**Product Perspective Discovery and Evaluation by Analyzing Tweets**

**PROJECT REPORT**

Guided by:

Prof. Rong Liu

**Group 2**

Gaurav Sawant

Geetkaur Sukhmani

Omkar Desai

Sayan Mukherjee

# Introduction

Social Media has become this destination for people to not only interact with each other but present and discuss their thoughts and opinions about products. A wide range of people post comments about products. They are either individuals associated to an organization to promote their brands, reviewers and experts posting their take on the product or end users who are a part of the consumer base. The consumers could be prospective, seeking to buy the product giving us an insight on the demand of a particular product or existing ones who want to share their experiences.

While this presents us with a variety of information, looking at all of them together can be overwhelming. Additionally it is difficult to separate the useful information from noise and regular chatter. However, if we are too myopic and narrow with what information we collect, it could result in heavily biased data and uni-dimensional information.

This presents us an unique opportunity to gather results on the basis of certain keywords. For the purpose of this project we have gathered tweets leveraging the Twitter API. We targeted three keywords - “iphonex”, “iphonexr” and “iphonexs” associated to three of the current popular mobile devices by Apple - iPhone X, iPhone XR and iPhone XS, respectively.

# Motivation and Objective

Our intention with this project is to demonstrate how businesses can use a similar approach to gain insights about the perception of their product in the market. It can further help organizations understand the kind of demand or hype around their products and also, what are some of the prominent features which are generating interest. These themes discovered by our approach will help businesses invest in areas worth investigating.

The comments need to be processed and cleaned to collect those that are relevant to the product or brand itself. This will need techniques of Data Mining and Natural Language Processing to be employed and to classify the comments and break them down into clusters.

The comments either seem to align with the opinions of the reviewer, disapprove of the review or mention a competing product that they deem to be better. This gives us an opportunity for Opinion Mining and Sentiment Analysis to be done on these comments. At the very least, it will give us a sense of how the particular product is being received by the consumers.

Also, we intend to highlight the features that consumers like the most or dislike the most so that the company can work on it’s products more proactively. Another aspect we intend to address is the perception of the brand. Comments are valuable to understand what is the common perception and how the brand is seen by the consumer-base. Sometimes these comments can also help us realize what are the competing brands or products in this segment.

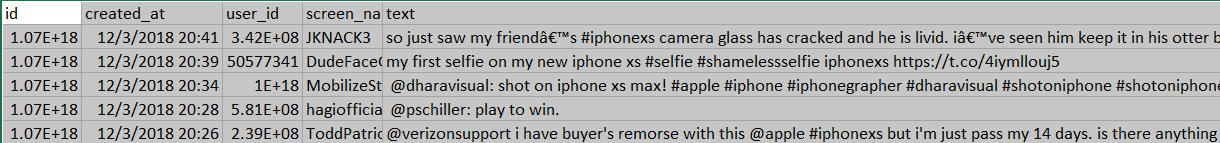
# Fetching Data

We have used the Python tweepy package to get tweets via the Twitter API. The tweets being fetched are done so using the *search* API endpoint that responds with tweets matching a certain keyword.

Our script continues to poll the endpoint and keeps getting tweets till it matches the number of required tweets (10k in our case). So we end with three distinct data sets for the three keywords. We will analyse these data sets independently.

The script then writes the tweets in a CSV file with the name as “<keyword>\_twitter.csv”. The CSV will comprise of the following columns - *id, created\_at, user\_id screen\_name, tweet.*

We capture the date along with the user name and the tweets so that we can later use it to derive some insights of when these topics are generated and the number of distinct users behind such topics.



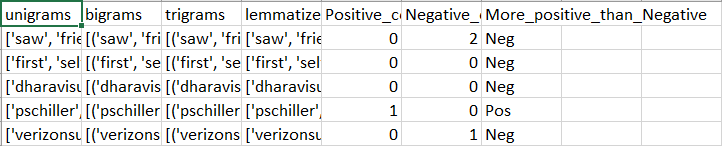
# Data pre-processing and Clean up

Since our approach is to explore prominent themes in the tweets, we have consciously decided to not purge duplicates. The presence of duplicates (or retweets here) indicate that multiple people agree or share the opinion.

In data processing we first performed tokenization. Tokenization is the process of breaking a stream of textual content into words, themes, symbols or some meaningful elements called tokens. Here we generated unigrams, bigrams and trigrams for each and every row or tweet in the data. Then we created three new columns in the dataset containing these unigrams, bigrams and trigrams of the tweets present in the respective rows.

The idea behind this tokenization was to better understand the tweets. We also performed lemmatization on the tokens generated from the tweets and created a new column to store this lemmatized words.

After applying the pre-processing techniques we get data that looks like this



# 

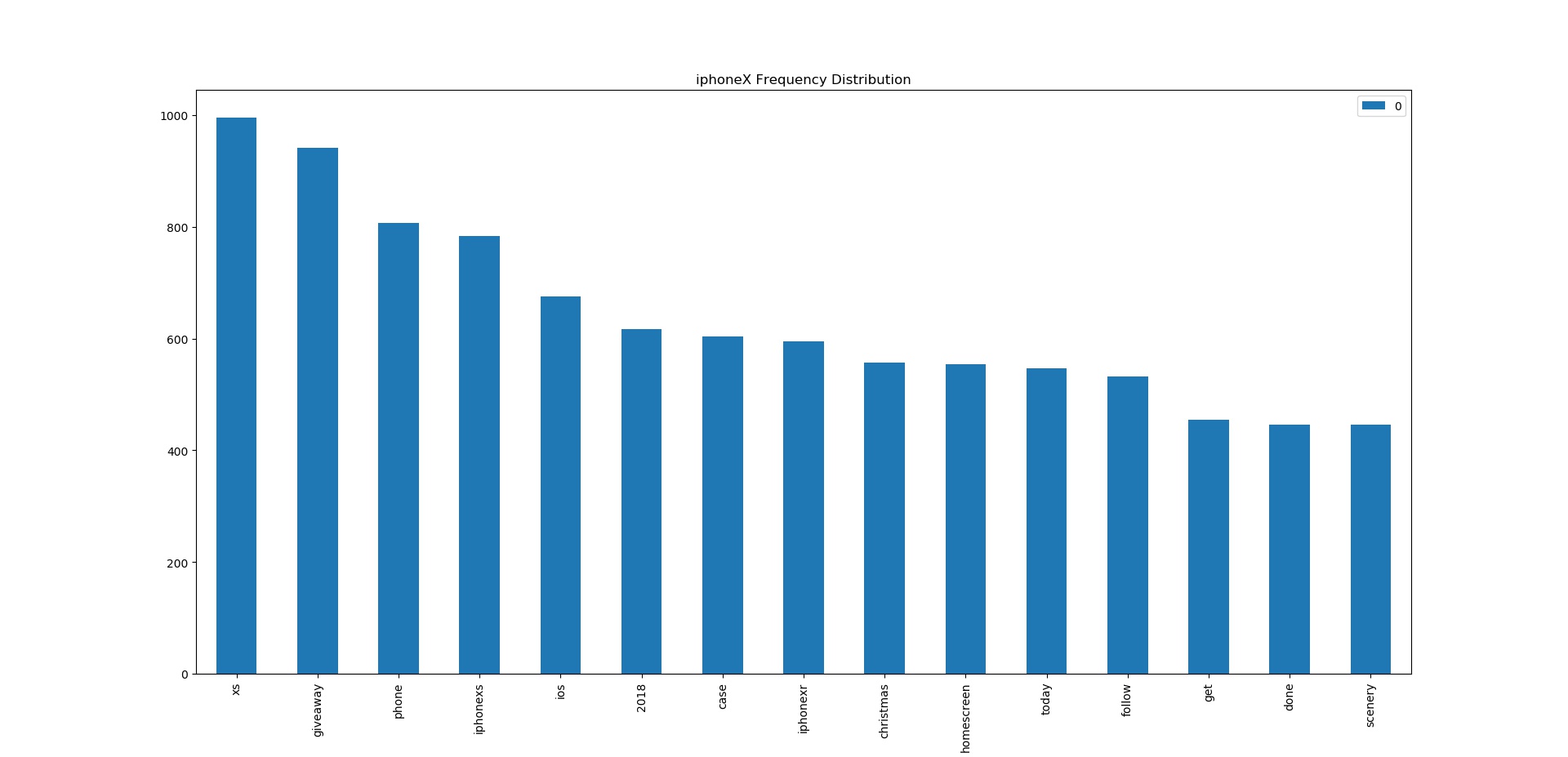
# Exploratory Data Analysis

During this phase of the project we did an in-depth analysis of the data and came up with several hypothesis based on the analysis.

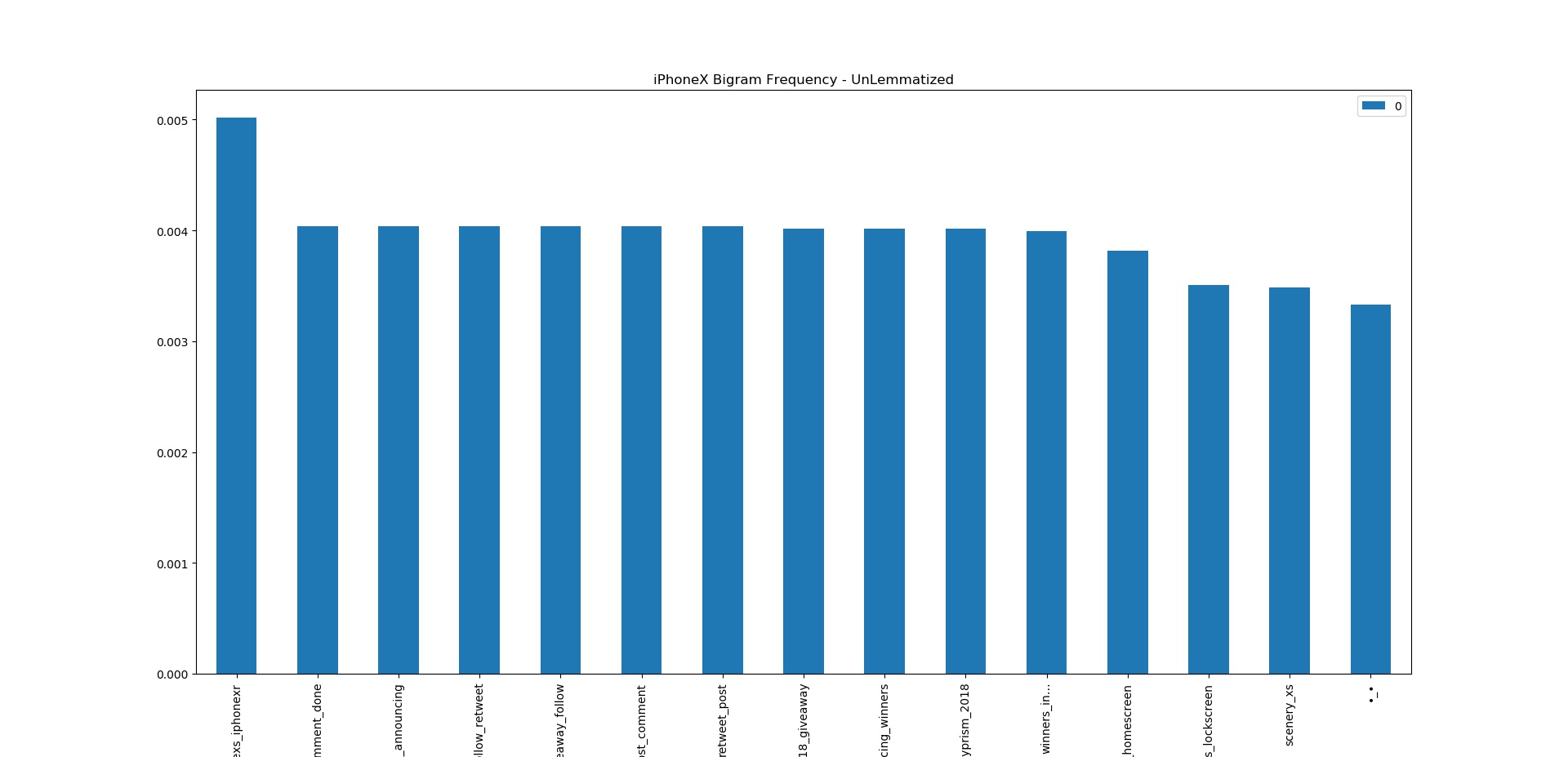
The first step was to utilize the lemmatized tokens and plot a frequency distribution graph. We hoped that the distributions would help us postulate the type of themes or topics that are frequently discussed. On initial analysis some words like http, vs, play, win, iphone appear very frequent but are not relevant as they come from tweets sharing links or tweets asking people to join a game or website to win free merchandise. These keywords are thus added to the existing list of stopwords.

The next step in the EDA process is to generate frequency distribution plot for bigrams. These will give us an idea of which tokens appear together. We use raw frequency as the bigram measure and then pick the top 15 bigrams to be plotted. An interesting observation is that upon lemmatization, we get some bigrams like "x\_x" which make no sense (in the XR data set). This goes to show that lemmatization and stemming should be done carefully and are not approaches that should be applied blindly to data sets.

1. The frequency distribution of the iPhone X data set suggests that tokens like xs (short for iphone XS), iphonexs, iphonexr, case etc. are extremely common. These give us some idea of what are some relevant subjects that we might uncover on further investigation.

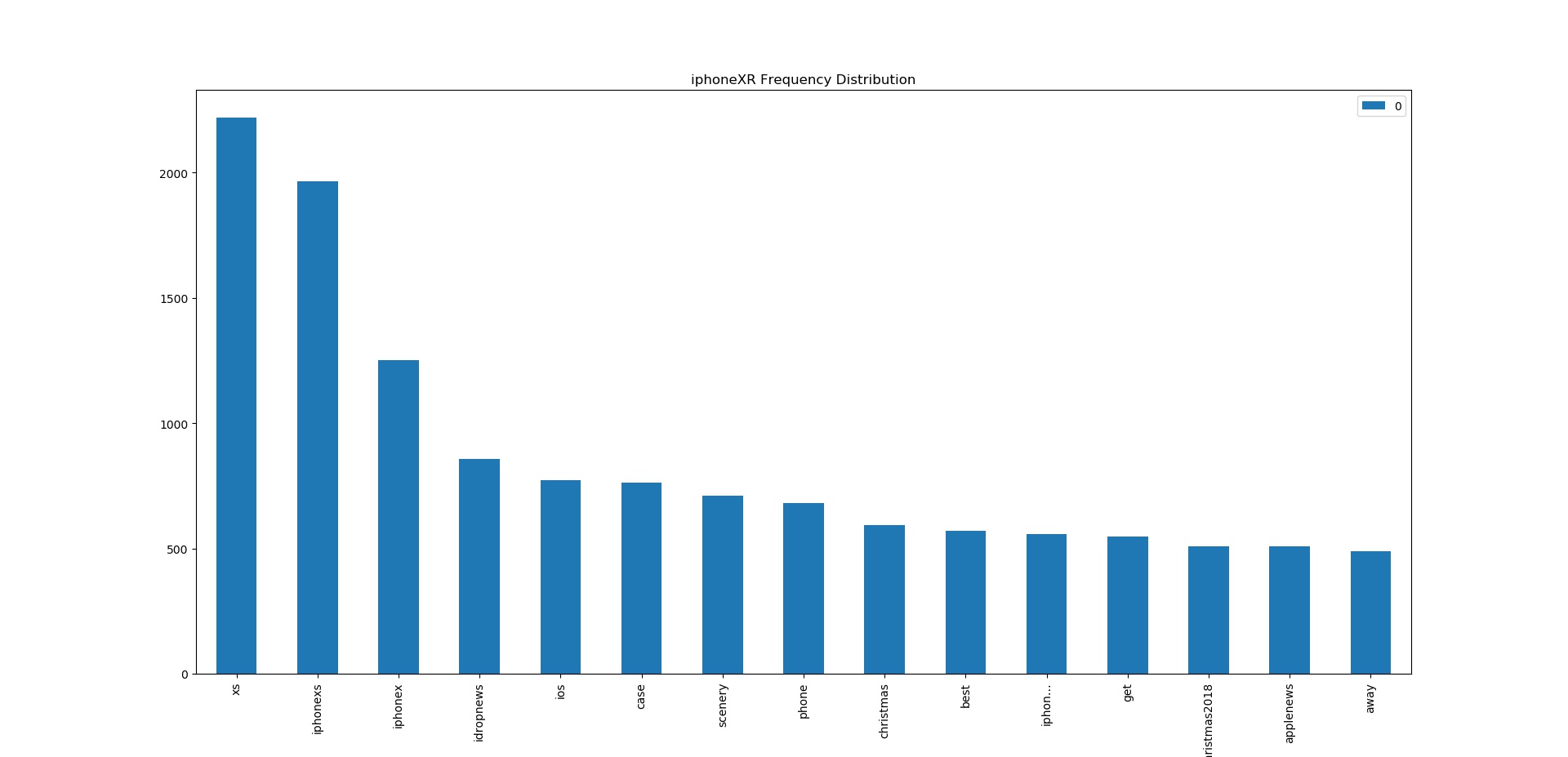


*iPhone X token Frequency Distribution*

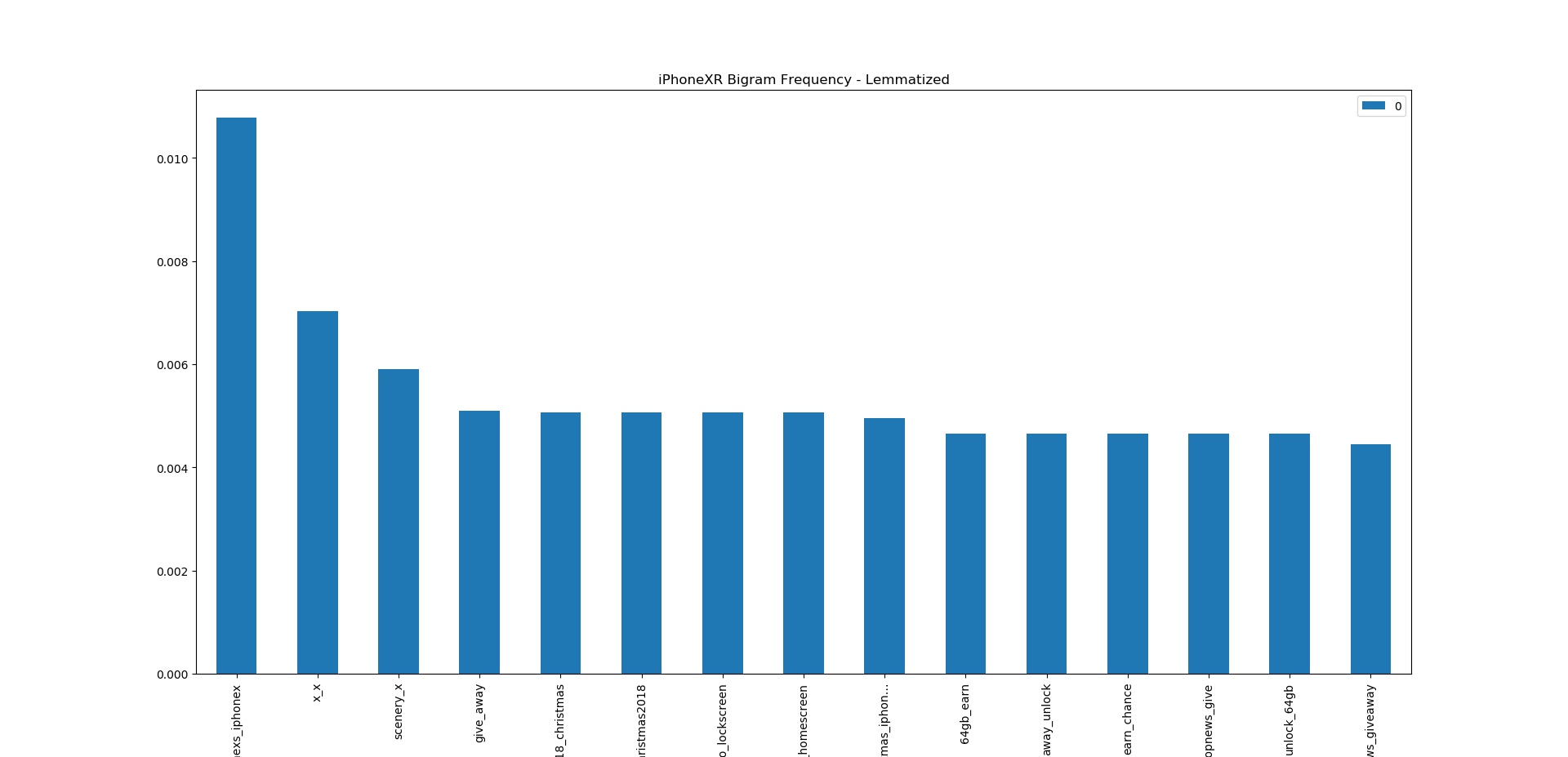
**

*iPhone X bigram (un-lemmatized) frequency distribution*

2. Similarly the data set for iPhone XR displays a frequency distribution that indicates high volume of tweets related to competing products like iPhone X and iPhone XR. Another interesting topic seems to be tokens related to influencers like idropnews and applenews. Lastly, there are tokens like christmas and christmas 2018 that suggest that consumers might be considering this as an option for buying or gifting during the holidays.

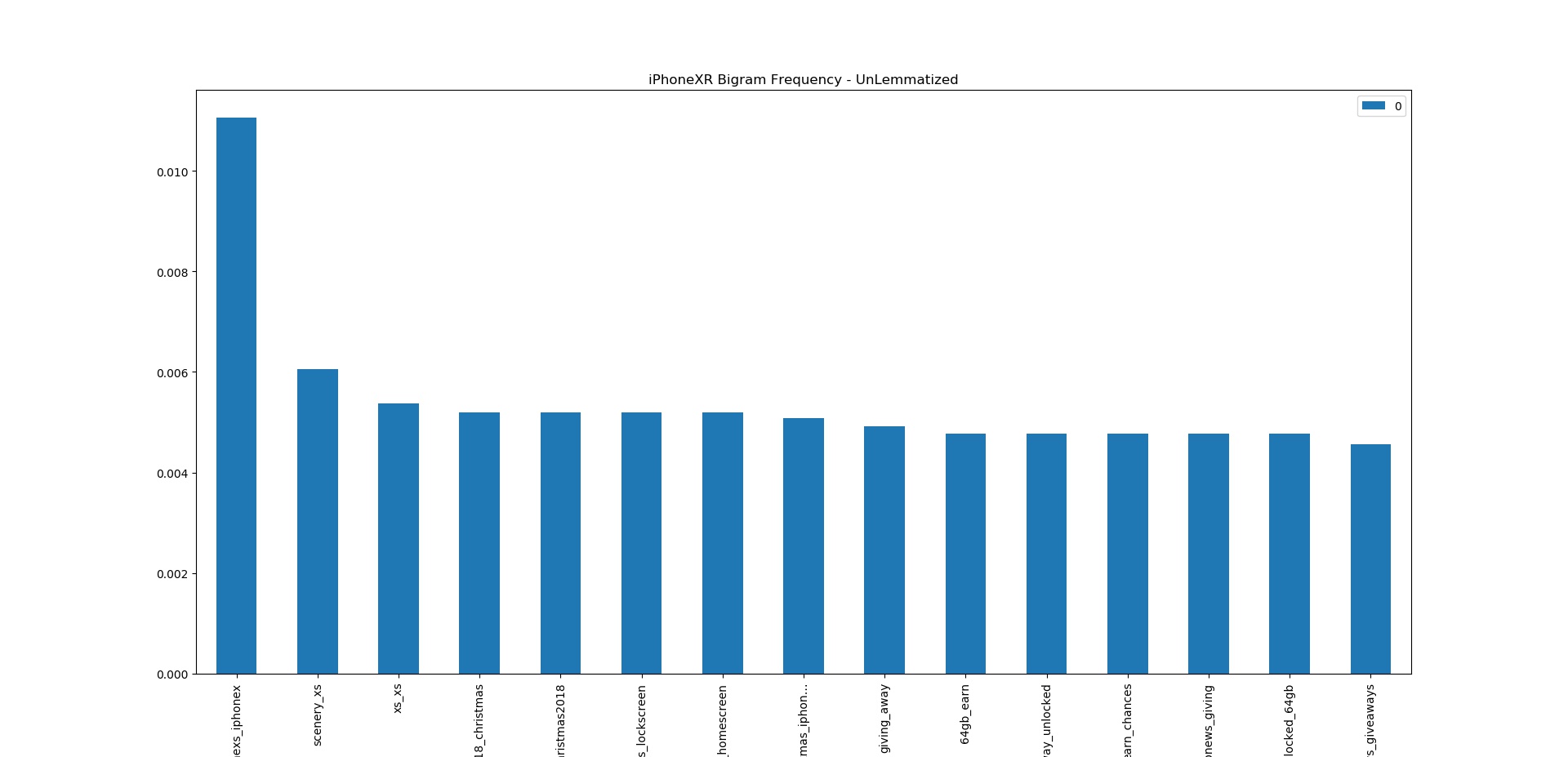


*iPhoneXR token Frequency Distribution*

**

*iPhone XR bigram (lemmatized) frequency distribution*

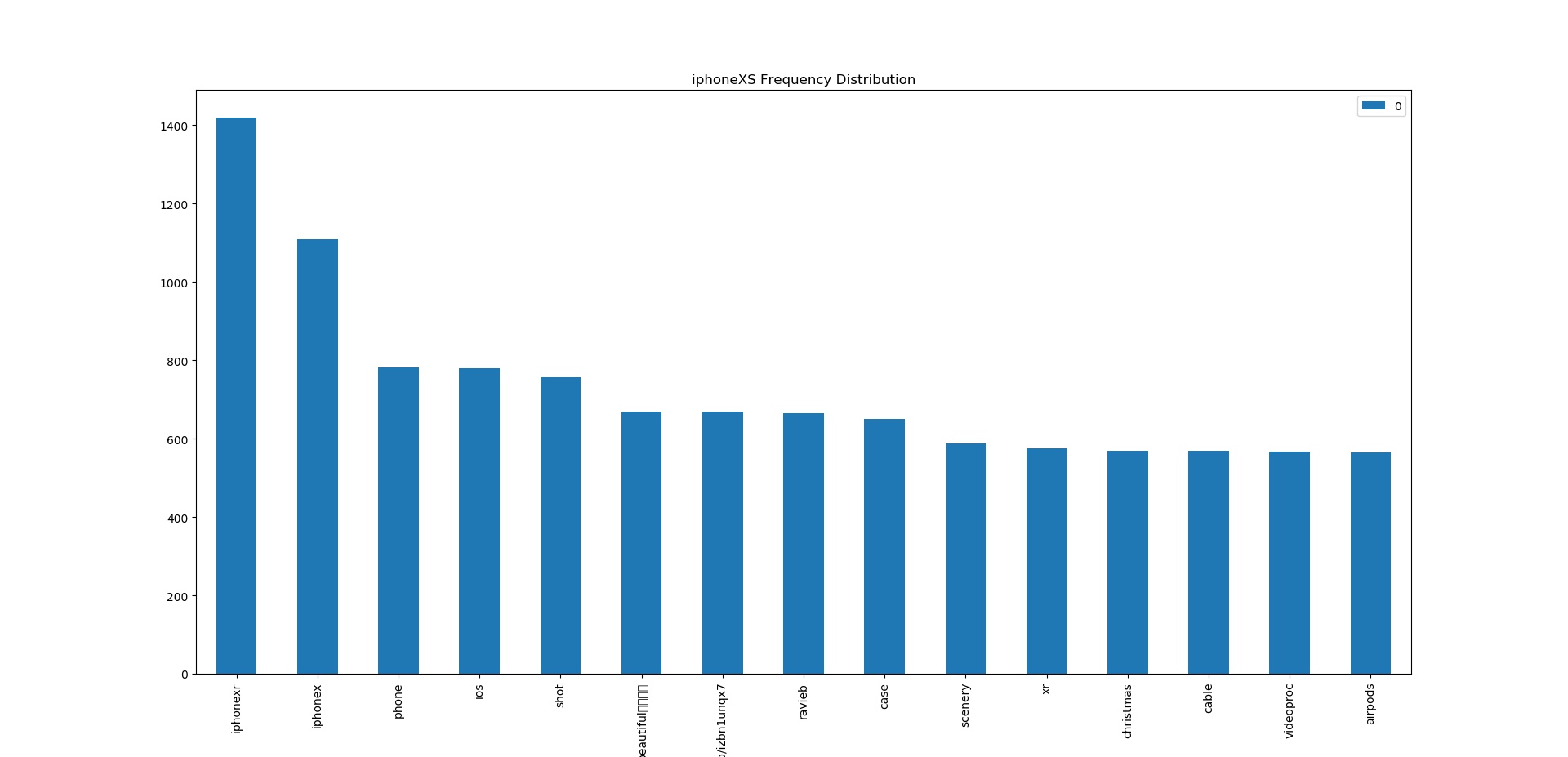
*(Notice the weird bigram x\_x upon lemmatization)*

**

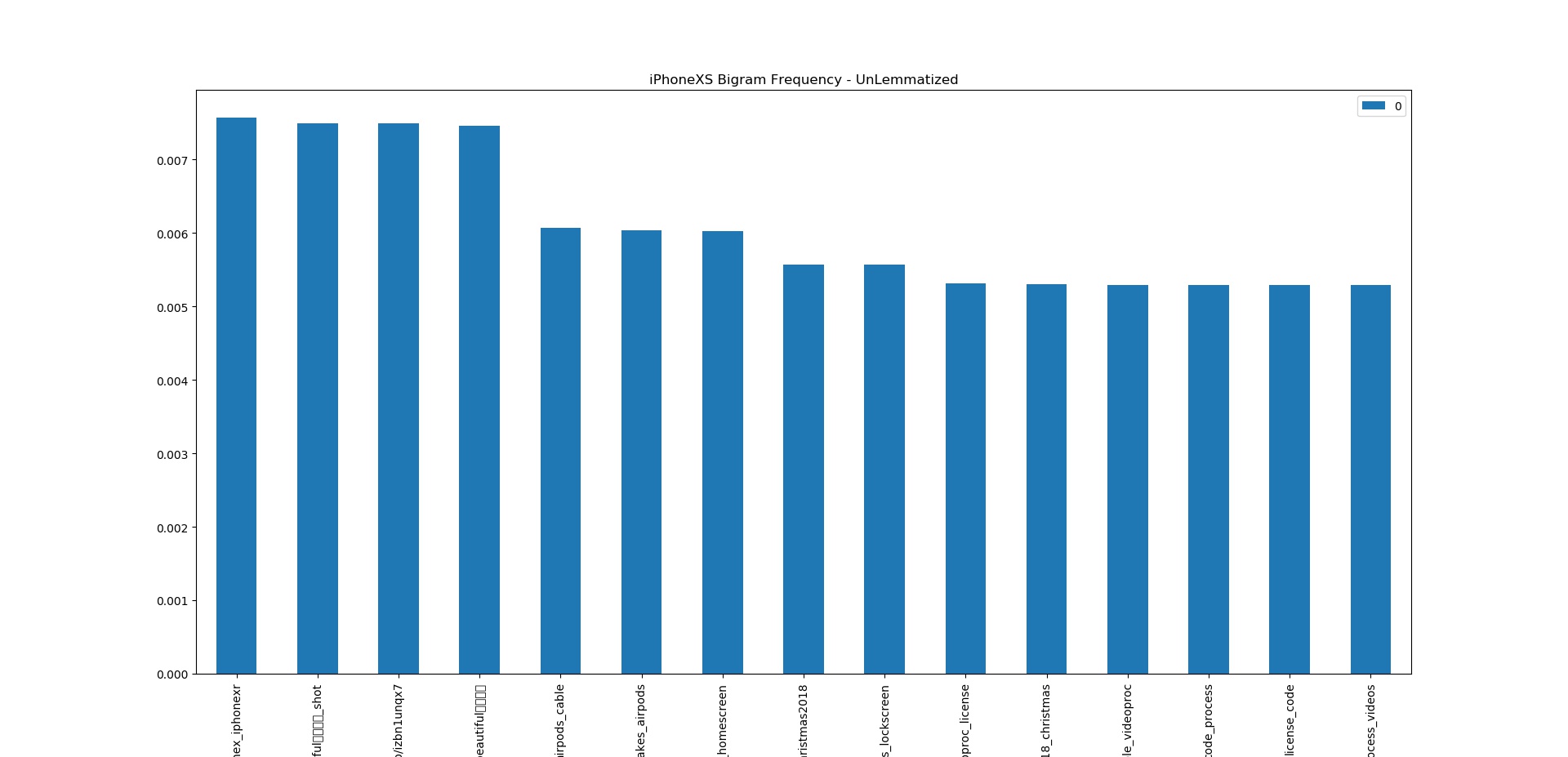
*iPhone XR bigram (un-lemmatized) frequency distribution*

*(Here we see that the bigram x\_x was actually xs\_xs)*

3. If we look at the iPhone XS data set, we discover keywords like xr, iphonexr, iphonex which point towards competing products while quite a large number of tokens like shot, beautiful and scenery that indicate posts about a feature, say the camera. Another group of tokens like case, cable and airpods could suggest a topic of accessories.



*iPhone XS token frequency distribution*



*iPhone XS bigram (un-lemmatized) frequency distribution*

Our exploratory data analysis has uncovered some interesting themes and keeping those in consideration we implement Topic Modelling.

# Topic Modelling

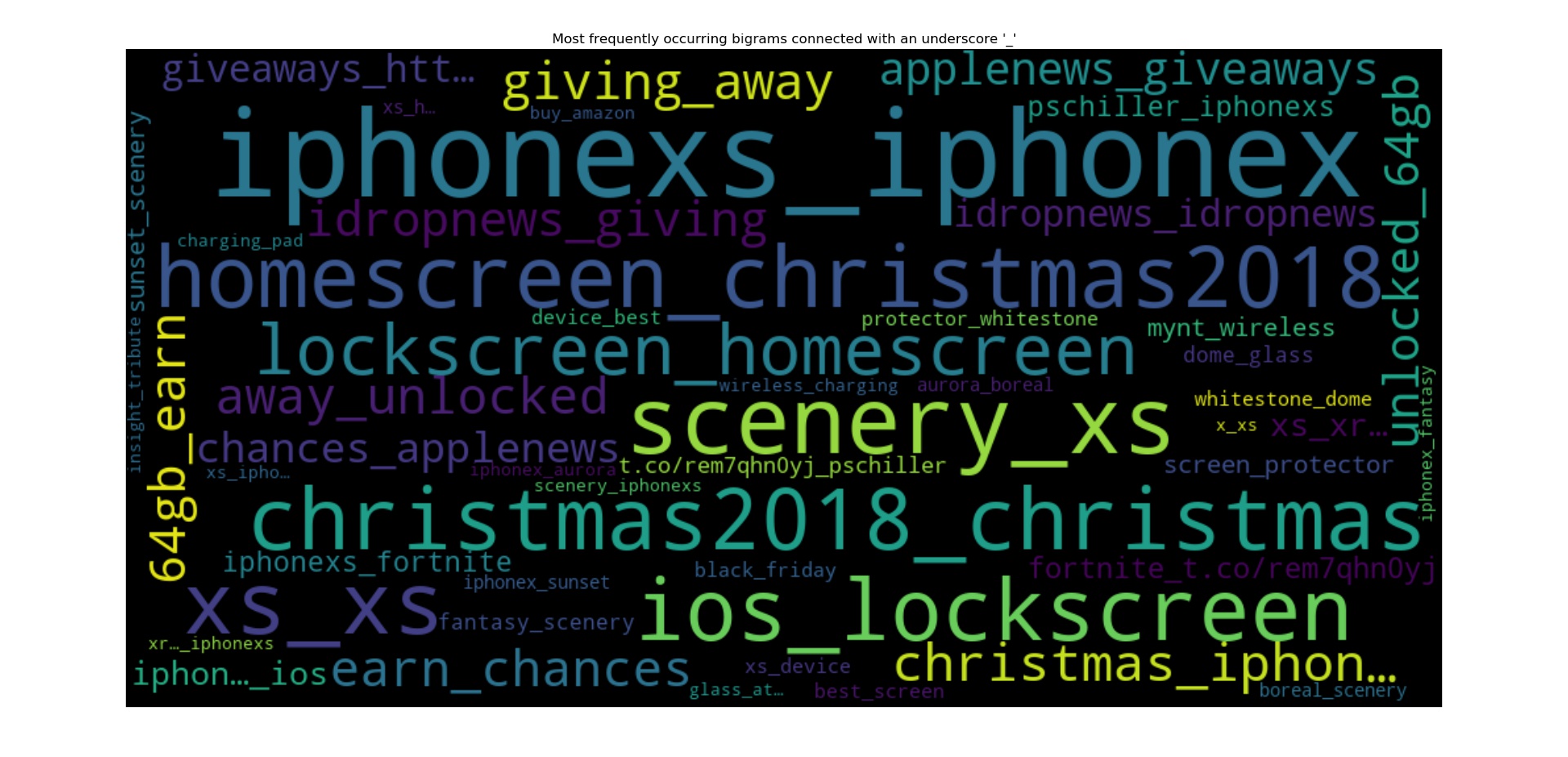
# After discovering the common tokens and bigrams that suggest the presence of certain tokens together we were able to form a hypothesis regarding the kind of topics we can expect to find from the data set. But we need a formal model to generate these topics. For this purpose we implement Latent Dirichlet Allocation using two different Python packages - sklearn and gensim.

# The first step is to create an LDA model to label the topics in the document. We cluster the document of size 10000 into 3 topics. The reason we have selected the number of topics as 3 is that more number of clusters do not generate any coherent results and the topics become very sparse. This stops us from inferring any meaningful insight from the topics.

# 

# *3 distinct topics for iPhone XR (50 words/topic)*

The word cloud of unigrams provides a better sense of what the topics are and what words are associated to them. This is in line with our hypothesis based on the token and bigram frequency distribution. The 3 topics seem to be Demand (Topic 0), Competitors (Topic 1) and Influencers/News (Topic 2). We can compare this with the bigram cloud.



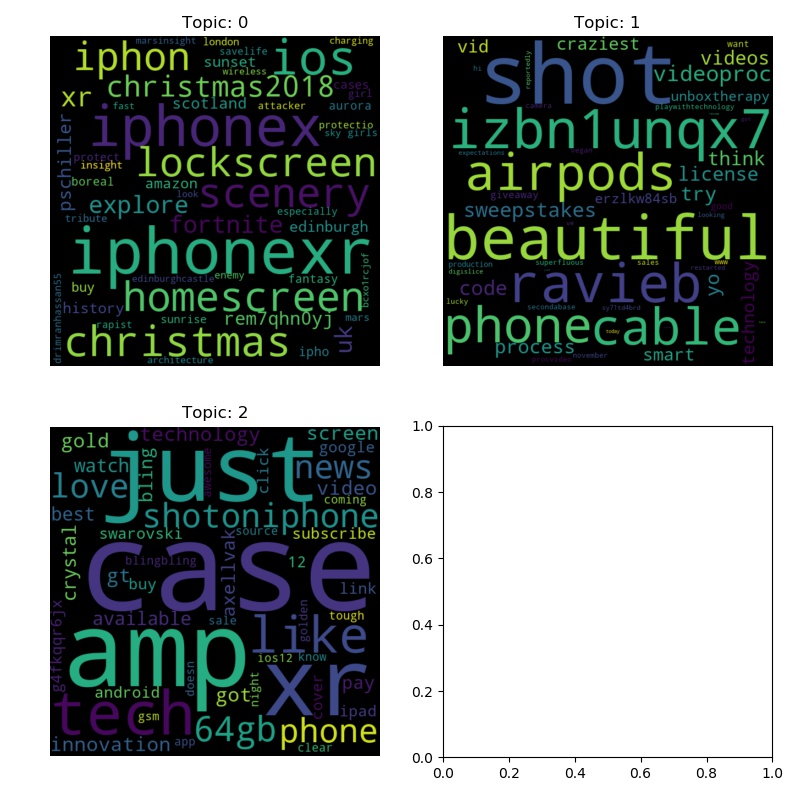
*iPhone XR bigram word cloud for comparison with the topics*

The same approach was applied to the iPhone XS data set. From that we discovered three distinct clusters. The topics inferred are - Competitors (Topic 0), Camera (Topic 1) and Accessories (Topic 2).



*iPhone XS word cloud for comparison*

These bigrams and the manner in which they appear in the word cloud would help us deduce the meanings in the topics and also compare them with our initial hypothesis.



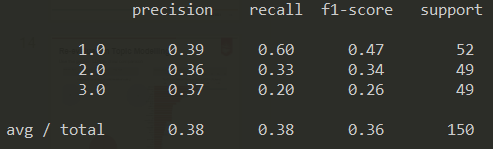
*3 distinct topics for iPhone XS (50 words/topic)*

We can thus summarise that the topics generated are in line with our initial hypothesis. But we need to validate and evaluate the model from a statistical viewpoint.

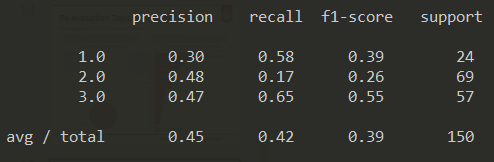
**Evaluating the LDA Model**

To evaluate the model, we labelled 150 documents (50 for each cluster) and derived a classification report from it. The topics were assigned by the Majority Vote rule.

For the iPhone XR data set we obtained the following metrics



# While, for the iPhone XS data set we obtained the following metrics



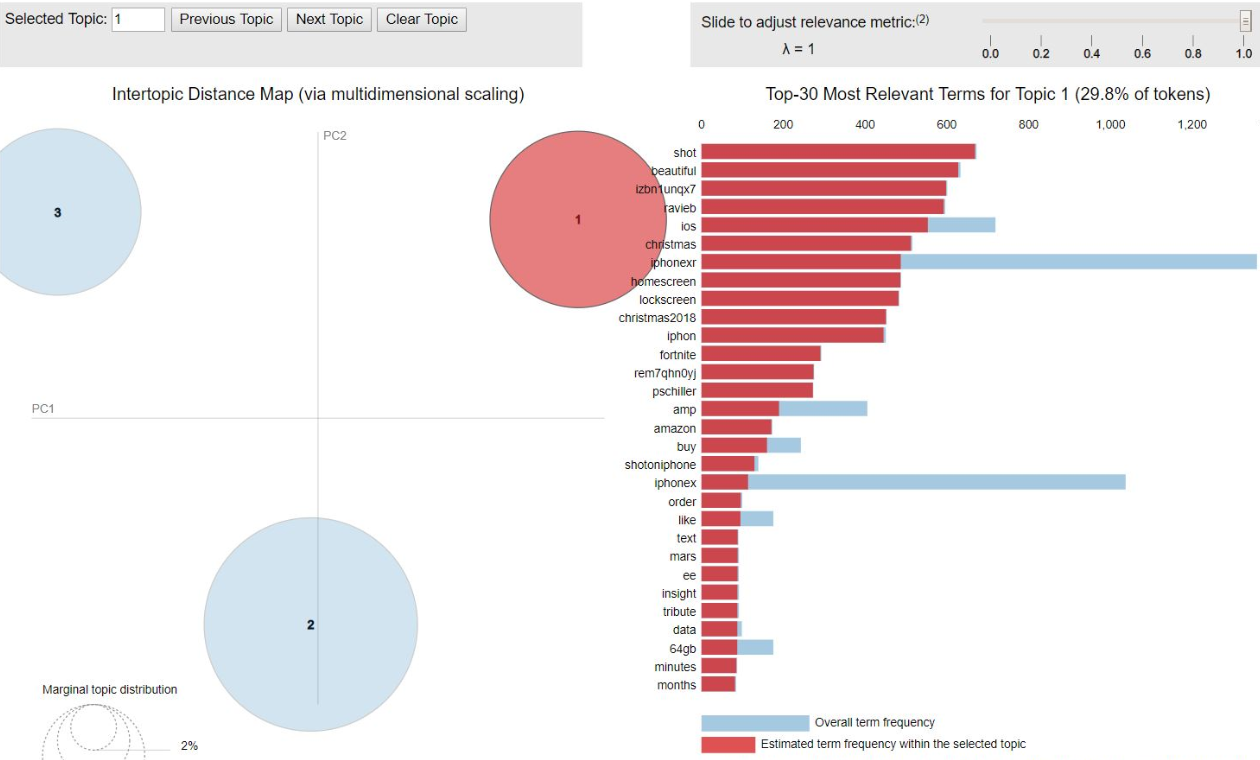
Even though these results do not indicate a high performing model, we believe that the model did the job of discovering relevant topics pretty well. The performance could be significantly improved if the labelling was done by industry experts as is the norm.

Also, if a larger labeled data set was available, we would have tried a supervised approach by training our model on the labeled data. This would have resulted in a more robust model with ideally better performance.

**Gensim Latent Dirichlet Allocation:**

To compare our findings and validate them we chose the gensim LDA package to generate another model. The gensim LDA package lets us visualise the data in a more comprehensive manner. In the results, we can see that we get 3 very distinct topics with very few shared words. This further helps us understand the prominent topics from the tweets and infer a theme.

# 



*Topic clusters for iPhone XS data*

# Again, these distinct clusters match what we achieved from the other LDA model. This goes on to depict confidence in our models and initial analysis of the data. The gensim LDA model was tweaked using the following lambda values - 0.2, 0.4, 0.6, 0.8, 1. Another thing worth noticing was smaller alpha results in more sparse distribution of words in topics.

# Clustering:

## Objective

Used Clustering as we were aware that our data is unlabelled data extracted using the twitter API.

We used SciKit Learn’s K-means clustering algorithm

We wanted to get an Insight on what kind of tweets are being tweeted on the topic.

## Work

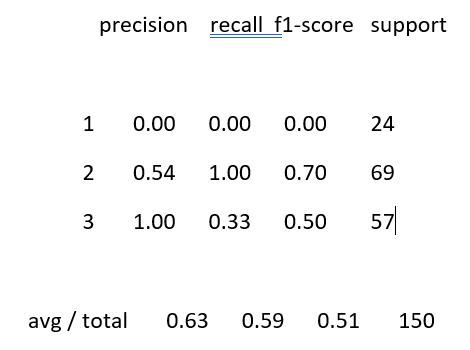
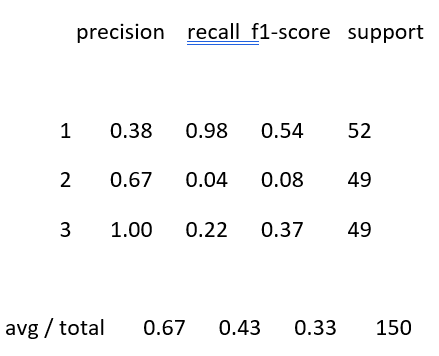
We tried different number of clusters on the data sets we had.

3 and 4 clusters returned a promising result in terms of the seperation we observed.

We finally settled for 3 clusters as we had manually labelled small number of tweets to find out performance metrics.

We measured performance metrics and they are shown in the figures below.

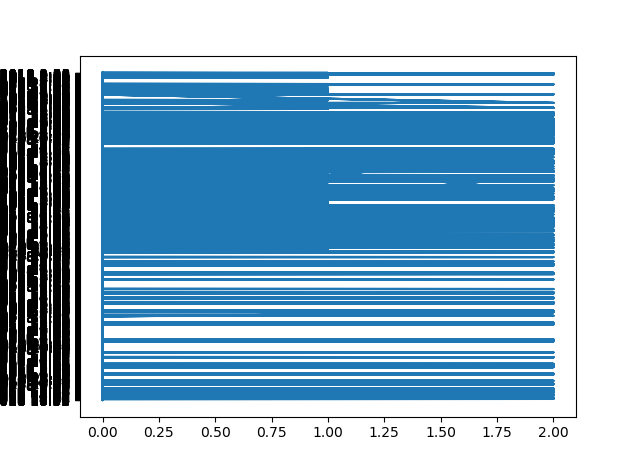
iPhone XS iPhone XR



## Analysis

As Tweets on Twitter are highly dependent on events in the real world, We wanted to get an Idea of how people talk about the iPhone XR and iPhone XS over a time period.

In order to do this, we plotted the Cluster concentration vs timestamp. This gave us an interesting result.

Ledgend: 0- Competitors/Buying/Accessories

1- Features

2- News

#### Fig: Cluster concentration vs timestamp

The figure shows time in increasing order from top to bottom on the Y-axis.

On the X-axis, We have tweets represented by the cluster number where they belong.

### **We observed the following results**:

In the initial periods of the release, people are talking about the competition, Buying as well as the News and reviews of the phone.

In the middle, We have a lot of people talking about the News, reviews and features of the subjects under consideration (iPhone XS,XR).

The remainder of the period, people are again talking about buying it, or the accessories for the phone.

## Inference

This study gave us an interesting insight on how trends move on twitter about a certain product and this can be an extremely valuable insight for the marketing teams for these products.

# 

# 

# 

Hierarchical clustering

Hierarchical clustering is a type of unsupervised machine learning algorithm used to cluster data which is not labeled ,in accordance with the similarity factor among the data point . Hierarchical clustering groups together the data points with similar characteristics as in the k means clustering. In lot of cases the results of hierarchical and k means are similar, but hierarchical gives more in depth insights and clusters of the data.

As our data is not labelled for iphone XR and iphone XS , we will use k means and hierarchical clustering to compare the performances of the unsupervised algorithms for finding some outstanding insights. One of the other reasons for choosing hierarchical clustering is to find a complete range of cluster solutions.

The hierarchical algorithm can be briefly summarised into the following steps:

1. The algorithm begins with treating each data point as an individual cluster.
2. 2 closest data points are clumped up together to form a cluster.
3. 2 closest clusters are grouped together to form a bigger cluster.
4. This process is repeated till one big cluster in formed.

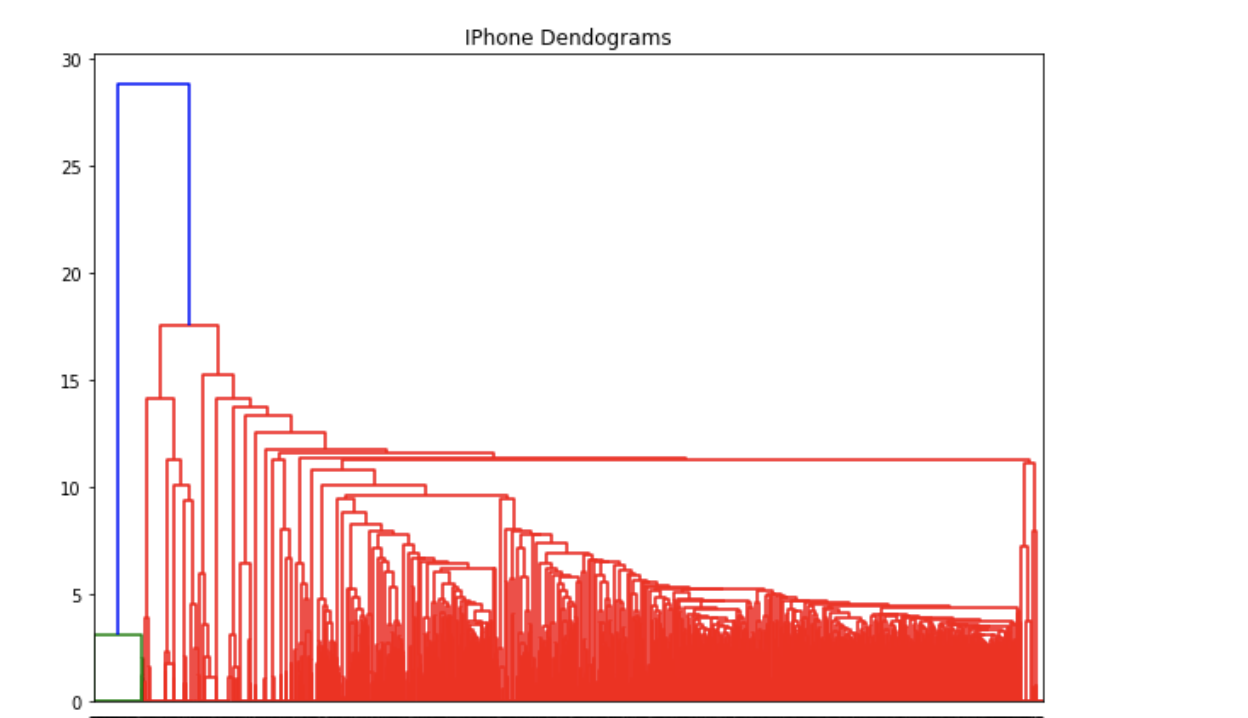
Details used for agglomerative clustering:

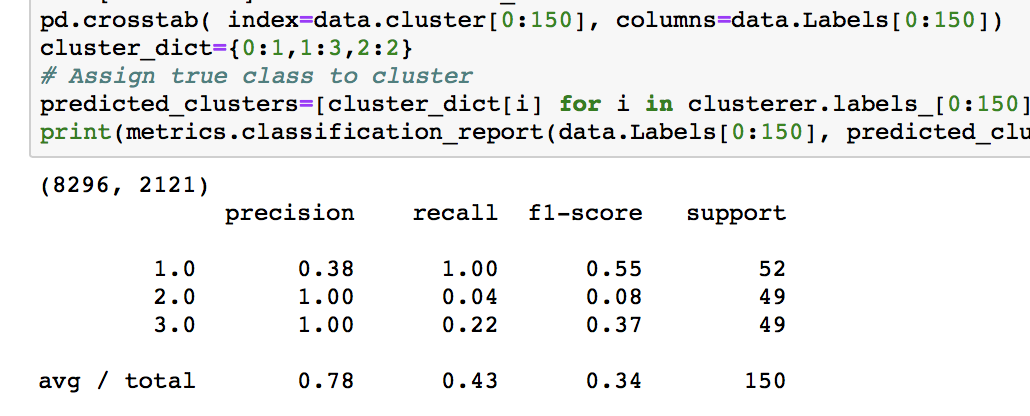
* Ward: Ward is one of the linkage methods that forms spherical clusters of similar sizes , extremely differentiated from other groups. It works only with euclidean distance.
* Number of clusters : We tried to achieve some insights by forming 3 clusters.
* Affinity : Metric used to compute the linkage. We can use euclidean, l1,l2,manhattan,cosine or precomputed.

Following is the function used for agglomerative clustering:

*clusterer = AgglomerativeClustering(n\_clusters=3, affinity='euclidean', linkage='ward')*

As the data size is huge, it is hard to understand the dendrogram visualization of the cluster, which is a tree like structure representing the cluster distribution of the data points.

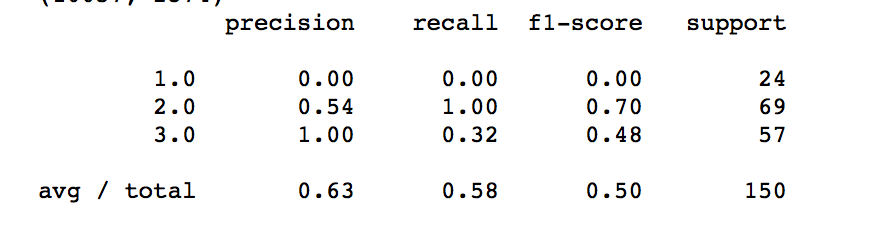
The dendrogram visualisation iPhone XR can be found as follows:****

The results after the clustering the text for the iphone XR data can be summarised in the following fig****

*Data: iPhoneXR\_twitter\_eda.csv*

When we compared the performance metrics of Hierarchical clustering and k means clustering for iphone XR we found out that the performance of hierarchical clustering is better than the k means , even though hierarchical clustering was computationally more expensive than the k means clustering .

The efficiency of the hierarchical clustering algorithm was tested by labelling 150 entries of the same data set and forming the crosstab of cluster number and the corresponding labels.

Hierarchical clustering was used for examining the iphone XS data as well and the process was similar as that was used for the iphone XR but with the iphone XS data . Its results can be summarised in the following figure

Data:iPhoneXS\_twitter\_eda.csv

**Hierarchical vs K-means clustering:**

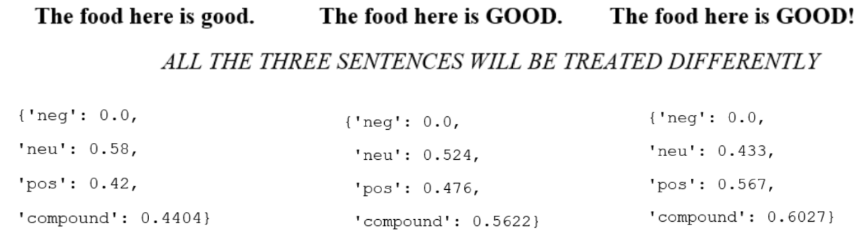
Both the unsupervised algorithms have pros and cons. For our particular case, we can confidently summarise that we found more proficient and accurate results in the hierarchical clustering than the k means clustering even though it is computationally more expensive and time consuming.

# Sentiment Analysis

In addition to topic modeling and Clustering we also performed some sentiment analysis on the tweet data for understanding people’s opinion about the products. VADER is a lexicon and rule-based sentiment analysis tool and out of all the sentiment analysis approaches available, we chose VADER because of the following reasons-

* VADER works extremely well on social media text
* It doesn’t really require any training data
* It handles things like exclamation marks, capitalization, emojis, slangs and acronym text very well which is commonly used in social media text

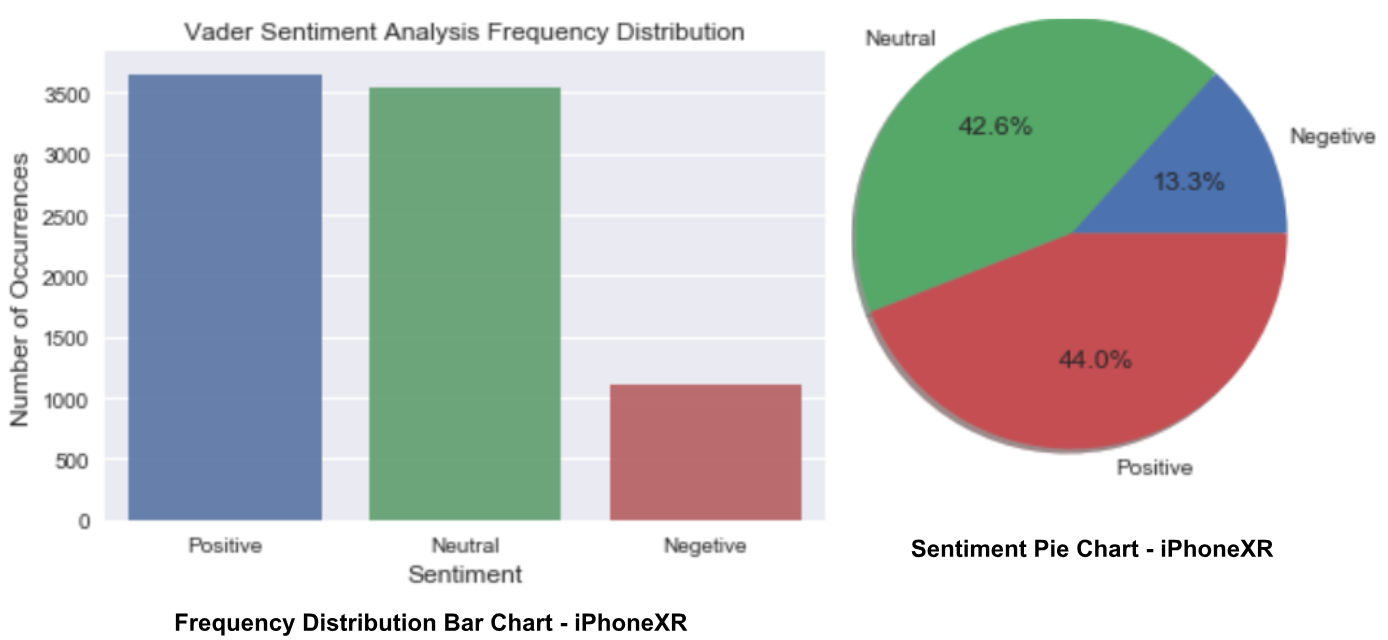
In VADER 4 scores are generated for each piece of text. Positive score, negative score, neutral score and a compound score. The positive, negative and neutral scores represent the percentage of work falling into these categories. The compound score is a metric that calculates the sum of all the lexicon ratings which have been standardized between -1 (most extreme negative) and +1 (most extreme positive). The figure seen below shows how VADER deals with Capitalization and exclamation in texts.



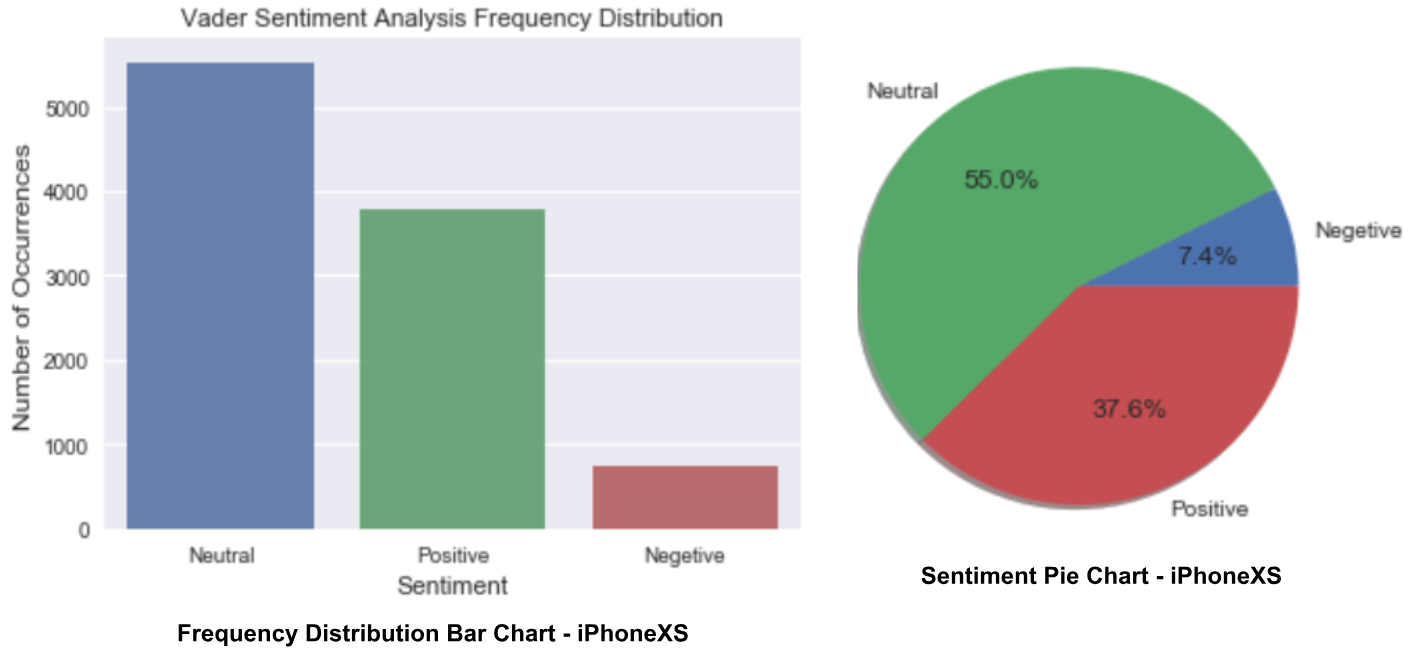
*How VADER treats capitalization and exclamation marks*

**Sentiment Analysis VADER results:**

The two images seen below show VADER results for iphone XR and iphone XS respectively. In both the cases VADER sentiment analysis has been performed on the entire dataset of tweets of the two respective products. This score thus represents the overall sentiment that the the tweets has about these two products. It can be concluded from the graphs seen below that: Percentage of positive and negative tweets for iphone XR are almost similar. But in case of iphone XS the percentage of neutral tweets have increased and at the same time the percentage of negative tweets has also reduced.



*VADER results for iphone XR*

**

*VADER results for iphone XS*

# Existing Work vs Our Work

We explored a lot of existing work during the ideation phase of our project and not during the implementation. We knew that we wanted to do something using the Tweets regarding new products but we were looking for ideas and that is why we went through some existing work. Below are the few pieces of work that we had referred to during the ideation phase.

* Liu B., Zhang L. (2012) A Survey of Opinion Mining and Sentiment Analysis. In: Aggarwal C., Zhai C. (eds) Mining Text Data. Springer, Boston, MA
  + <https://doi.org/10.1007/978-1-4614-3223-4_13>
* Norman Booth, Julie Ann Matic, (2011) "Mapping and leveraging influencers in social media to shape corporate brand perceptions", Corporate Communications: An International Journal, Vol. 16 Issue: 3, pp.184-191,
  + https://doi.org/10.1108/13563281111156853
* Katja Hutter, Julia Hautz, Severin Dennhardt, Johann Füller, (2013) "The impact of user interactions in social media on brand awareness and purchase intention: the case of MINI on Facebook", Journal of Product & Brand Management, Vol. 22 Issue: 5/6, pp.342-351
  + https://doi.org/10.1108/JPBM-05-2013-0299

Lets quickly see some differences between the existing body of work that we studied and our work.

|  |  |
| --- | --- |
| **Existing Work** | **Our Work** |
| In 2011, Booth and Matic in their paper, Mapping and Leveraging Influencers in Social Media to shape Corporate Brand Perceptions tried to identify and focus on **social media influencers** and understand their impact over target audiences | We did not focus on a specific group rather we studied the views of all people. Results may have been different if we had restricted our study to a particular |
| Katja Hutter discusses in her paper, The impact of user interactions in social media on brand awareness and purchase intention discusses how brand user interaction activities can affect the perception of a brand | We have chosen a brand, i.e. apple but we do not necessarily focus on brand to user interaction |

# Analysis of Experiment Results

1. **What part of your methodology worked (or didn’t work)?**

The performed topic modeling and clustering to identify what topics or subjects are being talked about in the Tweets by potential customers. But the results of Topic Modelling as well as clustering could have been better is what we concluded.

1. **Why did your methodology work or (didn’t work)?**

I guess we did not get great results because of the following factors-

* We used unsupervised machine learning algorithms in which labeling for the test data was done manually by us. We did not have any experts labeling our data. This could be one of the reasons why we did not get great results.
* More importantly we were only able to label a small amount of data as we did that manually. Increasing the number of labeled data will definitely give us better results on the algorithms mentioned above.
* Lastly, since the labeling was not done by experts but rather by us, it is possible that some sort of bias got introduced in the labeling process.

1. **How to improve?**

The following things can be done-

* Label more data to be used as a test set
* Seek expert advice on labeling appropriately

1. **How to utilize your results? What business insights can be derived from your analysis?**

From our results it is possible to identify what are people are talking about at a particular time on Twitter about the new products. We tried to cluster the different topics that people discussed regarding both the iphone XR and iphone XS. This data can be used to understand public interest in topics regarding the product and this can be used by marketing and advertising campaigns. This data can be used to identify desires that customers may have about the product and these desires can be fulfilled in the subsequent iterations of these phones. Also we learn about other interests that people generally have in phones like accessories.

# Conclusion and Future Work

Although our results can be used to identify what people think about the new products by clustering their tweets, the results of our experiment can definitely improve. The fact that we ourselves have labelled the data and we did not have any expert to label our data can be considered as one of the factors that explains our results. We also believe that increasing the number of labeled data will definitely improve our results.

**Future Work**

* In the future we plan to extract texts from each cluster and then perform sentiment analysis on each and every cluster to give us a better idea about the text.
* We are also considering that in the future we can narrow down our scope to tweets of social media influencers and the subsequent retweets that these tweets receive.