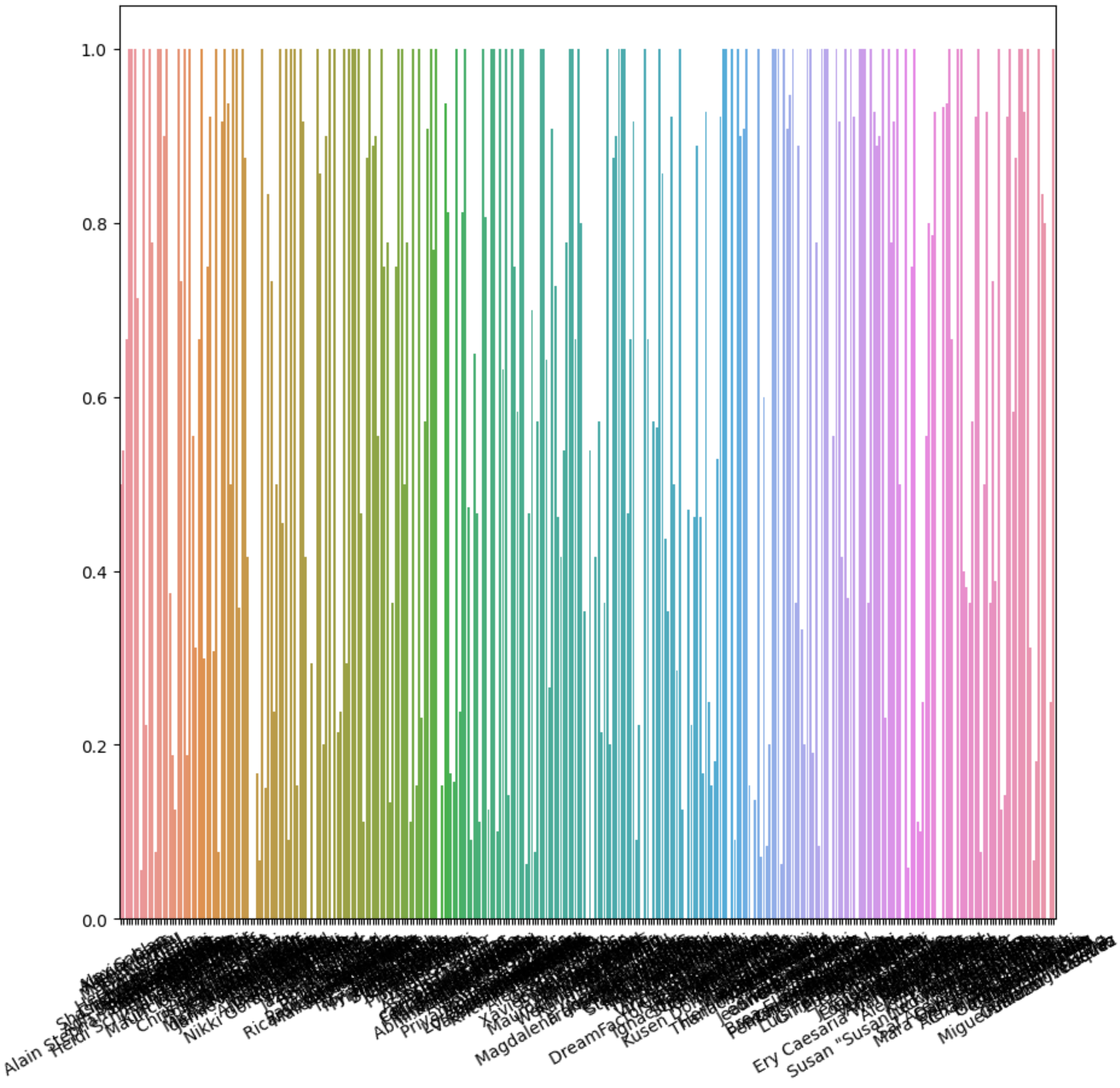
Inference- Observing the high peaks in the graphs here are we can see that most of the usernames have a high similarity score that is 184 users have 0.8 or above jaro winkler metric score this means that lot of People like to keep quite similar usernames. Also lesser peaks in levenshtein metric plot is expected evident of the lower mean than jaro winkler.

Facebook-Instagram Identity Resolution Plot-

Jaro Winkler Metric Plot-



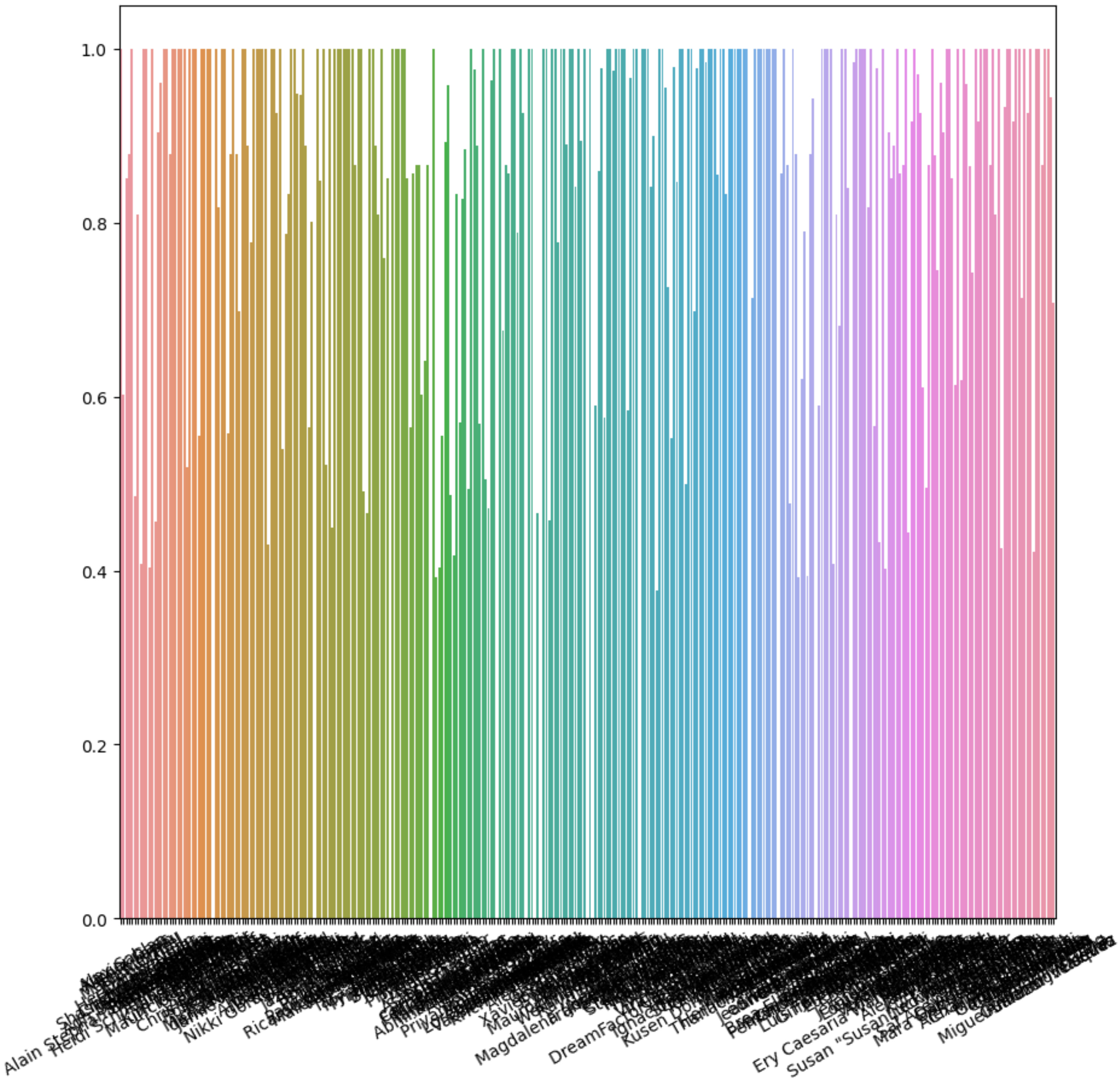
Levenshtein Metric Plot-



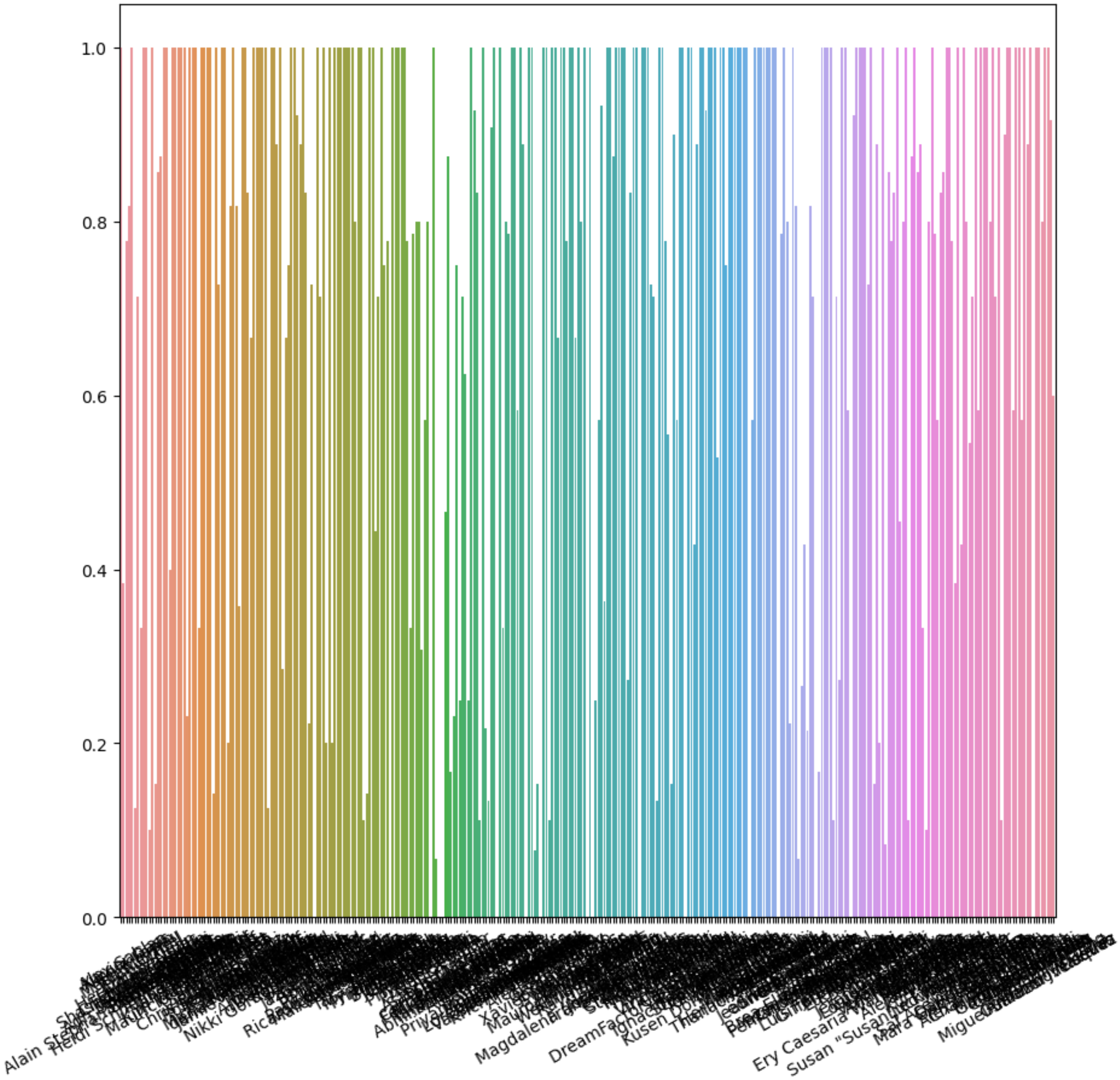
Inference- Looking at the graph here we can also see that most of the usernames have a high similarity score that is 199 users have 0.8 or above jaro winkler metric score this means that lot of People like to keep quite similar usernames. And also high mean of 0.78 shows the same thing.

Twitter-Instagram Identity Resolution -

Jaro Winkler Metric Plot-



Levenshtein Metric plot-



Inference- Very high mean of 0.84 and 241 people having similarity jaro winkler score o.8 or above tell us that for twitter and instagram most of the people keep most similar user names .