Individual portfolio

Week 2:

During this week, our team focused on exploring the World Cup context. We split the datasets, distributed them to all the members (me and Sarayu are responsible for the Instagram dataset), and cleaned the dataset by removing unnecessary columns and duplicates. Timestamps were separated from datetime (time and date). We delayed our commit changes because not a lot of work has been finished.

In terms of research, We came up with some potential questions that can be used for the projects and took our time to think about the meeting of the mentors, as we could come up with any idea about the project.

We also worked on the date that we were going to meet up so that we could work on the project on that date, and it needed to be a fixed date for everyone. We added each other's social media accounts so it could be easier to communicate.

Some insights that I can see are that Instagram's data seemed less complicated than Twitter and Facebook. This week marks the beginning of our journey, and I'm eager to see our insights evolve in the coming weeks.

WEEK 3:

During this week, we explored more of the practical use of Commit to Google Colab and created our own repository. We learned how to effectively use Commit to Google Colab and set up our own repository. This new knowledge will be essential in applying it to our project, reflecting our collaborative efforts. While my personal progress was not that great, my team made a lot of progress in various aspects of our project.

In terms of research, we made the strategic decision to divide our timeline into three segments: pre-World Cup, during the World Cup, and post-World Cup. My specific responsibility this week was performing a deep clean on the data, primarily within the 'content' column. The process of dealing with emojis presented some challenges and required a significant amount of time and effort because there are many languages.

One of the hard things this week for me was that we did not have any meetings, so I could not finish my task in time, but I am very lucky that my group has successfully completed many tasks and graphs.

An interesting insight came up from our exploration of the Instagram dataset, and potentially other datasets as well - the presence of various languages, including Chinese, Arabic, and more. This observation underscores the need to account for linguistic diversity as a potential factor influencing our analysis.
Despite facing challenges and limited individual progress this week, our collective efforts have propelled us forward in our project.

t[8]:

	content	from	location	date	device	hits	likes	comments
0	The FIFA World Cup Qatar 2022™ kicks off today...	vivo_global	NaN	2022-11-20 22:40:23	vivo	NaN	430.0	8
1	FIFA World Cup Qatar 2022™ is only 3 days aw...	hisense_international	NaN	2022-11-17 12:06:09	Hisense	NaN	161.0	66
2	The FIFA World Cup Qatar 2022™ kicks off in 3 ...	vivo_global	NaN	2022-11-17 23:00:53	vivo	NaN	662.0	11
3	The FIFA World Cup Qatar 2022™ is one week awa...	hisensesports	NaN	2022-11-12 12:03:57	Hisense Sports	NaN	74.0	2
4	We are 10 days away from the FIFA World Cup Qa...	hisensesports	NaN	2022-11-10 23:00:50	Hisense Sports	NaN	46.0	3
...	...	...	...	...	...	...	...	...
7629	The tensions are high and Skyworth is here t...	skyworth_my	NaN	2022-12-13 16:00:04	Skyworth Malaysia	NaN	5.0	0
7630	Morocco's World Cup magic potion, in part, is ...	freemalaysiatoday	NaN	2022-12-13 11:34:00	Free Malaysia Today (FMT News)	0.0	1413.0	6
7631	Watch the FIFA World Cup Qatar 2022 ™ like ne...	officialhlive	NaN	2022-12-05 13:28:22	Official Hlive	NaN	47.0	5
7632	Many eyes will be on Iran's side in the World ...	carnegiemiddleeast	NaN	2022-11-23 01:00:04	Carnegie Middle East Center	NaN	5.0	0
7633	The most expensive FIFA World Cup ever kicked ...	dwdocumentary	NaN	2022-12-02 23:54:54	DW Documentary	NaN	1357.0	56

:

	content	from	location	date	device	hits	likes	comments
0	The FIFA World Cup Qatar 2022™ kicks off today...	vivo_global	NaN	2022-11-20 22:40:23	vivo	NaN	430.0	8
1	FIFA World Cup Qatar 2022™ is only 3 days awa...	hisense_international	NaN	2022-11-17 12:06:09	Hisense	NaN	161.0	66
2	The FIFA World Cup Qatar 2022™ kicks off in 3 ...	vivo_global	NaN	2022-11-17 23:00:53	vivo	NaN	662.0	11
3	The FIFA World Cup Qatar 2022™ is one week awa...	hisensesports	NaN	2022-11-12 12:03:57	Hisense Sports	NaN	74.0	2
4	We are 10 days away from the FIFA World Cup Qa...	hisensesports	NaN	2022-11-10 23:00:50	Hisense Sports	NaN	46.0	3
...	...	...	...	...	...	...	...	...
7629	The tensions are high and Skyworth is here to...	skyworth_my	NaN	2022-12-13 16:00:04	Skyworth Malaysia	NaN	5.0	0
7630	Morocco's World Cup magic potion, in part, is ...	freemalaysiatoday	NaN	2022-12-13 11:34:00	Free Malaysia Today (FMT News)	0.0	1413.0	6
7631	Watch the FIFA World Cup Qatar 2022 ™ like nev...	officialhlive	NaN	2022-12-05 13:28:22	Official Hlive	NaN	47.0	5
7632	Many eyes will be on Iran's side in the World ...	carnegiemiddleeast	NaN	2022-11-23 01:00:04	Carnegie Middle East Center	NaN	5.0	0
7633	The most expensive FIFA World Cup ever kicked ...	dwdocumentary	NaN	2022-12-02 23:54:54	DW Documentary	NaN	1357.0	56

Before and After

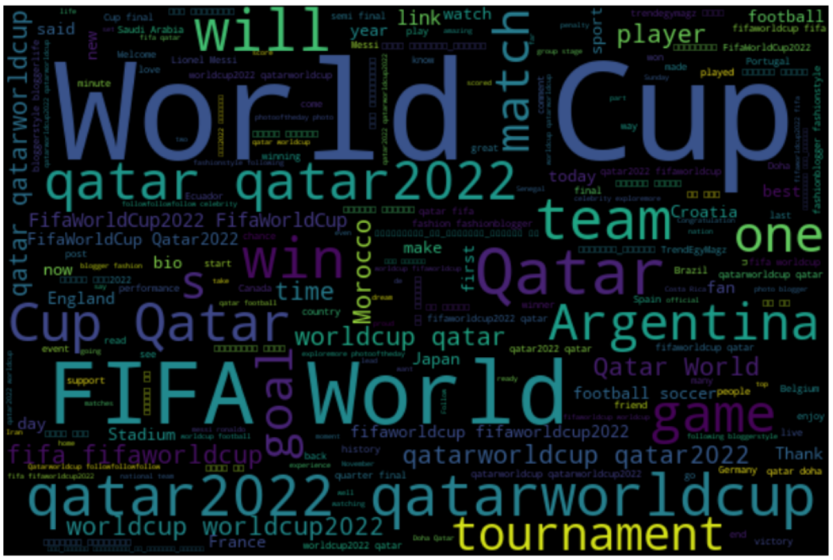
Week 4:

This week was marked by several noteworthy achievements and productive discussions within our team. I took on the responsibility of creating a word cloud for our team. While others focused on cleaning their datasets, including publications and hashtags, my task involved generating a word cloud to identify the most frequently mentioned words over time.

In a Wednesday meeting, we brainstormed the idea of categorizing our data into players or countries that attended the World Cup. We identified a dataset containing all the player names, which will aid in our analysis. Our team meeting this week proved exceptionally productive. We discussed updates to our code and shared ideas on how to effectively present our findings.

While working on my task of creating a word cloud, I encountered some challenges. The word cloud generation required the download of multiple packages, leading to a few bugs that needed to be addressed. It took some time, but I eventually resolved these issues.

Overall, we are heading in the right direction, and our progress is encouraging. We should continue with our current efforts and build upon the valuable insights we're uncovering.



WEEK 5:

This week, progress was minimal due to the heavy assignment workload that many of us were struggling with. We decided to take a break, allowing everyone to manage their assignments and deadlines effectively. However, even during this break, we maintained engagement by periodically reviewing our project's status and contemplating our next steps.

Despite the break, one of our team members, Sarayu, managed to make significant contributions. She successfully created two insightful graphs. The first graph highlighted users with the most likes across the entire time span, offering an insight into who produced the most engaging content. The second graph focused on users with the most comments, presenting an alternate perspective on user engagement. Sarayu's work added valuable value to our project.

We had a constructive discussion with Jefery and Zach, during which we recognized the need to refine and clarify our research questions. It became evident that the current formulation of our questions might lead to vague outcomes and a lack of clear research directions. This realization prompted us to rethink our project's overarching objectives and come up with more precise and focused questions.

As we approach the final stages of our project, the need for clear and well-defined research questions has become increasingly apparent.

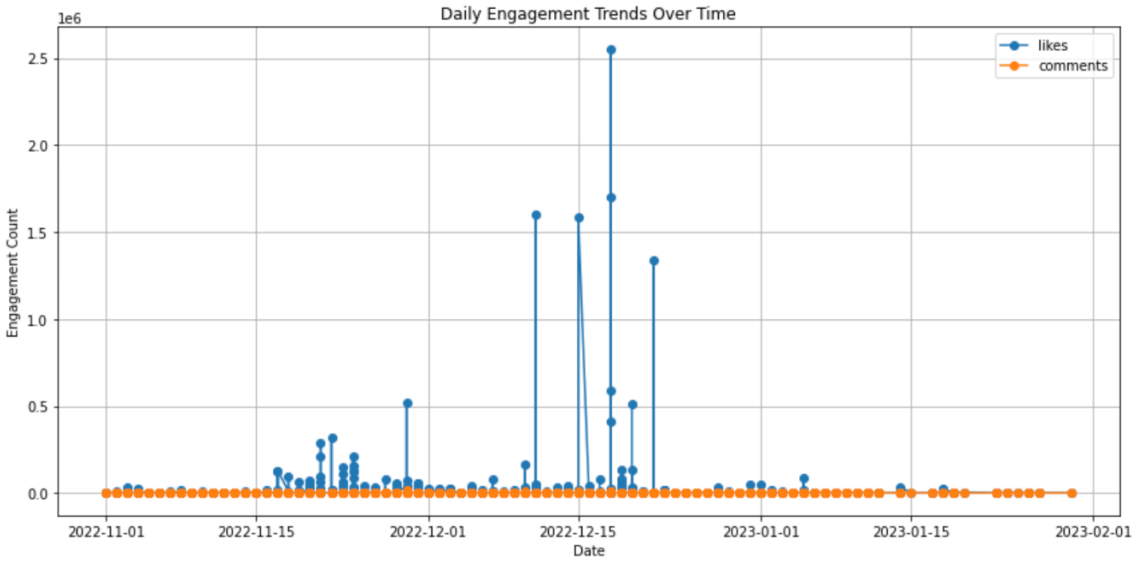
Week 6:

This week, at the meeting, we decided to gather and come up with the questions that were needed and assign the jobs to the members to see whether the question is viable or not. My question is, ‘How do the trending topics evolve throughout the tournament’s progression? Are there specific periods where certain controversies are most highlighted?”.

While researching this question, I employed the Pandas library for data manipulation and Matplotlib for data visualization. These tools allowed me to explore and analyze the daily engagement trends over time. Specifically, I examined metrics such as likes and comments to gain insights into how user engagement evolved throughout the tournament. I could discover which days had the highest engagement, whether there were spikes in engagement during specific events or matches, and whether there was a decline or increase in engagement as the tournament progressed.

This week, we also started work on the project report by starting some of the sections, such as the background, We also began to comment out code, for better understanding among team members regarding each other's contributions.

It seems like we're moving in the right direction, but it's crucial for us to enhance the specificity of our research questions. This will enable us to maintain a clear and focused approach to our project.



Mid-semester Reflection (Week 7) :

I've learned several valuable lessons while doing this project. Having a clear main question or goal is crucial. In the past, I worked mostly with numbers, so dealing with text data in this project was a new experience. Combining Data for Better Insights is a very important skill. We had three separate sets of data to work with, and bringing them together effectively was a big deal. It helps us look at all the data as a whole and understand it better. It's like putting all the puzzle pieces together to see the complete picture. At the same time, it's beneficial that each of us has conducted independent analyses on our respective datasets, and somehow we have different ways of observing the dataset and coming up with many different ideas.

A significant challenge we faced was creating clear research questions. Our project's nature is quite wide and open, so it makes it very hard for us to focus on what we are doing. When it comes to coding, we haven't encountered significant struggles since our project doesn't require us to implement predictive models. However, the limitations in our dataset give us so many challenges. As we move forward, my hope is that by the end of the year, we can discover and apply more advanced techniques to extract better insights from our dataset.

My primary involvement centered around the Instagram dataset. Collaborating closely with Sarayu, we jointly worked on the code that is currently in the repository. In particular, my focus was on tasks such as generating word clouds, removing emojis, and conducting sentiment analysis. Additionally, to ensure ongoing communication within the team, I took the initiative to host regular Tuesday Zoom meetings.

To do better, I should communicate more with my teammates. Overall, I think we're on the right track to present our findings in an interesting way.

WEEK 8

I ventured into importing a new dataset to identify the frequency of mentions for specific players and countries. The objective was to discern which players or countries were most talked about. We encountered multiple dataset anomalies. For instance:

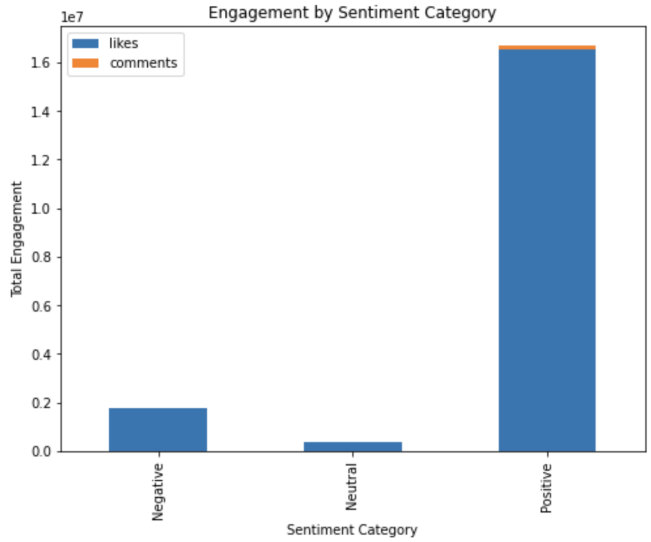
- There is variability in naming conventions for countries across posts.

- Player names are too general. Recognized the need to be more specific as the surnames needed for accurate identification.

One of our teammates, Macka, introduced an innovative idea regarding sorting the player names, which added immense value to our approach.

During this week, my primary focus in the coding domain was sentiment analysis, and I was fortunate to have Jeffrey's expertise to guide me through the complexities. While I made significant strides by deciphering the sentiments present in our Instagram dataset's content, I sought additional depth by turning to AI resources, specifically ChatGPT. Their choice of employing the 'vader' methodology for sentiment analysis intrigued me. Jeffrey posed an insightful question, prompting me to reflect on the reasoning behind the model's decision to split the threshold for its analysis. This contemplation not only deepens my understanding but also encourages me to further refine our sentiment analysis techniques.

Sentiment_Category	likes	comments
Negative	1742634.0	41411
Neutral	378326.0	6639
Positive	16531854.0	143601



While leveraging ChatGPT, I was reminded of an essential lesson: it's crucial to not just accept what the AI provides, but to truly understand and interpret its output. This ensures that the insights gained are both meaningful and relevant to our specific research context.

For the coming week, we must ramp up our research efforts, especially considering we're already into week 8. While our progress on the report and the PowerPoint presentation seems satisfactory, it's imperative that we maintain momentum in other areas. If research doesn't advance as anticipated, it's crucial to ensure all other aspects remain on course.

WEEK 9:

This week provided illuminating insights into the distinct nature of different social media platforms. We discovered that word clouds from each platform mirror their inherent characteristics. For instance, Instagram thrives on hashtags, a cultural norm to amplify visibility and engagement, while Twitter's word cloud is dominated by news-related terms, reflecting its status as a real-time information hub.

This observation led us to an intriguing research question: "Which hashtags garner the most traction in posts? Does the inclusion of certain hashtags correlate with increased engagement?" In my investigation, I uncovered that hashtags related to the LGBT community were banned in some host countries. To gauge the impact of hashtags on user engagement, I embarked on an analytical journey. I started by computing the average number of likes a user received. Then, I differentiated these averages based on posts with hashtags versus those without. To ensure robustness in our findings, we only considered users with a minimum of three entries for both categories. This stratification allowed us to compare the mean likes for each scenario, offering a clearer picture of the role hashtags play in user engagement.

Twitter

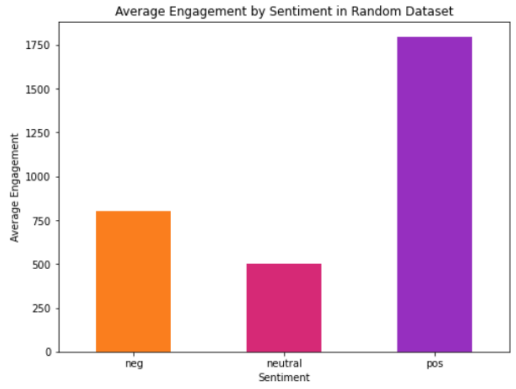
Instagram

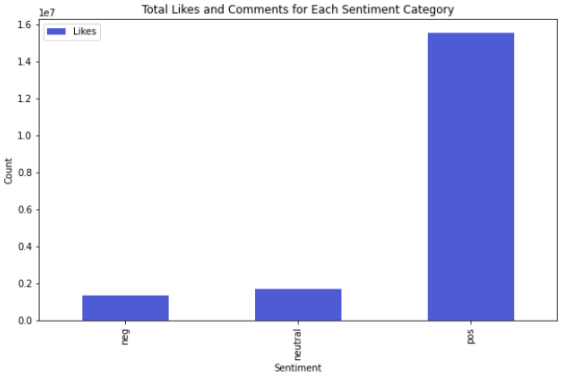
Our focus narrowed to exclusively analyzing the Instagram and Twitter datasets, primarily because hashtags are more prevalent on these platforms. Utilizing word clouds, exploratory data analysis (EDA), and external resources, we unearthed significant insights regarding the influence of hashtags on engagement. An interesting dichotomy emerged: on Twitter, posts with hashtags consistently garnered higher engagement compared to those without. However, Instagram showed a contrasting trend, with posts containing hashtags receiving, on average, fewer likes than those without.

We're quite pleased with our progress this week, especially having a good research question that holds promise. However, given the upcoming deadline, we must capitalize on this momentum and delve deeper for more insightful findings.

WEEK 10:

This week, on Zack's suggestion, we delved into predictive modeling, although determining what to predict posed a challenge. Simon's recommendation steered us towards the correlation between sentiment and engagement, leading to our research question: "Does content with positive sentiment garner more engagement than negative?" The relationship between content sentiment and engagement is intricate and may differ by context and platform. Using the textblob package for sentiment analysis, we assigned sentiment scores between -1 and 1 to posts. Based on this, we trained two models: a naive Bayes classifier and an SVM model. While textblob gauges sentiment, employing both models ensures comprehensive sentiment analysis, especially for intricate emotions.
Me and other team member worked on the Instagram dataset and this is the insight that we got: The Naive Bayes Model yielded an accuracy of 0.75, whereas the SVM Model outperformed with an accuracy of 0.8. When tested on a dataset of 300 random sentiment entries, the data suggested that posts with positive sentiments generally garnered the most engagement. This observation was solidified by another graph, which showed that positive sentiment posts in the complete dataset consistently received the highest combined likes and comments.



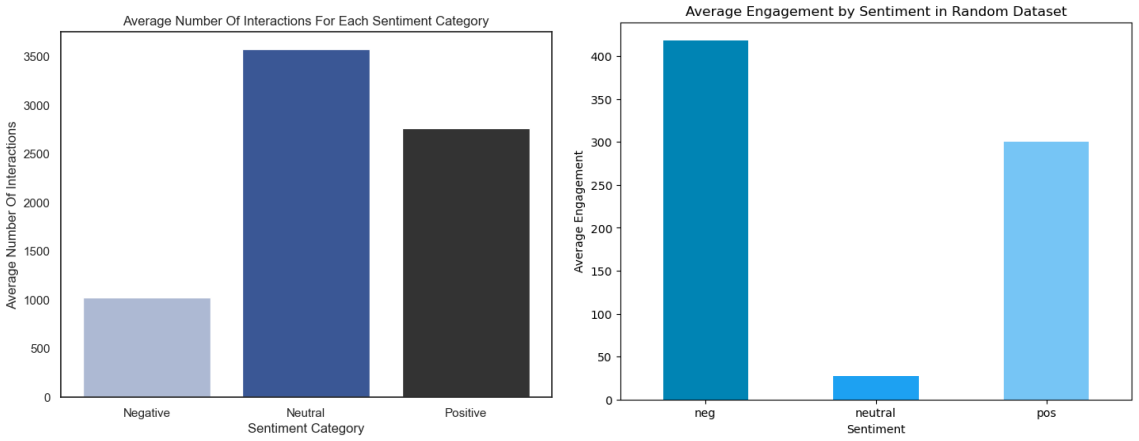


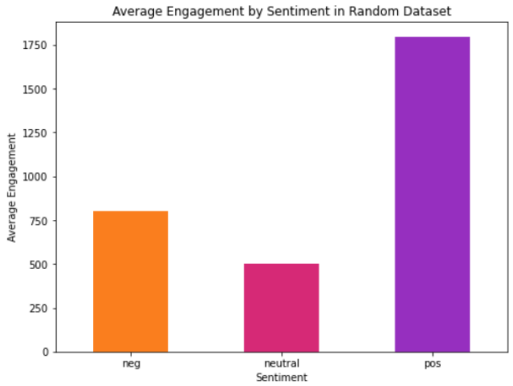
We're quite pleased with our progress this week, especially since we just came up with another good research question that is very useful to our project. The other team members that worked on the other dataset also came up with very positive outcomes.

WEEK 11:

This week, we wrapped things up for our presentation.A noteworthy contribution from the team was the introduction of an insightful question about the evolution of trending topics during the tournament. Instead of diving deeper into coding, our primary focus transitioned to refining our PowerPoint slides and organizing the content of our report. Each team member was assigned a specific section to present, ensuring a well-distributed and coherent presentation. A challenge we faced was the need to streamline our content due to time constraints, which required us to prioritize and highlight the most essential information. Beyond this, we dedicated time to anticipate and prepare for possible questions, ensuring we'd be ready to address any question.

This week, we took a break from coding. Instead, I assisted my teammates by ensuring their graphs aligned with our theme colours, enhancing the overall aesthetic and cohesion of our presentation.





We're quite pleased with our progress this week and feel well-prepared for our upcoming presentation. The primary focus now is refining our PowerPoint to ensure it's polished and engaging. While our report still requires significant work, given that we need to integrate

materials from three different sources due to our individualized approach, we generally believe we're on the right track.

WEEK 12:

In this final week, our presentation went reasonably well, and I'm genuinely satisfied with the outcome. The feedback provided was enlightening and will serve as valuable lessons moving forward. A challenge I grappled with was effective team communication, but despite the hurdles, we collaborated successfully to reach our goal. Throughout the project, I often found myself uplifting the team's spirit during challenging phases. While we occasionally faced moments of doubt, especially regarding our final model,i think my optimism was helping. Moving forward, I recognize the need to enhance my communication skills and to voice my concerns more assertively. Overall, this has been one of the most fun and interesting projects I've undertaken.