Reddit Data Science Study: Sentiment, Trends and Communities Dynamics

1. Introduction

The problem statement at hand revolves around leveraging Reddit data to gain valuable insights and trends which is bolstered due to the fact that Reddit supports a vast variety of discussion and content without many restrictions that enables users to put forth their opinions free of reservations. Analyzing such posts and comments on Reddit can serve as a good opinion when compared to obtaining information from a single source which may be tainted due to restrictions/guidelines. So an opinion derived from the Reddit ecosystem can be considered as a fair representative of the views of the general population. One crucial problem we aim to address is the sentiment analysis of subreddits, as it is essential to understand how different topics evoke distinct sentiment patterns within their discussions. By analyzing sentiment in these online conversations, we can uncover whether certain topics tend to attract more positive or negative reactions, shedding light on the emotional undercurrents of the Reddit community. Additionally, identifying trending topics across various subreddits through topic modeling is crucial for staying updated on current interests and discussions in the Reddit ecosystem.

2. Background

Since the inception of transformers in the field of Natural Language Processing [1], there has been a substantial proliferation, leading to notable advancements. This transformation has yielded significant achievements in various tasks, including Sentiment Analysis, Question Answering, and Information Retrieval. In recent literature, a prevailing trend involves the application of transformer-based architectures to scrutinize social media platforms, seeking insights into emerging trends. For instance, during the COVID-19 pandemic, researchers conducted a study [2] to investigate the community's receptiveness to vaccination by analyzing sentiment in subreddits.

The outcomes of this study shed light on the sentiment analysis of these online communities during a critical period. Additionally, another study[3] utilized data from university-related subreddits and harnessed advanced natural language processing methods, encompassing RoBERTa and Graph Neural Networks (GNNs), to delve into the influence of the COVID-19 pandemic on emotional and psychological states. The findings of this research revealed a notable increase in negative sentiments during the pandemic, particularly in universities where in-person teaching was prevalent.

3. Experimental Setup

To realize our research objectives effectively, we rely on the cutting-edge techniques within the field of Natural Language Processing, with a particular focus on transformer models. These models have revolutionized the way we handle text data, offering unprecedented capabilities for understanding the sentiment within textual content and extracting trending topics from the vast landscape of Reddit discussions. After a comprehensive review of prior research, we determined that the two primary sources for Reddit data [4] were the PushShift API and Reddit API (utilizing PRAW). However, these sources became unviable due to recent Reddit policy changes. Consequently, we have detailed the dataset chosen for our study in Section 3.1.

3.1. Dataset

Given the recent restrictions imposed on accessing the PushShift API, we explored alternative data sources and acquired multiple Reddit Data Dumps to support our research endeavors. Among the various data dumps procured, one particularly comprehensive dataset aligns with the project's objectives. This dataset encompasses fundamental Reddit information, such as author details, score (computed through an undisclosed algorithm

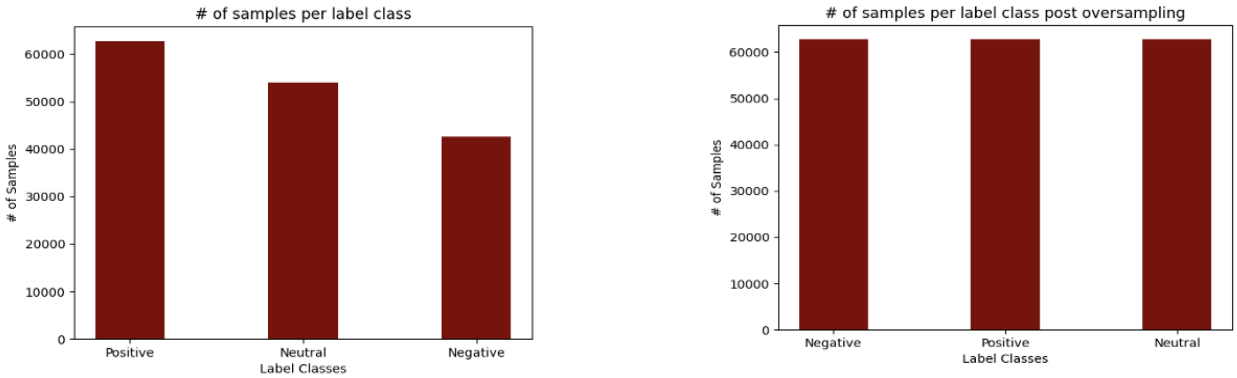

Fig.1 : Shows the class imbalance in the dataset(left). Oversampled dataset to resolve class imbalance (right).

utilizing up and down votes on posts), titles, subreddits, and more, organized into submissions and comments. The dataset includes posts belonging to all the subreddits, spanning from June 2005 to December 2022, and is substantial, totaling approximately 2.4TB of data stored as .zst files. To further supplement this we have data for all subreddits during the month of September 2023. Given the impracticality of processing the entire dataset, we employ filtering strategies to select specific subreddits of interest. To validate the methodology outlined in Section 3.2 of our research paper, we intend to utilize the Reddit data collected from posts between the years 2020 to 2022.

To facilitate effective preprocessing of our data, we initially convert the information from the .zst to .csv format, carefully selecting fields based on their relevance to the type of data under consideration. For submissions or posts, we extract the title, created_utc, and self_text, while for comments, we focus on the body and created_utc. This curated information is stored in .csv format and subsequently retrieved as a structured data frame. To enhance the text data we perform feature engineering, by first, preprocessing and then combining the title and self_text fields, facilitating a more comprehensive analysis.

Our preprocessing pipeline involves the removal of unnecessary special characters, links, and URLs from the combined textual data. Additionally, we address emojis and emoticons, converting them into textual representations of the corresponding feelings they express. Extraneous elements such as flags, map symbols, hashtags, user tags, and Chinese symbols are also systematically eliminated. Moreover, we convert the created_utc field into datetime format, streamlining temporal analysis and filtering processes.

For trending topic analysis, given the extensive dataset, we narrowed our focus to the years 2020 to 2022. Further refinement involves stop word removal, tokenization using the Natural Language Toolkit (nltk), and sentence creation. Subsequently, we filter out sentences with a length below five, ensuring that the retained text is of sufficient length to derive meaningful insights. This meticulous preprocessing approach lays the groundwork for robust and insightful trending topic analysis and sentiment analysis on the curated textual data.

3.2. Approach

We address two specific issues through analytical examination, constructing distinct models for each problem. Section 3.2.1 elucidates the modeling approach for sentiment analysis, while Section 3.2.2 outlines the methodology employed for trending topic identification. In Section 3.2.3, we conduct an experiment aimed at gaining a comprehensive understanding of the predominant topics within diverse subreddit communities and their corresponding sentiments.

3.2.1 Sentiment Analysis

In our investigation into sentiment analysis on Reddit posts, we have formulated a two-phase methodology to thoroughly evaluate the sentiments expressed. Initially, detailed in Section 3.1, we conducted data preprocessing and sourced posts from the r/happy and r/sad subreddits. Leveraging Vader's algorithm [5], sentiment labels were assigned. To ensure a robust ground truth we extracted positive labels exclusively from r/happy,

negative sentiments from r/sad, and neutral sentiments from both sources. To address class imbalance, as depicted in Fig.1, an oversampling technique [6] was implemented.

In our research methodology, we partitioned the data extracted from r/happy and r/sad subreddits, post oversampling, into 80% training data, comprising 150,000 rows, and allocated the remaining 20%, totaling 19,000 rows, for testing purposes. To enhance the predictive capabilities of handling new posts, we conducted fine-tuning on two Large Language Models (LLMs): specifically, BERT [7] and RoBERTa [8].

The baseline model, $BERT_{BASE}$, is characterized by 12 layers, 12 attention heads, and a hidden layer size of 768. As for the competing model, we fine-tuned a $RoBERTa_{BASE}$ model, sharing architectural similarities with $BERT_{BASE}$ but employing a distinct training approach. Both models underwent a 2-epoch training process, with a maximum sentence length set to 256 to mitigate potential training bias. Additionally, to address class imbalance and minimize skewness, balanced datasets were employed, ensuring an equitable representation of samples for each sentiment class.

3.2.2 Trending Topic Identification

For our examination of trending topic analysis on Reddit posts, our primary aim is to discern prevalent keywords and phrases, elucidating trends and their temporal evolution. To achieve this objective, we established a baseline model that employs the *Universal Sentence Encoder (USE)* on textual data. The USE generates 512-dimensional sentence embeddings using a transformer architecture, adept at handling diverse tasks and continuously refining embeddings based on encountered errors. These embeddings undergo K-Means clustering, effectively categorizing them into distinct topics. We then computed the distance between the centroid and each word in the vocabulary. Furthermore, for every word in the vocabulary, we computed its embedding by leveraging the USE model. The cosine distance between the centroid embedding and each vocabulary word was then calculated, revealing the top closest words associated with each topic.

In our pursuit of an advanced model, we employ *BERTopic* [9], a state-of-the-art methodology that harnesses BERT-based transformer models for refined topic modeling and analysis. BERTopic proves particularly advantageous for our objectives, thanks to its capacity to grasp contextual nuances and detect subtle shifts in language patterns. The process involves embedding the textual data using a sentence transformer with 768 dimensions.

The embedded data is then subjected to dimensionality reduction through *umap*, enhancing computational efficiency. Subsequently, clustering is performed using *hdbscan*, enabling the identification of coherent topics within the data. We have utilized hdbscan as it excels over K-means by automatically determining cluster shapes and sizes, offering robust performance in identifying clusters with varying densities and shapes, a flexibility that K-means, with its assumption of spherical clusters, may struggle to achieve. To further refine our analysis, we employ *c tf-idf* and we extract the top words from each topic along with their respective sizes. This comprehensive approach ensures that our trending topic analysis is not only sophisticated but also adept at capturing nuanced linguistic dynamics inherent in the evolving textual data.

Documents	Baseline	BERTopic	Llama2
['Which is better? Air pods gen 1 or gen 2?', Fake airpods pro 2', 'Airpods or beats?']	airpods,pro, noise, gen, beats, case	airpods, earpods, pods, earphones, headphone	AirPods and related accessories
['Macbook for gaming', 'Please help, is it right time to buy a Mac Air', 'Help with my MacBook']	macbook, apple, mac, help, macos	macbook, mac, for, apple, is, m1	Apple Product Support
['Dynamic island is crushing it', 'New possibilities with Dynamic island', 'Dynamic island more like pill']	pill, activities, island, dynamic, peninsula	dynamic, rollercoasters, patterns, reachability, interact	Dynamic Island Activities
['Voice to text', 'Sync Voice iPhone Voice Memos to MacBook without iCloud', 'the voice']	voice, memos, recognition, recording, dictation	voiceover, spoken, speech, voicenote, voicememohelp	Voice Recognition Technology

Table 1: Comparison of the labels generated from the various model utilized using data from the Apple Community

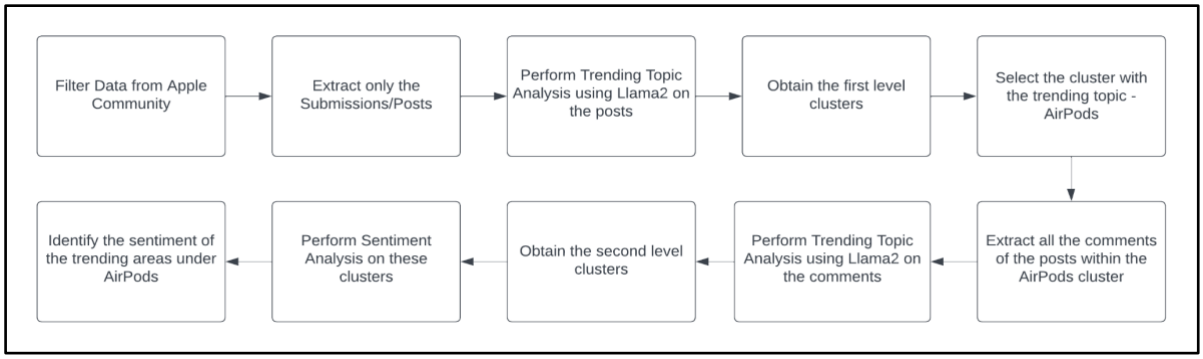

Fig. 2: Flowchart showing the steps for performing sentiment analysis on the identified topics.

During our model exploration, manual annotations were applied to designate cluster names. To enhance the coherence of these labels, we adopted Llama-2, a generative model, for cluster annotation. Harnessing the sophisticated capabilities of Llama-2, we effectively assigned meaningful descriptors to the clusters, establishing a robust framework for comprehending the inherent structures within our dataset. This methodology not only facilitates the interpretation of intricate cluster formations but also elevates the overall lucidity and interpretability of our research findings. The incorporation of Llama-2 represents a noteworthy progression in our analytical approach, significantly contributing to the precision and depth of insights gleaned from our cluster analysis. Table 1 shows a comparison of the clusters/labels generated for the same set of data from the Apple Community between Baseline, BERTopic, and Llama-2.

3.2.3 Sentiment Analysis on Identified Topics

In order to achieve a more thorough analysis, we extensively explored a specific domain, employing both of our models in an effort to extract valuable information. For this, we decided to go solely with the Apple community and obtain data from multiple subreddits connected with the Apple community including r/apple, r/applecare, r/appletv, and r/applesupport. We filtered out the posts from 2020 to 2022 to ensure the recency of the information that we obtained.

The flow chart as seen in Fig.2 shows a visualization of the implementation logic. Initially we consider only the submissions or posts from the aforementioned subreddits. From section 3.2.2 we have established that BERTopic + Llama2 performed the best for identifying the trending topics. So, we utilize this on the posts and obtain the first-level clusters. Following that we dive deeper into the 'AirPods' cluster. To highlight the most discussed feature in this cluster, we perform a second-level analysis of the posts and comments under this cluster. To consolidate the community's feedback on the features extracted from the second-level analysis, we perform sentiment analysis using the RoBERTa model.

4. Results and Analysis

To facilitate a detailed examination of the experimental setups outlined in Section 3.2, we categorize our findings into three distinct subsections, as illustrated below. This division allows for a more focused analysis within each section. Additionally, this structured approach enhances the clarity and comprehensibility of our results, aiding in the extraction of meaningful insights from the research outcomes.

4.1. Sentiment Analysis

To gauge the performance of the sentiment analysis models we have used evaluation metrics like F-score, precision, and recall, on the four scenarios—BERT and RoBERTa, each fine-tuned on both balanced and unbalanced datasets. Table 2 shows the results obtained from the various experiments.

The performance metrics clearly indicate the superiority of RoBERTa over BERT in our sentiment analysis task. RoBERTa consistently demonstrated higher precision, recall, and F1 scores across all sentiment labels—Negative, Neutral, and Positive. Fig. 3 shows the confusion matrix of RoBERTa, fine-

SCENARIO	PRECISION			RECALL			F-SCORE			ACCURACY (%)
	POSITIVE	NEUTRAL	NEGATIVE	POSITIVE	NEUTRAL	NEGATIVE	POSITIVE	NEUTRAL	NEGATIVE	
BERT trained on unbalanced dataset	0.9847	0.9891	0.9805	0.9894	0.9791	0.9860	0.9870	0.9841	0.9832	98.50
RoBERTa trained on unbalanced dataset	0.9941	0.9900	0.9859	0.9908	0.9888	0.9922	0.9925	0.9894	0.9890	99.05
BERT trained on balanced dataset	0.9848	0.9879	0.9843	0.9875	0.9800	0.9893	0.9862	0.9839	0.9868	98.57
RoBERTa trained on balanced dataset	**0.9949**	**0.9965**	**0.9928**	**0.9973**	**0.9895**	**0.9974**	**0.9961**	**0.9930**	**0.9951**	**99.47**

Table2: Shows the evaluation metric results for all the four scenarios for sentiment analysis.

tuned on the balanced dataset. The enhanced performance of RoBERTa can be attributed to its optimized training methodology, which refines the model's ability to capture intricate patterns and nuances in sentiment expressions.

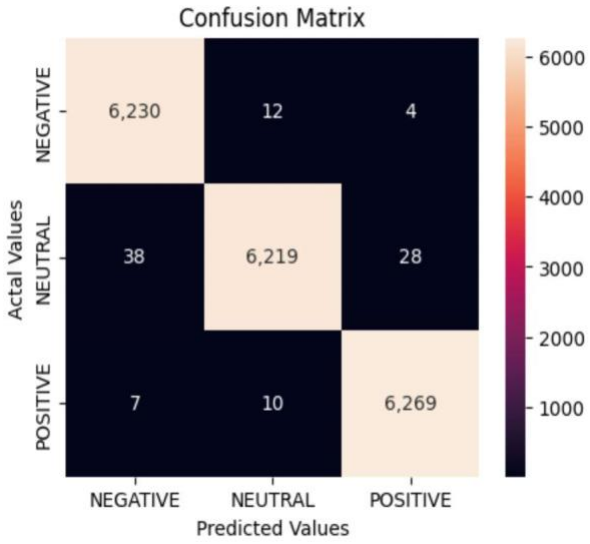

Fig. 3: Confusion Matrix for RoBERTa (balanced dataset)

4.2. Trending Topic Identification

To assess the efficacy of our trending topic models, we conducted experiments using data from two distinct subreddits, namely r/apple and r/explainlikeim5. The rationale behind selecting these subreddits lies in their divergent nature: r/apple, being more generic in discussions, and r/explainlikeim5, characterized by its potential for unique and random topics. The analysis aimed to provide insights into the both the models' effectiveness across different subreddit dynamics. For the r/apple subreddit, we initially applied our baseline model, generating clusters as depicted in Fig. 4.

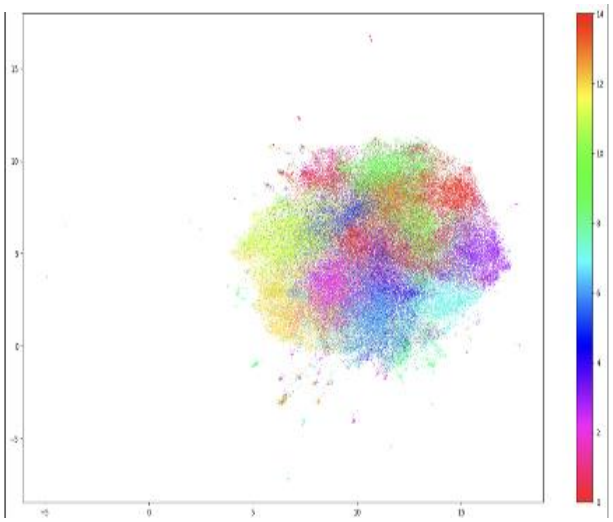

Fig. 4: Clusters formed using the baseline model

Subsequently, the same data underwent analysis using our BERTopic model, resulting in the cluster pattern illustrated in Fig. 5.

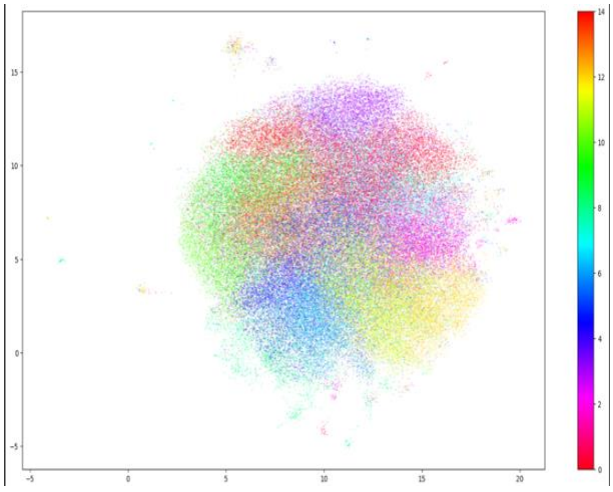

Fig. 5: Clusters formed using the BERTopic

Given the indeterminate number of topic groups, we selected the top ten topics based on their size for graphical representation. The extracted topics from both models were then plotted against their respective sizes in Fig. 6 and Fig. 7.

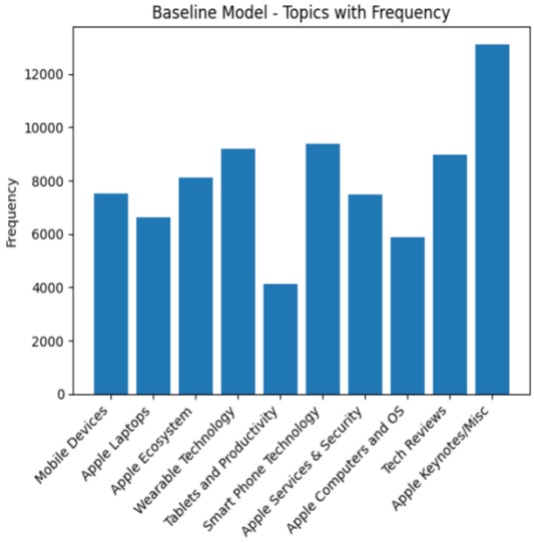

Fig. 6: Trending Topics for r/apple – baseline models

Notably, the clusters obtained from the BERTopic model exhibited clearer delineation compared to the baseline model. This is particularly evident in the accurate identification of trending topics, such as the legal dispute between Epic Games and Apple in the BERTopic model-generated topics.

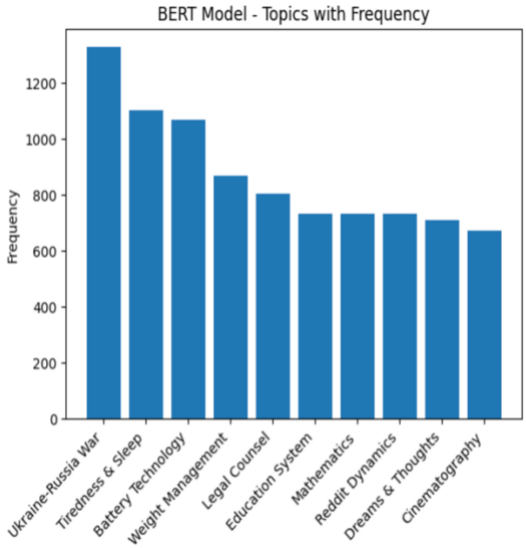

Fig. 7: Trending Topics for r/apple – BERTopic model

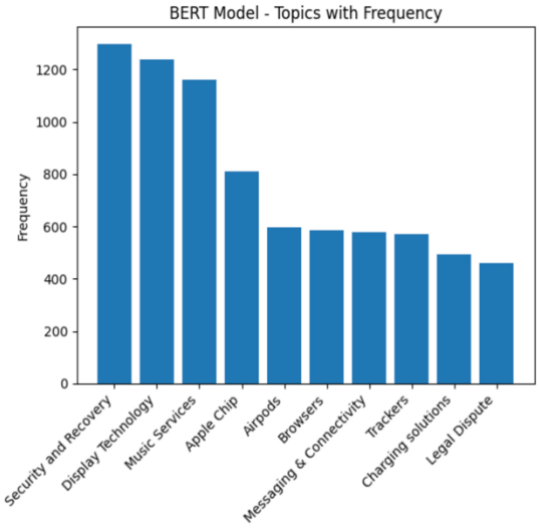

Fig. 8: Trending Topics for r/explainlikeim5 – baseline model

Moving on to the r/explainlikeim5 subreddit data, we subjected it to both the baseline and BERTopic models, yielding topics visualized in Fig.8 and Fig.9.

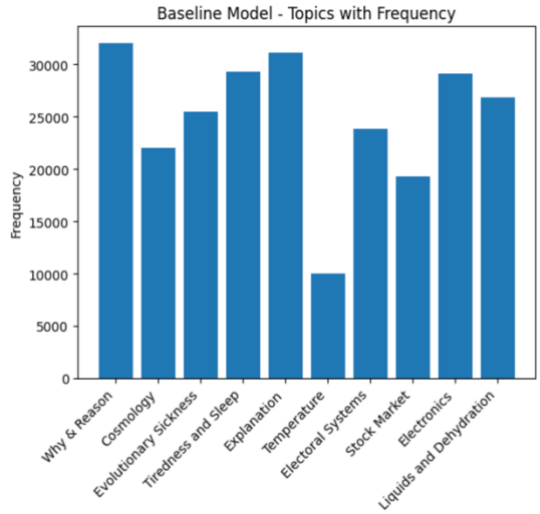

Fig. 9: Trending Topics for r/explainlikeim5 – BERTopic model

The BERTopic model demonstrated a distinct advantage, effectively capturing meaningful trending topics, exemplified by the identification of the Russia-Ukraine war—a prominent ongoing event during the selected timeframe. In contrast, the baseline model tends to produce topics dominated by common words like 'explaining' or 'reasoning,' which, while reflective of the subreddit's basic essence, lacked the specificity

required for discerning trending topics. This comparative analysis underscores the enhanced performance and relevance of the BERTopic model in extracting meaningful trends from diverse and dynamic textual data sources.

4.3. Sentiment Analysis + Topic Modelling

In our final analysis, our investigation commenced by observing the discourse within the Apple community on Reddit. Fig.10 illustrates the overall sentiment surrounding the top five trending clusters identified through the initial analysis outlined in Section 3.2.3. Notably, clusters such as 'Voice Recognition' and 'Dynamic Island Activities' exhibit significantly positive sentiment within the user community.

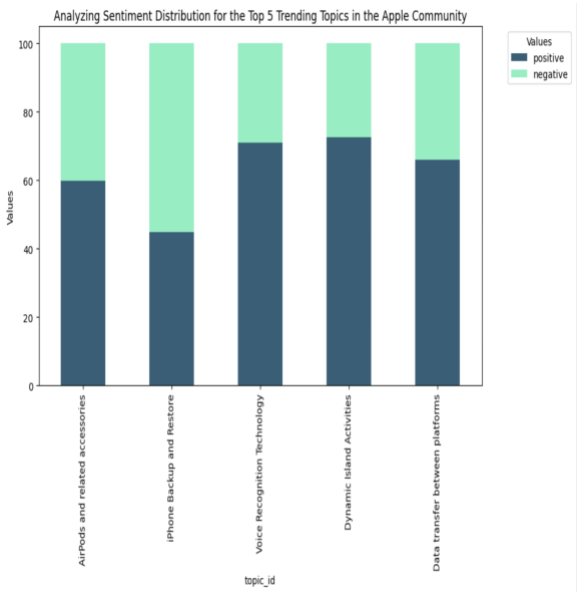

Fig. 10 : Sentiment Analysis for the top five trending topics in the Apple Community

To delve deeper into the nuances of each cluster, a second-level analysis was conducted. Specifically, we focused on the AirPods cluster, filtering posts and comments associated with this cluster to extract subtopics. These subtopics serve as representatives of features or issues related to AirPods, prompting a subsequent sentiment analysis to gauge the community's inclination.

The results of the sentiment analysis, detailing positive and negative sentiment values for each subtopic are obtained. Fig.11 elucidates the contribution of major features with positive sentiment, while Fig.12 illustrates the contribution of issues with negative sentiment.

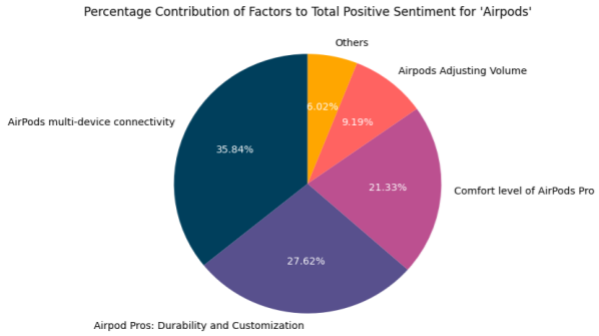

Fig. 11: Percentage contribution of Positive sentiment of the trending features in 'AirPods'.

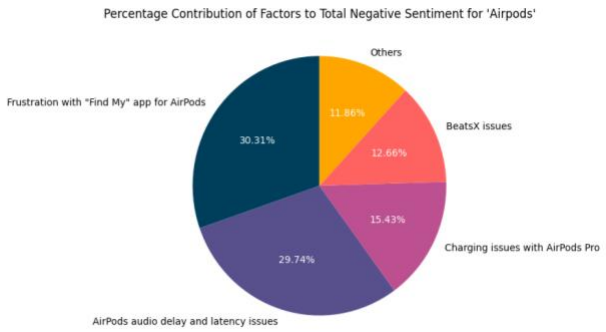

Fig. 12: Percentage contribution of Negative sentiment of the trending issues in 'AirPods'.

This granular sentiment analysis facilitates the identification of both the merits and concerns expressed by the community regarding AirPods. Notably, a prevailing topic with negative sentiment pertains to the 'Find My' device on AirPods. Examining data spanning from 2020 to 2022 reveals a persisting issue during this period, prompting Apple to address it through an update to the 'Find My' app for AirPods which can now identify the location of the AirPods to a pinpoint position. This adaptive response underscores the potential of leveraging such insightful information for product enhancement, illustrating the boundless possibilities inherent in utilizing the community feedback from sources like Reddit for continual improvement.

5. Conclusion

Within the scope of this research endeavor, our emphasis centered on the Apple community to gain profound insights into consumer feedback leveraging Reddit data. Employing state-of-the-art natural language processing techniques, specifically Transformers, we attained notable accuracy in our sentiment analysis task and successfully extracted trending topics from the relevant subreddits. This targeted approach, honing in on the Apple domain, afforded us the opportunity to discern the diverse subjects discussed by Reddit users, shedding light on the features and issues prevalent in their conversations on open-source platforms. This focused investigation enabled a comprehensive exploration of the nuanced discussions within a specific community, enriching our understanding of user perspectives.

6. Future Work

In the context of sentiment analysis, it is worthwhile to consider the utilization of Generative AI models such as Llama-2. These models have the potential to comprehend the intricate structure of Reddit posts and comments, thereby facilitating a more comprehensive and steadfast analysis. Subsequent phases of this project can involve extending the analytical scope to encompass various Reddit communities, including but not restricted to political, sports, and science communities. This broader exploration aims to discern the sentiment prevailing within these communities and delve into the factors influencing these sentiments through a second-level analysis.

7. References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I.

Attention is all you need. CoRR, abs/1706.03762, 2017.

[2] Chad A. Meltona, Olufunto A. Olusanyab, Nariman Ammarb, Arash Shaban-Nejad. Public sentiment analysis and topic modeling regarding COVID-19 vaccines on the Reddit social media platform: A call to action for strengthening vaccine confidence. In: Journal of Infection and Public Health 14 (2021) 1505–1512

[3] Yan T, Liu F (2022) COVID-19 sentiment analysis using college subreddit data. PLoS ONE 17(11): e0275862.

[4] Li, S.; Xie, Z.; Chiu, D.K.W.; Ho, K.K.W. Sentiment Analysis and Topic Modeling Regarding Online Classes on the Reddit Platform: Educators versus Learners. *Appl. Sci.* 2023, *13*, 2250.

[5] Is it possible to do sentiment analysis on unlabeled data using bert feat vader experiment

[6] R. Mohammed, J. Rawashdeh and M. Abdullah, "Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results," 2020 11th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 2020, pp. 243-248, doi: 10.1109/ICICS49469.2020.239556.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In North American Association for Computational Linguistics (NAACL)

[8] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “RoBERTa: A robustly optimized BERT pretraining approach,” arXiv preprint arXiv:1907.11692, 2019.

[9] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794 (2022). https://doi.org/10. 48550/ARXIV.2203.05794

[10] Fine tuning BERT for sentiment analysis

[11] HuggingFace – Roberta page

[12] HuggingFace – Llama-2