Untitled

GUI GAO

2023-10-28

Contents

1 Abstract

2 Background and Motivation

3 Purpose/objective of this project - Forethought Project background

4 Data and Methodology

4.1 Data Sources

The four datasets analyzed in this report comprise posts and comments are pulled from Facebook and Instagram social media, regarding the four brands Amaysim, Optus, Telstra, and Vodafone.

4.2 Data Cleaning

4.2.1 Preliminary cleaning:

In the data preparation phase, several vital functions are used. The `rep_str` function is used to set up a replacement text for Emojis. While the `str_replace_all` function is deployed to systematically replace Emoji texts in both the posts and comments datasets.

Additionally, the `gsub` function plays a crucial role in the substitution of spaces, commas, and other special characters within variables, ensuring data is tidy. For more targeted character adjustments, regular expressions are employed. The `separate` function is used for splitting a character column into two columns to meet specific requirements. Extracting specified strings is facilitated by the `str_extract` function.

Ensuring uniformity in date formats is a critical step. Both posts' publish dates and customers' comment dates are standardized using the `mdy` function. Furthermore, dates denoted as *two weeks ago*, *one day ago*, *three months ago*, and so on are harmonized into a standard date format using the `add_with_rollback` function.

4.2.2 Deep cleaning (for the development of our R Shiny application):

Utilizing the `select` function to filter out brand names and post URLs from the posts dataset, which are then stored in a separate dataset labeled *FB/IG_url*. To ensure uniformity, brand names are standardized within the posts dataset. The `pivot_longer` function converts the data for customer actions, comments, and shares on Facebook, as well as likes, comments, and views on Instagram, into a long format. It is convenient to use the `facet_wrap` function to draw multi-panel time series plots about different customer reactions when making R shiny. This feature allows for a clear and organized presentation of data, making it easier for users to compare and analyze trends across different response metrics.

The comments dataset and *FB/IG_url* are in-joined using the `inner_join` function to match which brand each customer's comments belong to. Furthermore, a crucial aspect of the data preparation involved performing an inner join operation using the `inner_join` function between the comments dataset and the *FB/IG_url* dataset. This operation could establish clear associations between customer comments and their respective brands.

Acquiring a set of stop words utilizing the `get_stopwords` function, employing the snowball lexicon. These stopwords are stored in a dataset named *stopword*. Both posts and comments are separated into individual words or pairs of words using the `unnest_tokens` function. This step is essential for further analysis. The *stopword* dataset is reintegrated using a back-join operation with the `anti_join` function, effectively removing the identified stop words. The resulting dataset is stored for subsequent use in generating wordclouds, sentiment analysis plots, and other analytical outputs.

Introducing a new variable pertaining to brands within the themes of posts dataset with the `append` function. Merging the datasets pertaining to themes of posts from both Facebook and Instagram, using the `rbind` function to ensure a unified and comprehensive view of brand-related.

4.3 Data Limitations

- Limited historical data: Access to historical data through APIs poses a constraint on long-term trend analysis. For example, Facebook's Graph API typically allows access to data from the past 90 days. This limitation necessitates a focused approach to extracting and interpreting data within this time frame.
- Sample bias: It is crucial to recognize that the customer base actively engaging on social media platforms may not fully represent the broader population. This inherent bias can influence the insights derived from the data, prompting a cautious approach to generalizing findings.
- Sparse data: Data sparsity emerged as a significant consideration. Not all customers furnish comprehensive information, and not all posts contain pertinent data for analysis. This scarcity can impact the depth and scope of analytical endeavors.
- Data quality: An additional challenge arose from the diverse and sometimes unstructured nature of customer-generated content. Comments often encompassed a range of linguistic variations, including slang, potential spelling errors, and non-standard language. These intricacies presented hurdles for natural language processing tasks.

4.4 Data Dictionary

4.5 Methodologies

4.5.1 Time series method

Employing a time series plot to illustrate the trend of customer reactions to brand posts over time on both Facebook and Instagram. Each brand is represented by a distinct color, offering a clear visual distinction. To further enhance visualization, use the `facet_wrap` function to create multiple panels. For Facebook, these panels encompass a combination of reactions, comments, and shares, while for Instagram, they comprise likes, comments, and views. This segmentation enables users to quickly discern the trends in customer engagement metrics over time. This approach provides an effective means to assess the popularity of each brand's posts across both platforms. The use of color coding and facet panels enhances the visual clarity, allowing for a detailed examination of customer reactions.

4.5.2 Natural Language Processing (NLP) method

While gauging the 'heat' of each brand's posts is informative, it is essential to delve deeper into the insights that social media data offers. Unlike conventional surveys or questionnaires, individual expressions on social media are highly personalized. Therefore, employing advanced Natural Language Processing (NLP) techniques becomes instrumental in extracting nuanced insights. By harnessing NLP methods, such as word cloud mapping, sentiment analysis, and customer behavioral analysis, users can gain a more profound understanding of prevailing social media trends and brand performance. These techniques can dissect the language and sentiments expressed by customers, providing a comprehensive view of public opinions and attitudes. Incorporating NLP methodologies not only unveils hidden patterns but also allows us to capture sentiments that might elude traditional analysis approaches. This method, in turn, equips a more accurate and insightful assessment of social media trends and customer behavior.

4.5.3 OpenAI Large Language Model API

In the realm of social media data analysis, using the power of the openAI Large Language Model API for textual summarization. API is a critical tool in extracting textual topics from the vast pool of social media content by filtering out extraneous and less relevant information and obtaining the most meaningful topics, providing users with a clear and concise overview for informed decision-making and strategic planning.

4.5.4 R shiny for visulization

Using the R shiny app to display the visualization results. As shown in the Figure 1, it is the layout interface of the whole R shiny. The title is Social Media Analysis, which corresponds to the entire project.

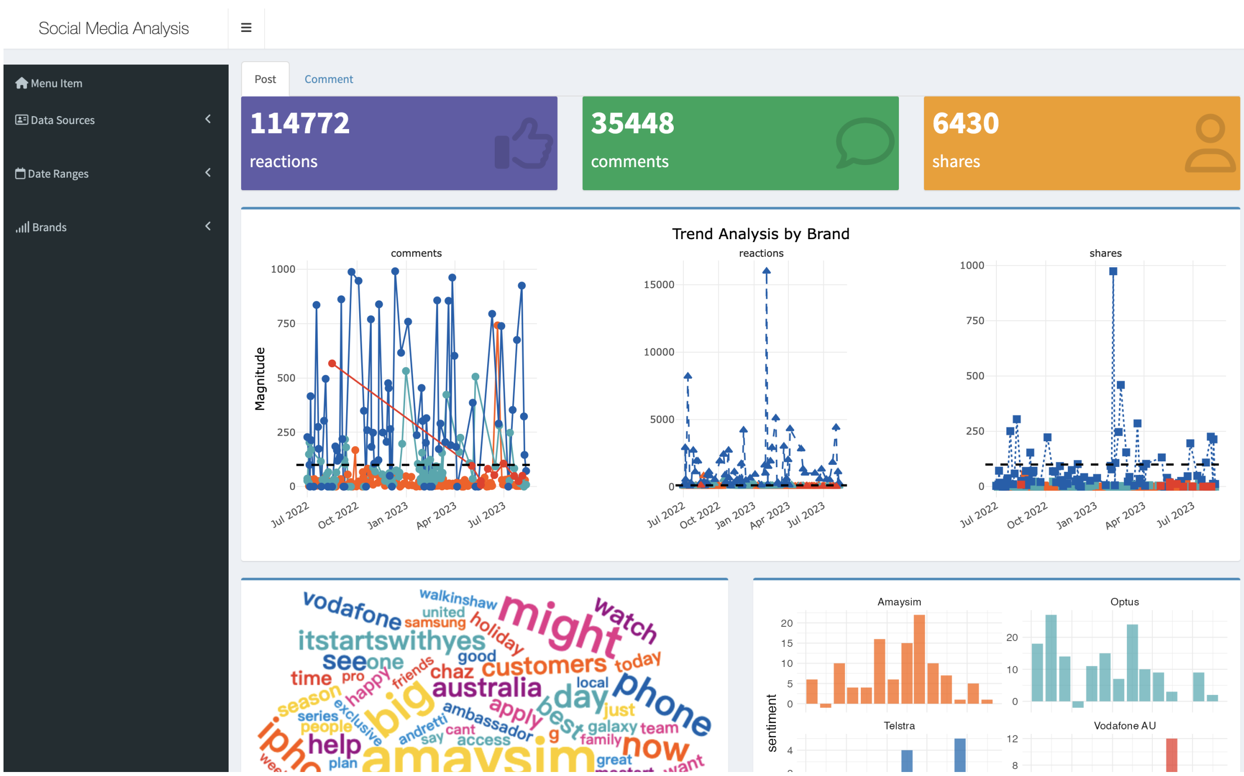

Figure 1: R shiny interface

On the left side, under Menu Item, users find a combination of R Shiny's dropdown controls. The first control as shown 2 is the data source, which allows users to switch between Facebook and Instagram datasets dynamically.

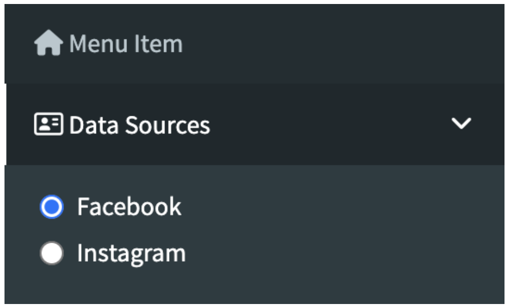

Figure 2: Data source control

The second, Date Ranges as shown 3, provides a feature-rich control for precise time range filtering, ensuring analysis accuracy.

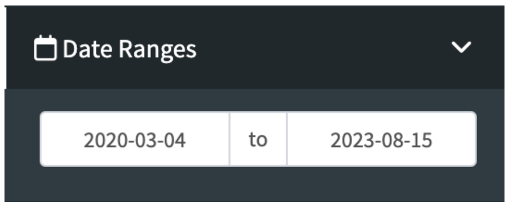

Figure 3: Date ranges control

The third control Brand as shown 4, allows users to explore specific telecom operators like Amaysim, Optus, Telstra, and Vodafone. To enhance usability, convenient *Select All* and *Clear All* options are provided for effortless brand selection management.

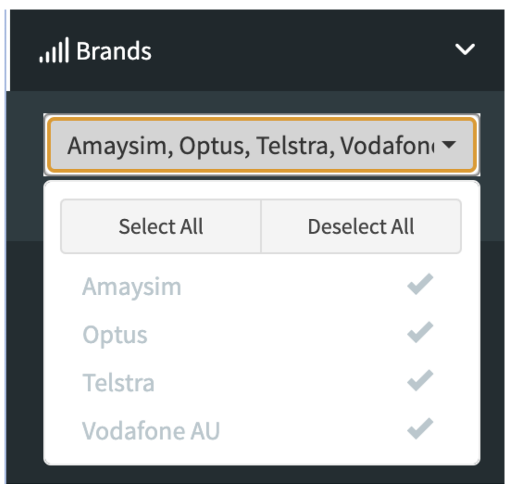

Figure 4: Brand control

5 Fingdings

5.1 Posts Findings

5.1.1 Value boxes

Using the `valueBox` function within R Shiny to craft three visually engaging value boxes, spotlighting key summary statistics related to customer responses over time. The three value boxes provide real-time insights into the evolving metrics: the count of reactions, comments, and shares on Facebook 5 or the aggregate sum of likes, comments, and views on Instagram 6, all contextualized within the current filtering parameters. Notably, the data underscores that posts on Instagram generate a notably higher level of customer interactions compared to Facebook.

Figure 5: Facebook value boxes

Figure 6: Instagram value boxes

Furthermore, an in-depth analysis reveals that posts affiliated with the Telstra 7 brand exhibit a remarkable surge in engagement, surpassing that of the other three brands. This insight emphasizes Telstra's notable influence and resonance within social media.

Figure 7: Facebook value boxes of Telstra

5.1.2 Time series plot

Time-series plots 8 are pivotal in providing a visual narrative of key customer engagement metrics within the data. They offer a clear avenue to monitor the ebb and flow of customer interactions, discern brands exhibiting notable engagement, pinpoint specific times and dates of heightened activity, and ultimately gauge the popularity and influence of social media content.

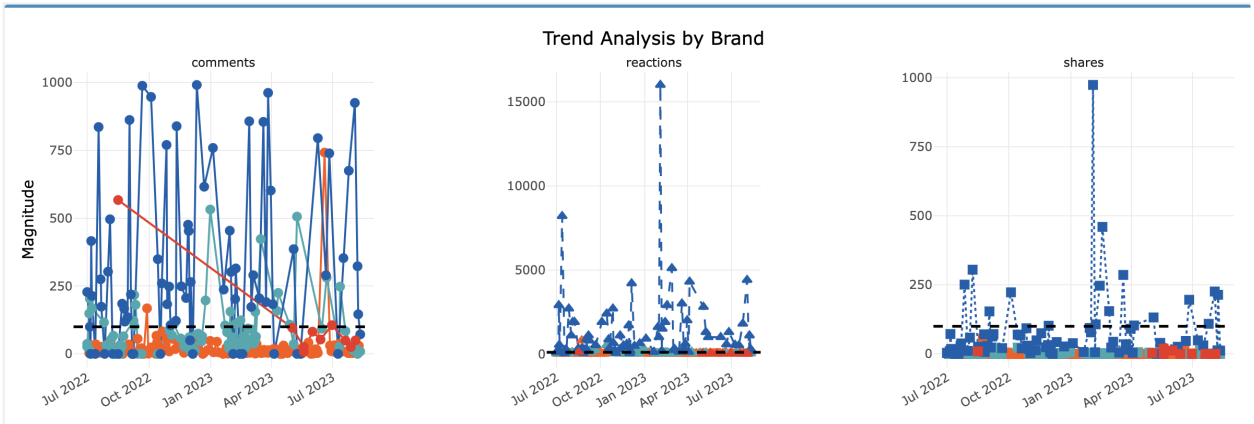

Figure 8: Time series plot

For instance, delving into the Facebook dataset, a standout observation was Telstra's post 9 on February 3, 2023. This post garnered an exceptionally high number of customer reactions and shares. The post, humorously captioned *POV every dads belt in 1999*, cleverly references a specific period (1999) and taps into a universally relatable parenting experience. By addressing a potentially severe topic like discipline in a light-hearted manner, the post strikes a chord with individuals who came of age during that era. This clever use of nostalgia triggers a wave of strong emotions and memories, making customers engage actively with the post. This analysis underscores the potent impact that well-crafted content can wield, resonating deeply with audiences and fostering heightened engagement levels.

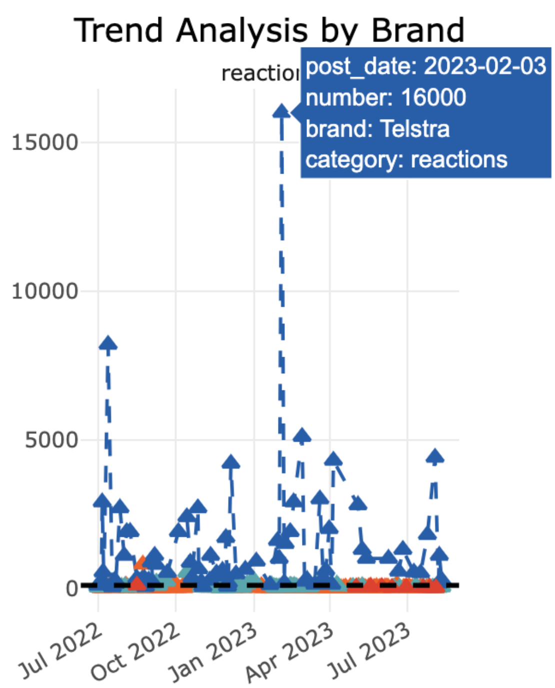

Figure 9: Time series plot of Telstra

In the time series plot 10 of the Instagram dataset, a vast gap arises in the likes number. This post, jointly published by Optus and the renowned Australian racing driver Daniel Ricciardo, garnered an astounding

330,000 likes. The involvement of a highly influential celebrity like Ricciardo significantly amplifies the post's visibility and impact. This observation underscores a strategic avenue for brands - collaborating with well-known personalities. Such partnerships enable brands to harness the extensive reach and influence of celebrities, expanding their follower base and extending their brand's outreach. This symbiotic relationship not only enhances brand visibility but also fosters a broader and more engaged audience.

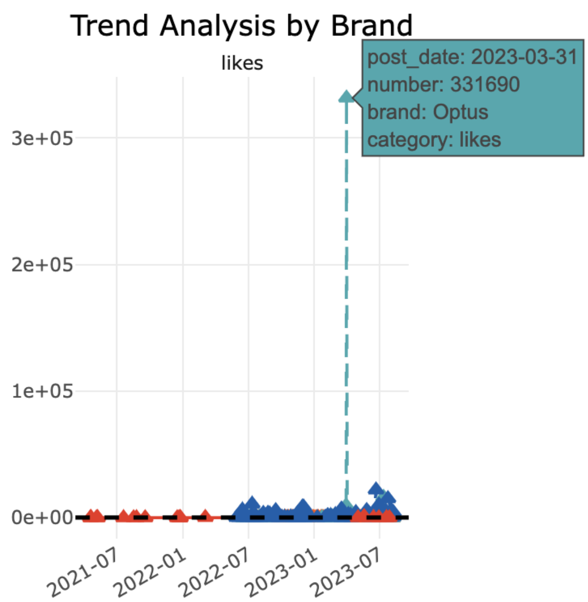

Figure 10: Time series plot of Optus

5.1.3 Wordcloud plot

The word cloud shows the top 60 words with the highest word frequency in the post text for the current filter. Words that occur more frequently are visually emphasized with larger font sizes in the word cloud map. In the Facebook dataset 11, the word *amaysim* emerges with the highest frequency, tallying 31 occurrences. Other commonly used words include *big*, *caption*, and *event*. Meanwhile, within the Instagram dataset 12, *optus* takes the lead with a frequency count of 70 instances. Additional prevalent terms encompass *big*, *love*, *amaysim*, and more. These frequently occurring words exemplify common themes prevalent in social media content. They serve as critical indicators of the prevalent topics and discussions within the realm of social media data.

Figure 11: Word cloud of Facebook posts

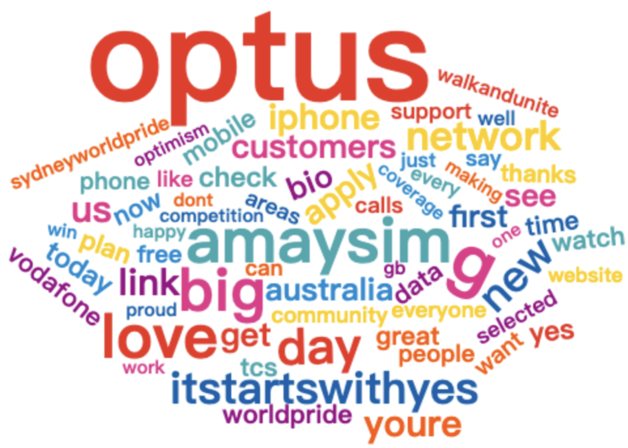

Figure 12: Word cloud of Instagram posts

5.1.4 Net sentiment plot

The net sentiment plot 13 serves as a visual representation of different brand posts over time about the balance between positive and negative sentiments. A positive value on the y-axis indicates that posts published by the brand within a given month exhibit an overall positive sentiment. In contrast, a negative value signifies a prevailing negative sentiment. Notably, a majority of posts across both Facebook and Instagram social media platforms manifest an overall positive sentiment. It is imperative to highlight a trend observed in the Facebook dataset. Telstra's posts experienced six months with a net negative overall sentiment. Overall, the net sentiment plot underscores a clear correlation between positive sentiment and heightened engagement rates. Customers exhibit a greater propensity to share content that evokes positive emotions, underscoring the pivotal role of sentiment in fostering customer interactions.

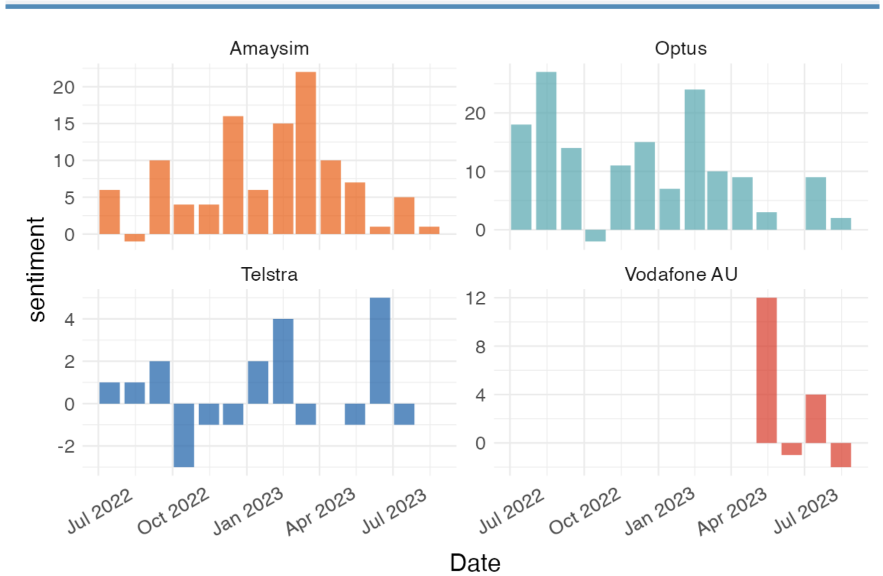

Figure 13: Net sentiment plot

5.1.5 Different themes plot

The Column plot, combines with the `facet_wrap` function for multi-panel display, presents a clear visual representation of the number of topics associated with different branded posts. This feature aids users in swiftly discerning and categorizing pertinent topics related to branded content across social media platforms. When correlated with the earlier time series plot, this visualization offers an additional layer of insight, shedding light on which types of topics wield the most resonance with customers. This valuable information empowers brands to adjust their marketing strategies, ensuring they remain attuned to evolving consumer interests and preferences.

In the Facebook 14 and Instagram 15 datasets, a distinct trend emerges: Amaysim predominantly centers its posts around themes of customer appreciation, engagement, and loyalty. This strategic focus indicates an effort by Amaysim to bolster customer satisfaction and fortify brand loyalty. The consistent expression of gratitude towards customers and positive interaction further strengthens these crucial relationships. This approach has profound implications for customer retention. Over time, individuals who experience higher levels of satisfaction and a sense of loyalty are markedly more inclined to continue availing of Amaysim's services. Their reduced likelihood of exploring offerings from competing brands underscores the enduring impact of Amaysim's customer-centric approach.

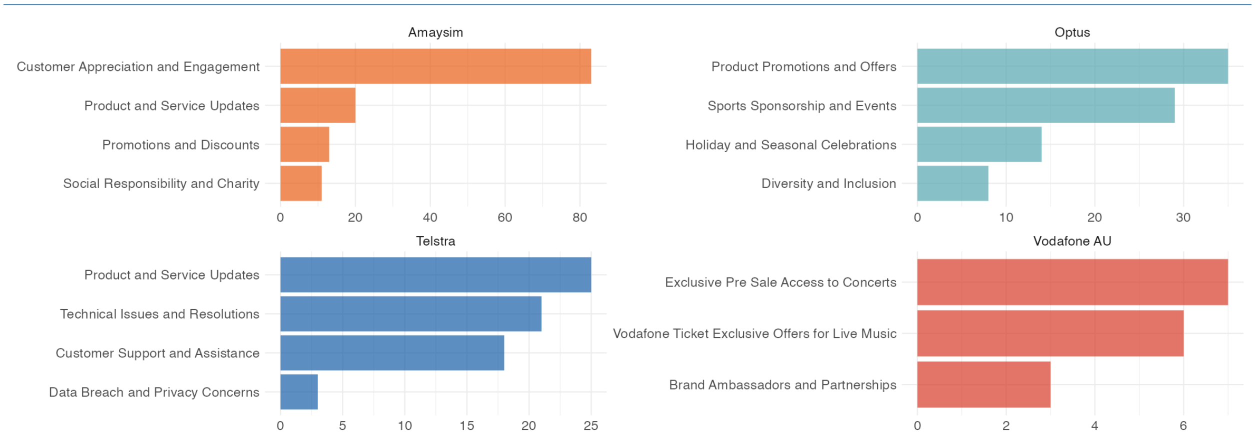

Figure 14: Themes plot from Facebook dataset

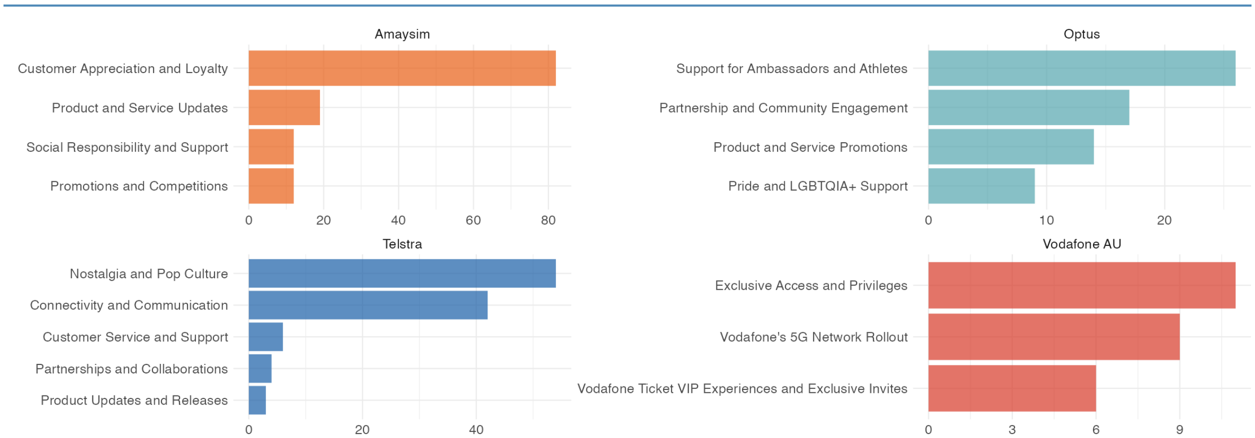

Figure 15: Themes plot from Instagram dataset

Optus and Telstra, in contrast, their posts predominantly focus on topics related to product and service promotions as well as updates. This strategic emphasis equips them to maintain a competitive edge in the market. By leveraging social media platforms, they keep customers well-informed about the latest product features and enhancements. This proactive approach drives sales and encourages customers to embrace new offerings readily. The deployment of promotional and update-centric posts further enables these brands to vividly showcase the value and benefits inherent in their products and services. This approach not only solidifies their existing customer base but also serves as a potent attractor for new customers.

5.2 Comments Findings

5.2.1 Bigram network

The bigram network 16 visualization method unveils the interrelationship between consecutive word pairs within customer comments. The intensity of gray shading surrounding more frequently occurring word pairs provides a visual cue. This technique is a powerful tool for extracting insights into co-occurring terminologies, illuminating specific language patterns and expressions commonly employed in customer comments. Bigram can highlight key phrases or expressions that are frequently mentioned in customer comments, which helps to identify recurring themes. The resultant bigram network plot serves as a window into specific sentiments and feedback associated with a product, service, or overall experience. This dive into customer behavior and preferences equips brands with a heightened comprehension of their audience. Remarkably, each brand's bigram network prominently centers on pivotal terms like *customer*, *phone*, and *service*, among others. This focal point underscores that customers frequently invoke brand-associated products and services in their comments, affirming their relevance and resonance in customer discourse.

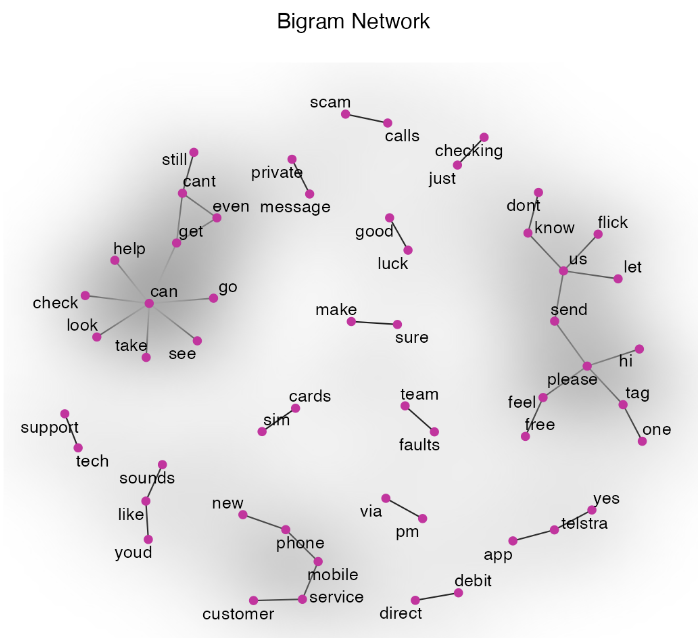

Figure 16: Bigram network from Facebook dataset

Furthermore, a specific examination of the bigram network 17 for the Amaysim brand underscores a predominant focus on the term 'Amaysim'.

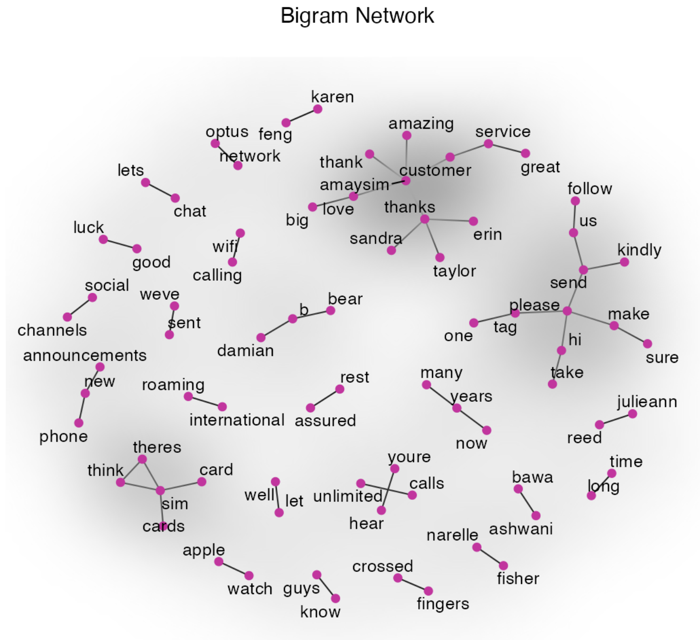

Figure 17: Bigram network of Amaysim

Within the Optus brand's bigram network 18, salient terms like *private* and *disappointing* emerge as central elements. This insight strongly suggests that customers frequently invoke the brand in their comments. Notably, the recurrent mention of *disappointing* in conjunction with *privacy* signals a noteworthy concern in customer feedback, warranting heightened attention from Optus in safeguarding customer information.

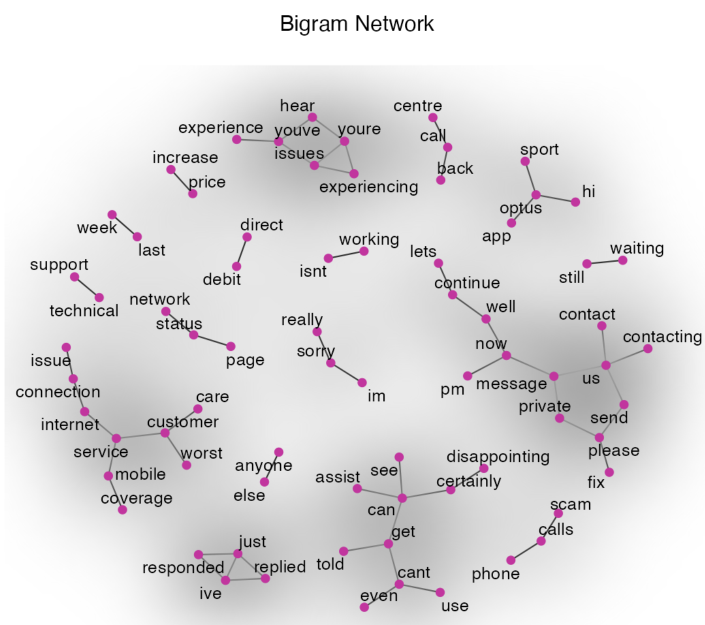

Figure 18: Bigram network of Optus

Additionally, the bigram network 19 for Vodafone AU pinpoints notable attention on terms like *check*, *issue*, and *disruption*, indicating a potential area of improvement in signal quality for their future products. This bigram network is a valuable suggestion for Vodafone AU to enhance signal performance in forthcoming offerings, further bolstering customer satisfaction.

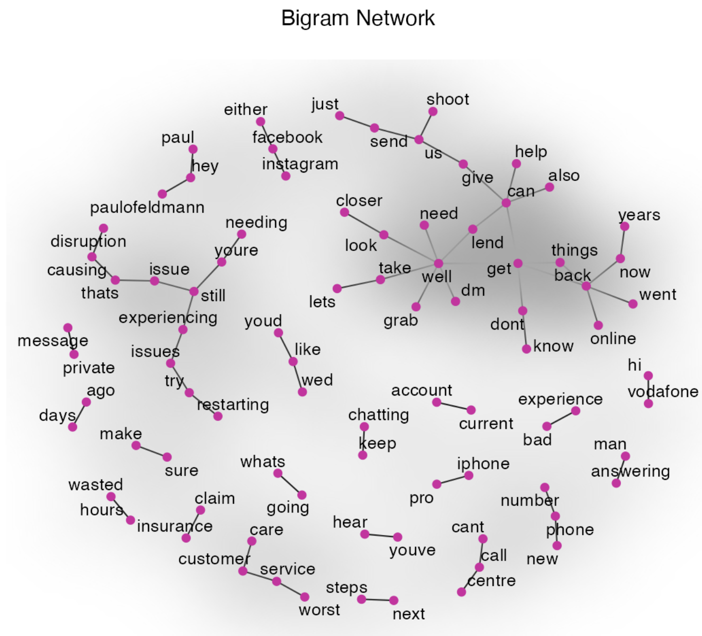

Figure 19: Bigram network of Vodafone AU

5.2.2 Word cloud

The figure 20 is another type of word cloud map that offers a unique perspective, showcasing words that recur with higher frequency in customer comments. This word cloud map offers a unique perspective, showcasing words that recur with higher frequency in customer comments. It is a potent visual representation of customer sentiment, providing brands with an intuitive means to discern and comprehend customer feedback about their products and services. Notably, negative sentiment is depicted in red, concentrated in the upper half of the word cloud, exemplified by terms like *issue*, *scam*, and *problem*. Conversely, positive sentiment is portrayed in green, predominantly occupying the lower half, encompassing terms such as *like*, *happy*, and *good*.

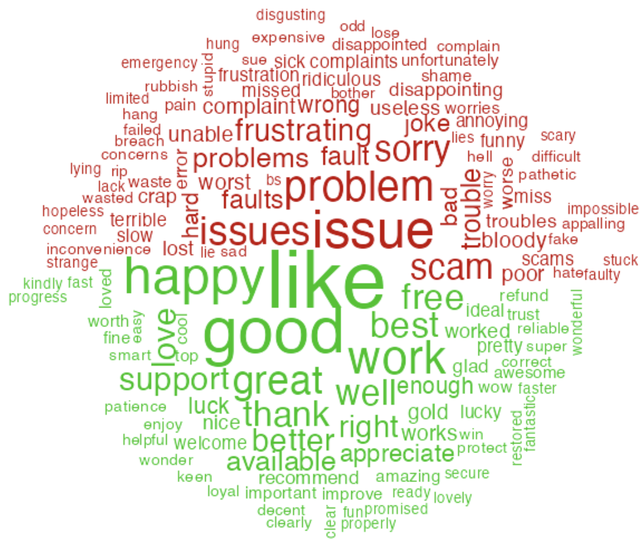

Figure 20: Word cloud of Facebook comments

For specific campaigns or products, delving deeper into the emotions they evoke can yield invaluable insights, allowing brands to cater to customer preferences and needs more effectively. This type of word cloud proves particularly invaluable in crises, providing brands with the ability to swiftly identify trends or issues that may be contributing to negative sentiment among customers. With this word cloud, brands can proactively implement improvements based on customer feedback, ultimately elevating their overall ratings and bolstering customer satisfaction.

5.2.3 Donut plot

This donut plot 21 is a visual tool for categorizing emotion words into distinct categories (such as positive, negative, angry, and so on). The visualization vividly showcases the distribution of these emotion categories as a percentage of total words in customer comments. This granular understanding of customer sentiment equips brands with a valuable metric to gauge customer satisfaction levels, thereby informing critical business decisions related to product enhancement, service optimization, customer experience refinement, and more. Upon analysis, it emerges that over 20% of the words in customer comments convey positive attitudes, underscoring a prevailing sense of contentment. Notably, words signifying anticipation, trust, and negative sentiment all constitute roughly 15% each.

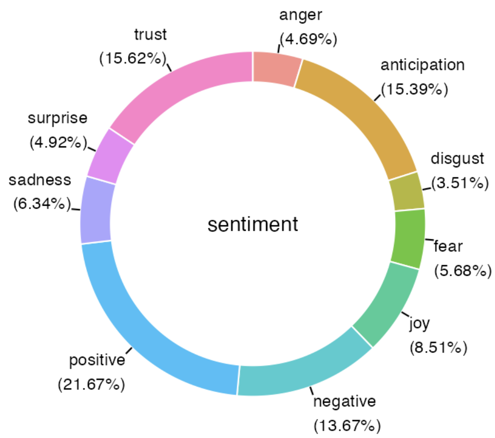

Figure 21: Donut plot of Facebook dataset

It is intriguing to note that Vodafone AU's customer comments 22 exhibit a slightly lower percentage of words expressing positive emotions compared to the other three brands, warranting further exploration.

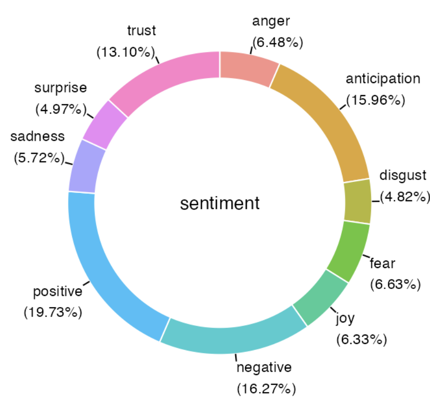

Figure 22: Donut plot of Vodafone AU

5.2.4 Tf-idf table

This table 23 presents vital information on term frequency (tf), inverse document frequency (idf), and their combination, tf-idf. Term frequency (tf) quantifies the frequency of a word's occurrence within a customer's comment. On the other hand, inverse document frequency (idf) assesses the word's prevalence across all comments, discerning whether it is a common or rare term in the corpus. The concept of tf-idf lies in its ability to spotlight pivotal words within the comment content, achieving this by diminishing the weight of commonplace words while elevating the significance of less frequently used terms across the entire comment collection or corpus. A higher tf-idf value signifies the heightened importance of a word in the context of the comments.

Show 8 entries Search:

	brand	word	n	tf	idf	tf_idf
1	Optus	sorry	98	0.09099	0.28768	0.02618
2	Vodafone AU	keen	25	0.06793	0.28768	0.01954
3	Amaysim	win	55	0.02539	0.69315	0.0176
4	Amaysim	kindly	31	0.01431	0.69315	0.00992
5	Telstra	gold	235	0.00644	1.38629	0.00893
6	Amaysim	amazing	47	0.0217	0.28768	0.00624
7	Telstra	faults	293	0.00803	0.69315	0.00556
8	Amaysim	welcome	14	0.00646	0.69315	0.00448

Showing 1 to 8 of 34 entries Previous 1 2 3 4 5 Next

Figure 23: Tf-idf table of Facebook dataset

In the current filter, the analysis highlights several keywords that carry significant weight across all customer comments. Notably, terms like *sorry*, *keen*, *win*, and *compatible* emerge as focal points. Intriguingly, the term *scam* also registers as a noteworthy factor as shown 24. To delve deeper into this observation, the `grepl` function was employed to conduct a targeted search within customer comments. This research result unveiled a cluster of comments where customers expressed dissatisfaction with the branded data roaming service, citing arbitrary charges as a point of contention. Customers perceive this practice as deceptive, an aspect that warrants immediate attention and resolution.

Show 8 entries Search:

	brand	word	n	tf	idf	tf_idf
1	Amaysim	compatible	6	0.0163	1.38629	0.0226
2	Optus	mad	2	0.01575	1.38629	0.02183
3	Optus	unavailable	2	0.01575	1.38629	0.02183
4	Optus	scam	3	0.02362	0.69315	0.01637
5	Vodafone AU	joke	5	0.02008	0.69315	0.01392
6	Amaysim	unlimited	6	0.0163	0.69315	0.0113
7	Amaysim	supported	3	0.00815	1.38629	0.0113
8	Amaysim	winner	3	0.00815	1.38629	0.0113

Figure 24: Tf-idf table of Instagram dataset

Additionally, there were instances of customers reporting information leakage concerns with calling cards, along with an influx of fraudulent text messages and calls. To proactively address this issue, the staff is taking commendable steps by providing a designated website for reporting scam SMS/calls, thereby empowering consumers to take action against potential scammers. This analysis underscores the imperative for the brand to redouble efforts in curbing all forms of scams and spam, thereby further elevating consumer satisfaction and trust.

6 Conclusion

6.1 Result highlighting

Upon an in-depth exploration and analysis of four distinct social media datasets, a comprehensive narrative has emerged, shedding light on the dynamic relationship between branded posts and customer comments within the realm of Australian telecom operators.

The time-series plot unequivocally illustrates the profound impact of innovative and emotionally resonant content, as well as collaborating with influencers to co-publish posts, in driving heightened customer engagement rates and amplifying brand visibility. Given that the level of customer interaction is higher for posts published on Instagram. Therefore, for brands seeking to maximize brand exposure, entirely allocating resources to create engaging post content on the Instagram social platform is a strategic task for brands.

Regarding sentiment and thematic analysis, it becomes evident that brands must maintain a vigilant watch on sentiment trends during content dissemination. Cultivating positive emotions within content not only bolsters customer engagement but also cultivates a more favorable brand perception. For telecom operators, striking a balance between customer-centric content and product/service updates emerges as a pivotal strategy. Focusing on customer appreciation, engagement, and loyalty fortifies satisfaction and retention while concurrently spotlighting product and service offerings ensures competitiveness within the market landscape.

In response to the data stories generated from customer comments, brands can make some appropriate adjustments in terms of services, products, and technological offerings. The common concerns of customers, such as data roaming charges, information security, and fraudulent calls/texts, should be of concern to brands. By actively mitigating these pain points, brands can effectively elevate customer satisfaction and foster brand loyalty. Notably, for Vodafone AU, which registers a comparatively lower percentage of positive sentiment in customer comments, considering improving the signal quality of the product is a step that should be addressed in increasing customer positivity and satisfaction.