Assignment 1
Analysing and Tracking the Sentiment and Topics on Social Media

Table of Contents

1. Introduction 2

2. Data Collection 2

2.1 Data Extraction and Description 2

2.2 Data Exploration 3

3. Pre-processing and Data Cleaning 4

3.1 Cleaning Outcome 5

4. Sentiment Analysis Method 7

4.1 Opinion Word Counting Sentiment Analysis 7

4.2 Vader Sentiment Analysis 8

4.3 Topic Modelling Approach 9

5. Analysis & Insights 10

5.1 Sentiment Analysis Insights 10

5.2 Topic Modelling Insights 11

6. Conclusion 12

7. References 13

1. Introduction

This report offers a thorough topic modeling and sentiment analysis of Reddit discussions around iPhones. With the use of cutting-edge text mining and Natural Language Processing (NLP) techniques, we hope to learn more about how the public views Apple's flagship product, as well as identify important conversation points and customer satisfaction levels. In addition to classifying sentiments as positive, negative, or neutral, our analysis will highlight recurring themes and patterns in user discussions. IPhone being a popular smartphone that gets a lot of feedback on different platforms. Effective techniques for methodically going over and deriving meaning from this massive amount of data are sentiment analysis and topic modeling. The knowledge gathered from this study will be helpful for stakeholders in formulating marketing plans, developing new products, and making general business decisions. This project aims to assist Apple in sustaining the iPhone's market position, improving user satisfaction, and directing future innovations based on consumer feedback and preferences by methodically analyzing a substantial volume of user-generated content. Moreover, revealing hidden or undercurrent topics can disclose the aspects of the "iPhone" that consumers are interested in, which is a useful source of data for marketing and new product development.

2. Data Collection

2.1 Data Extraction and Description

For the sentiment analysis of Reddit data related to the iPhone, the data collection step included using a Python script that leverages Restful Reddit API through the praw (Python Reddit API Wrapper) library. After that, submissions from the "iPhone" subreddit were retrieved using the Reddit client, with a focus on the hottest posts up to a 1000 threshold, which generated 952 posts and 9121 comments making our corpus size 10,073 documents. We extract and store every post's title, score, ID, URL, creation date, number of comments, body content, and word count. Along with retrieving each post's comments, the script also records information about each comment, including its ID, author, body, creation date, score, and word count. Our data spans from 28th of August 2024 to 6th of September 2024. The JSON file that is created from all this structured data will be used for additional sentiment analysis. This methodology guarantees all-encompassing data extraction, thereby enabling an exhaustive examination of public opinions regarding the iPhone.

2.2 Data Exploration

For our exploration, we examined the length of posts in words, it was found that the average length of posts was 65.98 which is a very decent length for a post, shows that people posting really mean their posts. Figure 1 below shows distribution of raw reddit post length.

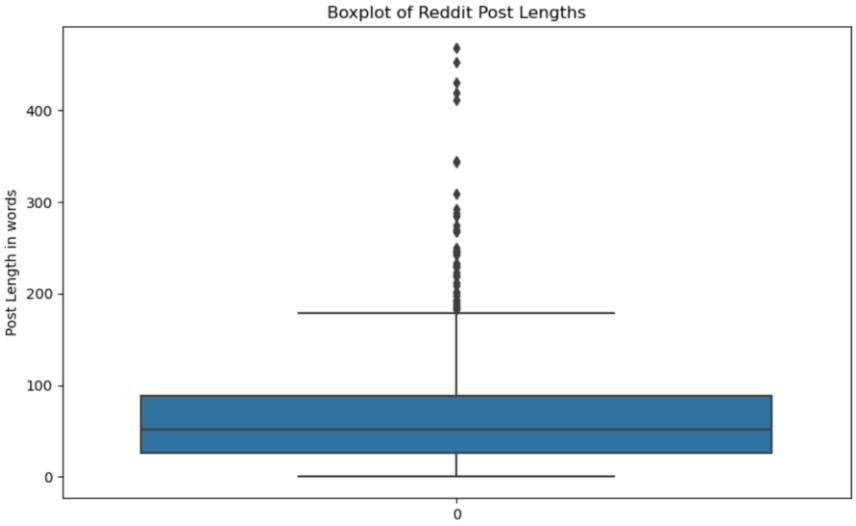

Figure 1: *Boxplot of Reddit Post Lengths*

We also looked at the number of posts about "iPhone" that were made each day. The graph shows how many posts were made on Reddit every day between August 28, 2024, and September 6, 2024. There is a discernible increase in posts at the end of August and the start of September, which could indicate an impending occasion given the pattern of new iPhone releases in September.

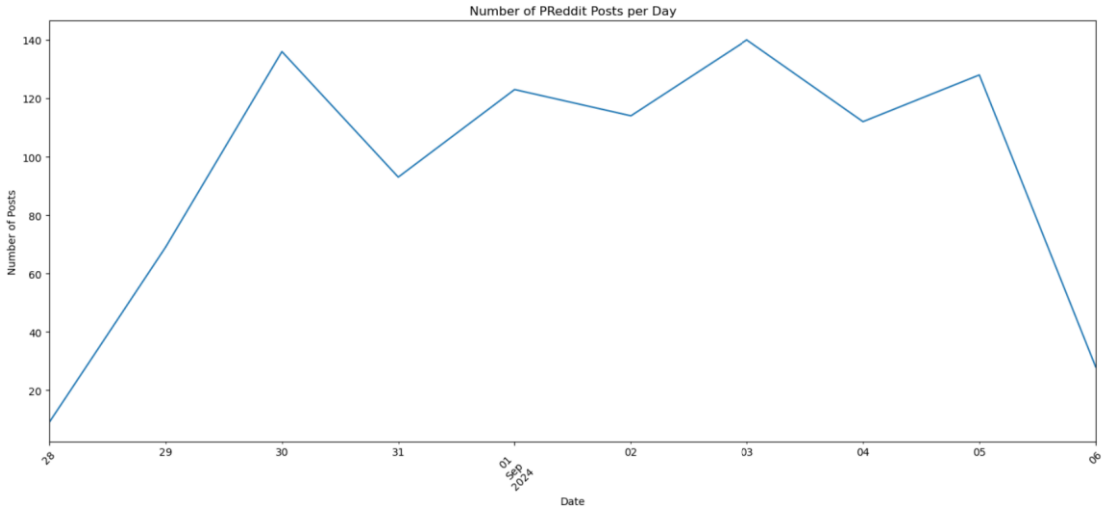

Figure 2: *Number of Reddit Posts per Day*

3. Pre-processing and Data cleaning

Pre-processing is one of the most crucial steps in any analysis as the effectiveness of our analysis hinges critically on the quality of our data. I have implemented a pre-processing pipeline to the data to clean the data preparing it for our sentiment analysis. We aim to remove noise (such as stop words, digits, URLs) and normalize text (through tokenization and lemmatization) to clean our data.

Our comprehensive pre-processing pipeline comprises of the below mentioned steps:

- Conversion to lowercase: It is to normalize text, converting all text to lowercase, establishing a uniform base for analysis.
- Tokenization: It is done to break down sentences/full texts into individual words/tokens, facilitating granular analysis.
- Trimming excess leading and trailing white space from tokenized words: It ensures to remove a whitespace (a space) before or after a token.
- Digits removed, URLs and special characters removed: Since digits, URLs, and special characters do not contribute to our analysis, we are removing these tokens.
- Stop words removed: We retrieve a list of common English stopwords from NLTK library. Words like ‘and’, ‘this’, ‘is’, ‘they’ which do not provide any significant meaning are removed. I have also included punctuation marks (e.g., ‘?’, ‘,’ ,’!’) list, extracted from ‘string.punctuation’. Also, I have included a custom list containing ['via', '...', '…', "", """, ",'⣿',' ',''].
- Lemmatization: It is the process of reducing words to their base or root form by utilizing ‘lemmatizer’, a tool in natural language processing (NLP). It is favored over stemming because unlike stemming, that simply cuts off the end of words, which sometimes result in non-dictionary forms, lemmatization transforms words into their proper base form maintaining the context and part of speech. This approach significantly enhances the accuracy of our analysis.

3.1 Cleaning Outcome

Below I have made a comparison between the post before and after pre-processing, the effect of our cleaning steps can clearly be seen on the post text. Stop words like ‘I’, ‘a’, ‘is’ etc. , URL and digit 15 have been removed, leaving us with a more refined set of tokens ideal for our analysis.

- **Reddit post (pre-clean):**

https://imgur.com/a/TtU1rYW
check out the clip , i have a 15 pro is my gpu possibly failing ? there is a weird artifact while benchmarking

- **Reddit post (post-clean):**

['check', 'clip', 'pro', 'gpu', 'possibly', 'failing', 'weird', 'artifact', 'benchmarking']

The most common terms in Reddit posts before and after data cleaning are compared, and the results demonstrate a notable improvement in data quality. The pertinent keywords were first hidden by common stopwords and non-content-bearing terms like "the," "to," and "that," which dominated the data. Following cleaning, these superfluous terms were eliminated to expose more pertinent content, such as "iphone," "phone," and "support," which are all closely associated with the topic at hand. The accuracy and applicability of the analysis that followed were improved by this process, which sharpened the focus on pertinent data.

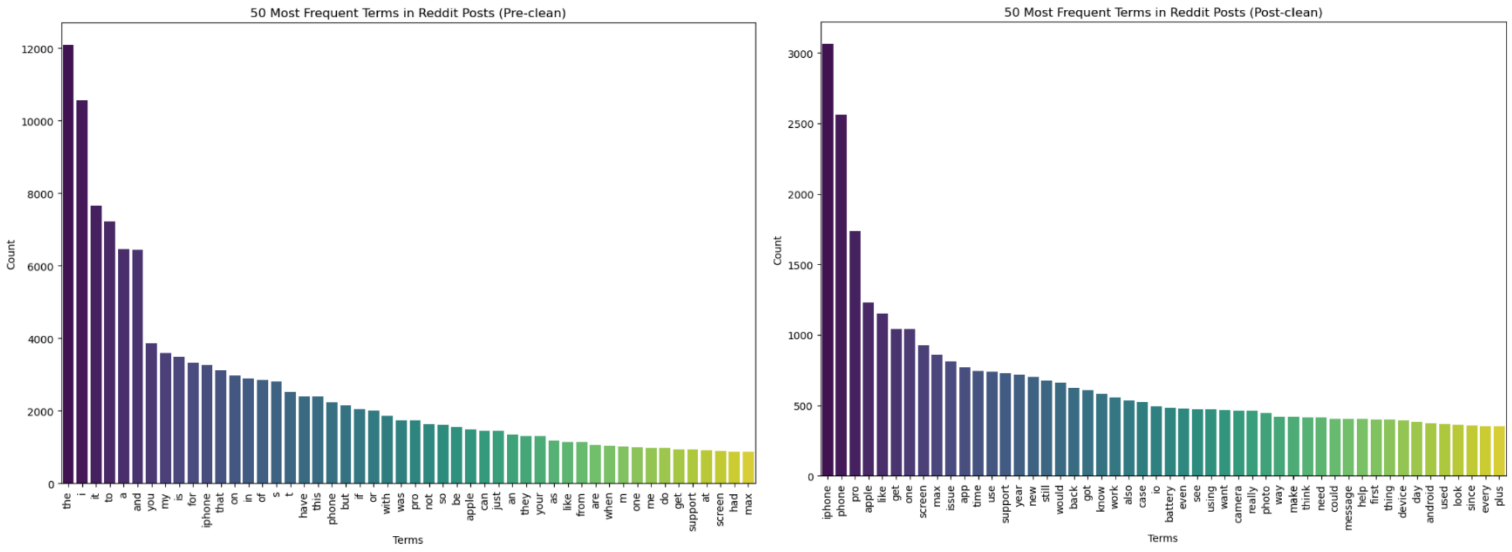

Figure 3: *The top 50 Frequent terms pre clean (Figure 3A, left) and post-clean (Figure 3B, right)*

4. Sentiment Analysis Methods

Sentiment analysis can be done in a number of ways. We concentrate on two techniques in this report: the Count method and the Vader method.

4.1 Opinion Word Counting Sentiment Analysis

- Post tokenizing (divided into individual words), the Count method assesses each token independently. Each word is given a score based on its inclusion in a custom dictionary of positive and negative words: 1 for positive, -1 for negative, and 0 for neutral. The total number of negative words is subtracted from the total number of positive words to determine the overall sentiment score. A positive score indicates a positive sentiment in the tweet; a negative score indicates a negative sentiment.
- While these dictionaries aren't designed with Reddit in mind, preprocessing the text to eliminate slang and acronyms reduces context-specific problems. We count the positive and negative words in each tokenized tweet and calculate the difference between these counts to get the sentiment score. An overall positive sentiment is indicated by a positive score, whereas a negative score denotes a negative sentiment. The course materials contain word lists that were taken from movie reviews. One drawback of this approach is that it doesn't recognize context or word relationships which in turn limits its ability to identify specific words.

4.2 Vader Sentiment Analysis

- Vader (Valence Aware Dictionary and Sentiment Reasoner) is a more sophisticated model, found in the NLTK package. Unlike, count method, which provides a simple positive or negative score, it uses a more detailed scale to rank a sentiment. It also incorporates special rules for boosting the sentiment score, by emphasizing on words in capital or accompanied by exclamation marks, reflecting the increased intensity of the sentiment being expressed.
- Vader also negation handling, a technique which recognizes a sentiment as negative is a positive sentiment word is preceded by a negation (e.g., 'not good'). Vader considers the context around words to provide a comprehensive sentiment score.

4.3 Topic Modelling Approach

- I've used LDA Model to perform topic modelling. LDA is probabilistic clustering model. It assumes that each document is made up of a fixed number of topics. Topics can generate words based on some probability. Topics are ultimately identified based on occurrences of words.
- LDA has several parameters and optimal values for these parameters are very volatile and change with corpus and documents.
- Latent Dirichlet Allocation (LDA) is a statistical model that we have implemented for topic modeling of Reddit posts on iPhone. It works under the premise that topics are distributions over words and documents are mixtures of topics. In LDA, terms are iteratively assigned to topics according to their patterns of occurrence in documents and throughout the corpus. The model uses Dirichlet distributions to manage the word-topic and topic-document probabilities, and it requires an initial specification of the number of topics. The advantages of LDA are its unsupervised nature, its capacity to reveal hidden themes, and its ability to generate results that can be understood. For large datasets, it can be computationally demanding and sensitive to parameter selections.

Notwithstanding these drawbacks, LDA is still an effective method for identifying latent themes and gleaning significant patterns from sizable text data sets.

5. Analysis & Insights

5.1 Sentiment Analysis Insights

I conducted sentiment analysis using two different methods, and the following are the key observations and insights gained from the analysis.

- Based on the two approaches, the sentiment time series, which is hourly plotted from September 28 to September 6, 2024, shows different patterns in sentiment behavior. Sentiment demonstrates significant swings using the Count Method (Figure 7A), with values falling between -20 and 40. Both positive and negative extremes are present throughout the period, indicating significant variability in sentiment captured by this method. Conversely, the Vader Method (Figure 7B) yields a trend that is more consistent, with sentiment values mostly remaining positive and falling between 0 and 25. Negative sentiment is, however, less common.

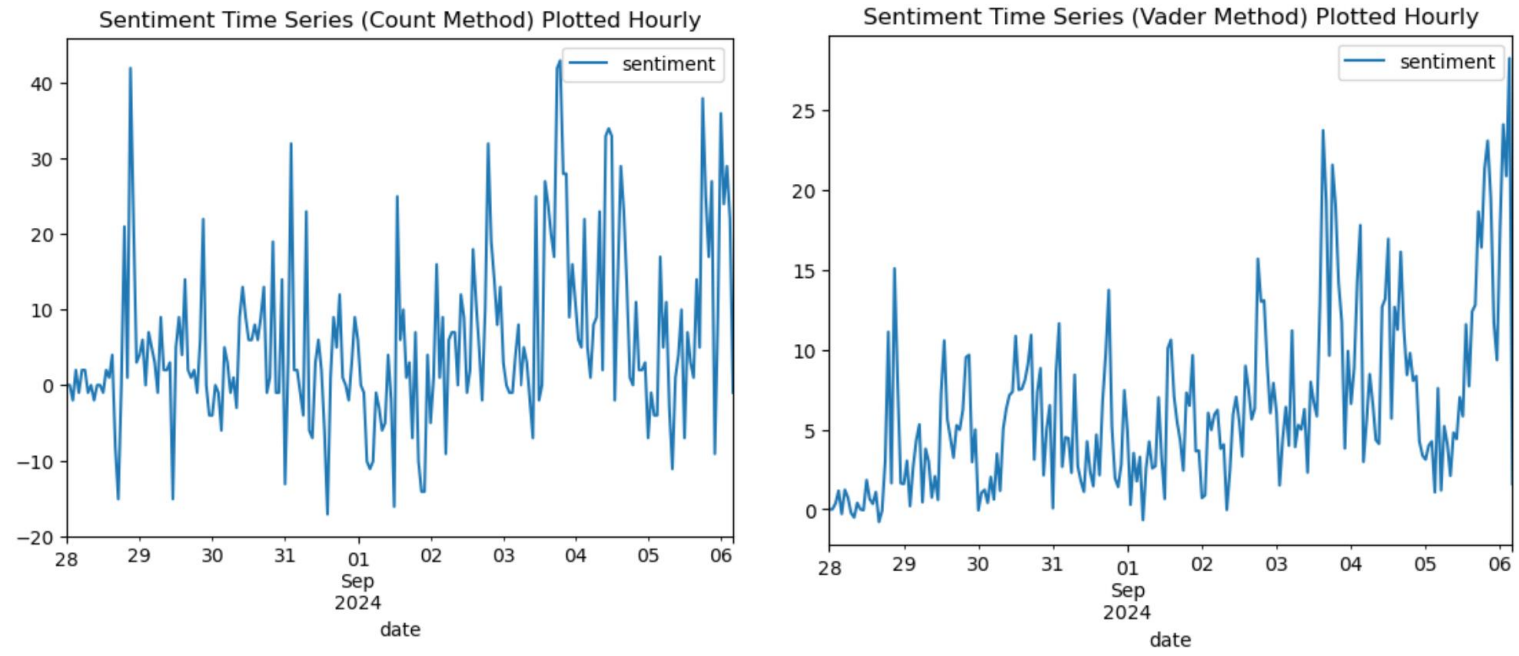

Figure 4: *The sentiment trend for the corpus using the Count method (Figure 7A, left) and the Vader method (Figure 7B, right) is displayed.*

- The time series of the Count Method reveals a much wider range of sentiment values, frequently alternating between positive and negative, pointing to an extremely erratic sentiment pattern. Notable peaks occur on August 29th and September 4th, when sentiment rises above 30. On the other hand, the Vader Method provides a more consistent positive sentiment trend with fewer spikes. For instance, while there are occasional spikes around comparable dates, they are typically less severe and stay below 25.
- The Count Method is more likely to record negative sentiment, with multiple values falling below zero, particularly on September 1st. However, the Vader Method mostly stays positive, indicating a more cautious method of identifying negative sentiment. This shows that while the Count Method is more sensitive to drastic changes in sentiment, the Vader Method might be more appropriate for spotting trends in neutral or positive sentiment.

5.2 Topic Modeling Insights and Analysis

Latent Dirichlet Allocation (LDA) was employed in this analysis to identify the underlying themes in our dataset. LDA is a well-liked topic modeling method that aids in locating word clusters that frequently occur together and offers insights into the primary themes found in the text data.

I found that 10 clearly separate clusters were achieved with the default values:

topicNum = 10

featureNum = 1500

Words such as ‘iphone’ and ‘phone’ were excluded before running the model because they were dominating the topics without providing meaningful insights into the actual themes of discussion.

We have included an Intertopic Distance Map that was made using multidimensional scaling to show the relationships and distinctness between the topics found by the LDA model. The distribution of the topics with respect to one another is shown in this map. The topics are clearly separated from one another on the map, suggesting that they each represent unique word clusters that are related.

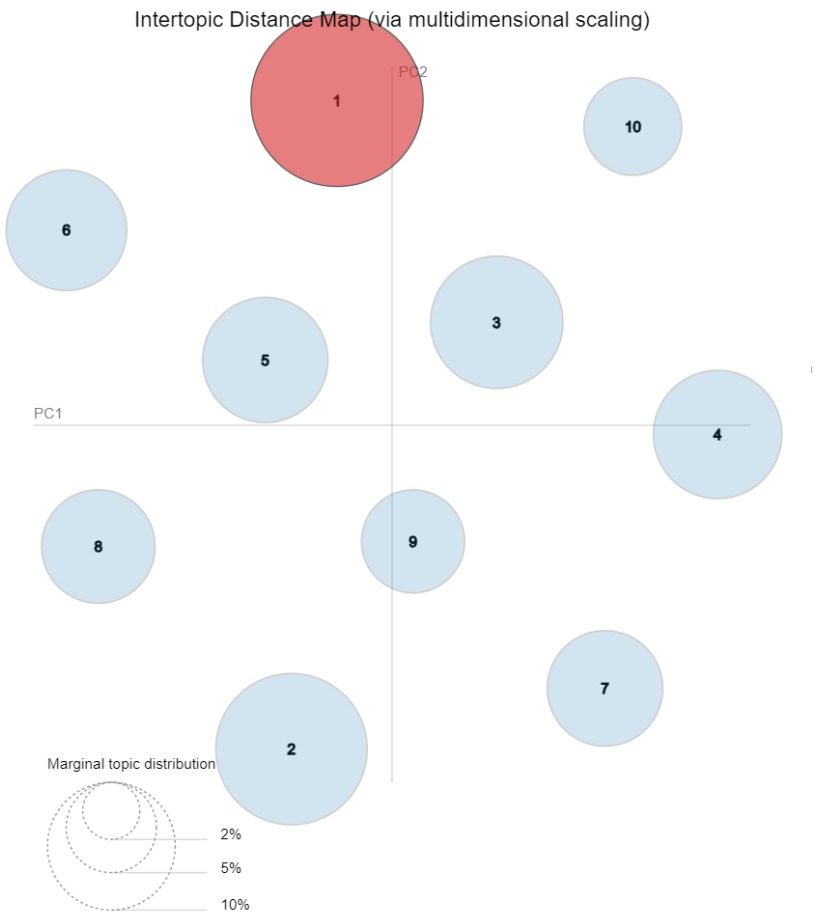

Figure 5: *Intertopic Distance Map*

Below are the found topics of discussion after running LDA:

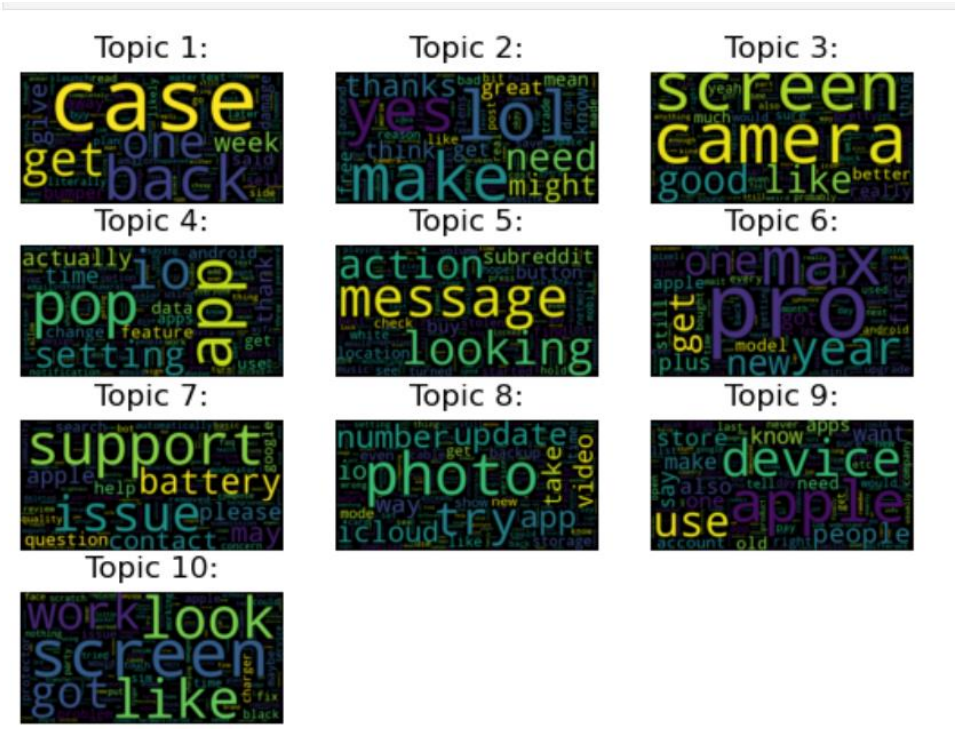

Figure 6: *Topics*

Following topics were identified:

Topic 0:

case back get one week give bumper said away damage sell literally side read likely

Topic 1:

lol make yes need thanks might think great get know mean free like lens bad

Topic 2:

camera screen good like really better much thing yeah sure pretty would probably sound well

Topic 3:

app pop io setting actually time thank feature data use android change get apps notification

Topic 4:

message looking action subreddit button buy location find see white check started turned stolen hope

Topic 5:

pro max year one new get plus first got apple model still upgrade every android

Topic 6:

support issue battery may contact please apple question help search google automatically removed concern bot

Topic 7:

photo try update app number icloud take way video io storage backup get even mode

Topic 8:

apple device use people store one know make want say also apps need old account

Topic 9:

screen look like work got problem fix apple sim charger issue protector black maybe time

Topic 0: The general use of apps and devices, including settings and data, appears to be the focus of this discussion. It probably has to do with conversations about the use and functionality of apps.

Topic 1: Reviews, updates, and the iPhone's battery life seem to be the main topics of discussion here. Terms pertaining to user experience and device performance are included.

Topic 2: It appears that this topic is about the hardware and UI components of the device, including the buttons, camera, and screen. It might have to do with how devices work and how users experience them.

Topic 3: Support and problems with getting in touch with them, fixing issues, and handling hardware or software problems are covered in this topic.

Topic 4: This section covers app management, which includes options for using and deleting apps.

Topic 5: This topic is more all-encompassing and appears to contain opinions and feelings, such as gratitude and remarks regarding the device.

Topic 6: This topic discusses the various iPhone models, including the latest iterations and their features, as well as their models, features, and general impressions.

Topic 7: It appears that this topic is connected to conversations about purchasing choices, queries about Apple products, and community interactions (such as subreddits and FAQs).

Topic 8: Device storage and associated matters, such as music, accessories, and backup, are covered in this topic.

Topic 9: Software updates, iCloud, and other relevant matters, such as glitches and version upgrades, are the main topics of discussion here.

All things considered, it seems that these topics encompass a wide range of conversations about iPhones, from features and individual viewpoints to technical problems and user assistance. These subjects offer a decent summary of the many topics people talk about when it comes to iPhones, which is helpful if the objective is to comprehend user attitudes and worries.

6. Conclusions

Our topic modeling results revealed distinct clusters of discussion topics, ranging from hardware and software features to user experiences and support issues. The Intertopic Distance Map highlighted clear separations between these topics, demonstrating the effectiveness of LDA in identifying unique themes within the dataset. The topics span various aspects of the iPhone experience, from device performance and app management to technical issues and customer support.

With the Vader Method offering a more consistent summary of positive and negative sentiments and the Count Method capturing larger sentiment fluctuations, the sentiment analysis provided a nuanced view of user emotions. This dual strategy gave a more comprehensive view of public opinions while highlighting the variation in user feedback.

All things considered, stakeholders looking to improve the iPhone's market position will find great value in the insights obtained from this analysis. Apple can more effectively address customer concerns, improve product features, and customize marketing strategies by comprehending the prevailing themes and user sentiments. In addition to helping to maintain user satisfaction, this analysis provides direct customer feedback for future innovations.

7. References

[1] J. Chan, " COSC2671 | Social Media and Network Analytics, Week 3”

[2] J. Chan, " COSC2671 | Social Media and Network Analytics, Week 4”

[3] J. Chan, " COSC2671 | Social Media and Network Analytics, Week 6”