Reflection Entry - Week 2

This week, our team focused on refining our dataset for analysis. We split the datasets and conducted basic data cleaning by removing unnecessary columns and duplicates, and separated timestamps from the datetime column. We made progress but decided not to commit to the repository yet, as more work remains.

In terms of research, we explored the context of the World Cup, and considered categorising data into pre, during, and post-World Cup periods. We also wondered what questions event organisers might have when working with similar datasets, refining our project objectives.

We met earlier during the week to ensure communication and collaboration remained strong. In the meeting, we discussed progress and shared the GitHub Repository for code and a Google Drive folder for research information from our mentor.

Some insights I found during my analysis is that the Instagram data appeared less extensive and responsive than Facebook. This difference may relate to content types on each platform, which prompted questions about platform choice and content's impact on user engagement.

Reflection Entry - Week 3

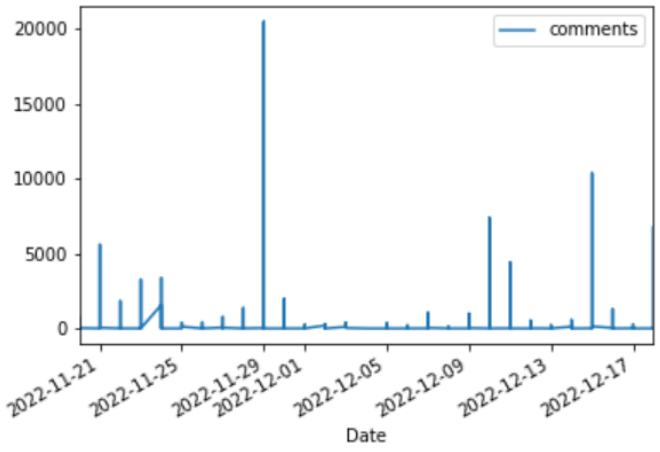
During Week 3, our team made significant progress in preparing the social media dataset for analysis. I began by examining basic statistics related to likes and comments, which provided us with a preliminary understanding of user engagement. To ensure data accuracy and workability, I converted the relevant columns from strings into datetime and floating-point formats. This step was crucial for effective data manipulation and visualisation. As part of our exploratory data analysis (EDA) efforts, we created a basic graph illustrating the sum of comments per day over the entire duration of the dataset.

This week, we conducted research into compiling a list of players who participated in the World Cup. This information could prove valuable for understanding user engagement and posts related to these players during the tournament.

We did not have any team meetings this week. However, we made a decision to combine the various reactions in the Facebook dataset into one consolidated column labelled "interactions." This decision was primarily driven by the fact that other datasets we were working with contained only "likes," making it easier to compare user engagement across platforms. Additionally, we shared some code related to natural language processing (NLP) analysis and continued our efforts in EDA, making progress toward our project's goals.

One interesting observation was made regarding the Instagram dataset, which contained a substantial number of hashtags. This finding has sparked our interest in delving deeper into hashtag analysis for content

Figures Generated:



Week 4 Reflection Entry

One of the major accomplishments this week was successfully downloading the NLTK package for text analysis. Additionally, we divided the dataset into three distinct timeframes: pre-match, during-match, and post-match dates. This categorization will enable us to analyse how conversations and sentiments evolved throughout the World Cup. Another completed task this week was the removal of emojis from the content. This step not only makes the text more manageable but also ensures that we are working with a clean dataset.

Furthermore, we found an external dataset containing a list of players who participated in the World Cup. This opens up the potential of using this data to gain insights into player interactions, engagement, and mentions.

Our meeting this week was particularly productive as we discussed updates to our code and shared ideas on how to present our findings effectively. We reached a consensus on creating a timeline visualisation to track trending topics before, during, and after the World Cup across our datasets. This will provide a comprehensive view of how discussions evolved over time.

An Interesting insight I found this week is that many of the other datasets we considered did not contain as many emojis as our current dataset. This might suggest that the use of emojis was more prevalent in conversations related to the World Cup.

Our code is steadily progressing. The removal of emojis not only enhances the readability of the content but also ensures that we are working primarily with English-language posts, which aligns with our research objectives. However, we still have some work ahead, including the removal of punctuation, addressing copyright issues, and dealing with hashtags.

Figures Generated:

Before

	content
7400	entered #Hisense amazonfiretv competition year...
7024	support successful delivery FIFA World Cup Qat...
7618	���World Cup trophies official football ⚽ ⚽ ⚽...
7141	卡達�� 世足賽2022要到了！你的足球魂已經開啟了嗎？ ⚽ Qatar World Cup2...
7402	Qatar's moment arrived. tiny emirate using ...

After

	content
7400	entered #Hisense amazonfiretv competition year...
7024	support successful delivery FIFA World Cup Qat...
7618	World Cup trophies official football for upcom...
7141	2022 Qatar World Cup2022 coming soon, ready ...
7402	Qatar's moment arrived. tiny emirate using ...

Week 5 Reflection Entry

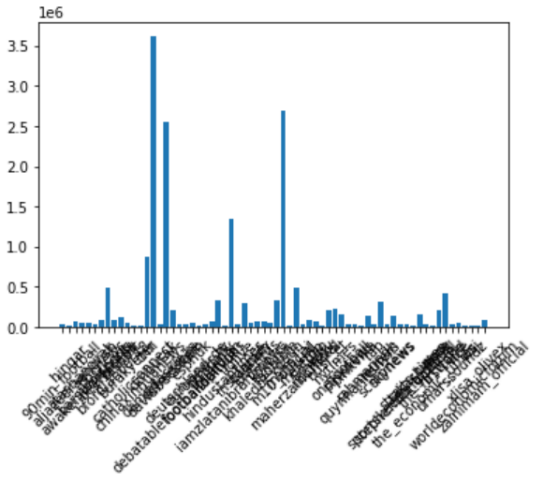
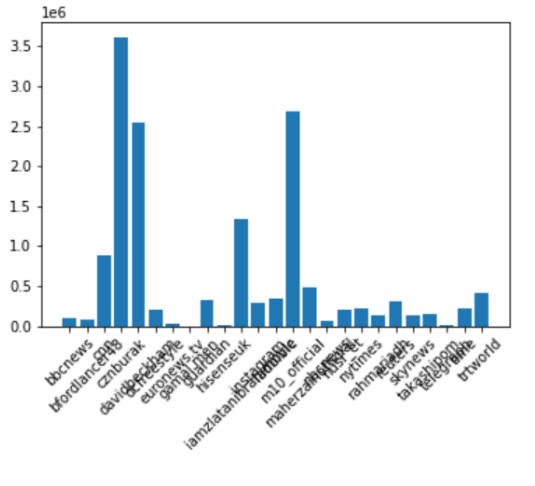
One of the major accomplishments this week was successfully removing stopwords from the dataset. Additionally, I created two important graphs that provide valuable insights. The first graph highlighted users with the most likes over the entire time period, shedding light on who had the most engaging content. The second graph focused on users with the most comments, offering a different perspective on user engagement.

Collaboration continued to be a key component of our work this week. I shared some code snippets related to the removal of punctuation and specific words or phrases that we want to exclude from our content.. Additionally, we created a graph illustrating the most mentioned players in the dataset, providing us with valuable insights into the key influencers in our data.

One notable insight from our data exploration this week was the prevalence of news article-type content within the Twitter and Instagram dataset. This discovery opens up the possibility of analysing whether certain types of content perform better on one platform compared to another.

The successful removal of stopwords is a significant milestone in our data preprocessing journey. This step will enable us to generate word clouds, which will help us identify the most frequently occurring words for further analysis. However, there is still work to be done, particularly in terms of further removal of punctuation and other words that may need to be excluded.

Figures Generated: (both need to be further edited for readability)

Users with most number of likes	Users with most number of comments
	

Week 6 Reflection Entry

One of the main accomplishments this week was generating a word cloud from the content section of our dataset. This allowed us to visualise the most frequently mentioned words and phrases, giving us valuable insights into the dataset's content. We also formulated a plan for additional EDA. This plan includes analysing the most used words by users to gain a better understanding of their content. Furthermore, we are considering creating a correlation matrix between likes and comments to explore potential relationships between these variables.

Generating the word cloud required us to delve into various resources to understand the process thoroughly. This involved installing and updating several new packages to facilitate the task.

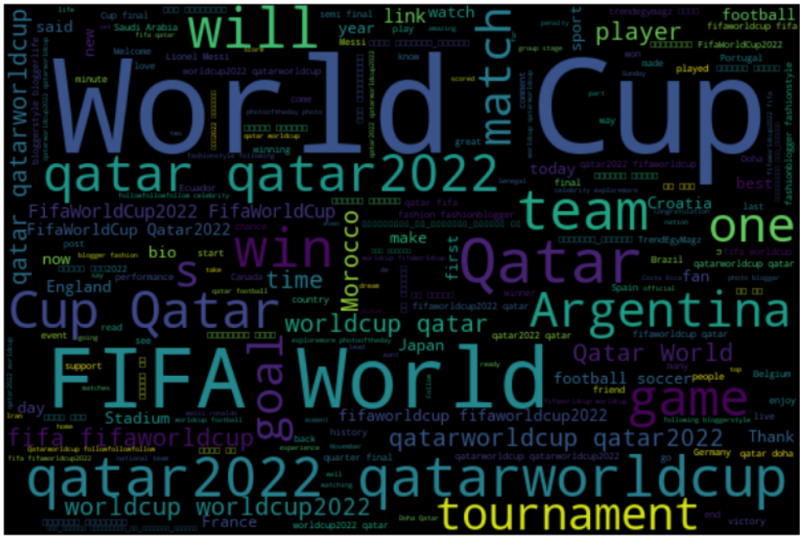
Our team continued to work collaboratively during this week. We brainstormed and developed three to four research questions based on our dataset, aiming to combine them into a coherent research project. We also uploaded our current files into the repository, ensuring we have a reliable backup and a reference point for code from other datasets.

In addition to these tasks, we started work on the project report by tackling some of the easier sections, such as the introduction and data quality assessment. We also began to comment out code, for better understanding among team members regarding each other's contributions.

Looking at the word cloud showed some interesting insights. Notably, we observed that the most frequently mentioned words included "FIFA," "World Cup," and "Qatar." While these terms are expected in a dataset related to the FIFA World Cup, we need to filter them out to reveal other meaningful words and phrases. Additionally, we are considering delving into the hashtags used within the dataset to uncover further insights and patterns.

Our code is currently in a relatively clean state, but there is still work to be done. A priority for the upcoming week is to separate the hashtags from the content itself.

Figures Generated:



Mid Sem Reflection:
Working on this project has taught me valuable lessons about the importance of data quality and cleaning before diving into dataset analysis. My previous experience was mainly with numerical data, which contrasts significantly with text data as it requires extensive cleaning. We are dealing with three distinct datasets, each necessitating thorough cleaning and filtering, underlining the significance of effective communication within the team. As each of us has conducted individual analysis on our respective datasets, it's good to compare our thought processes and understand why certain phrases were removed or specific graphs were chosen for representation. This emphasises the need for regular communication.

One of our primary challenges was defining a solid set of research questions. Our project is quite open-ended, providing flexibility but making it challenging to reach a consensus on what to focus on. Ultimately, we settled on a timeline approach to analyse World Cup trends as they occurred. While we've filtered down to a tentative set of questions, they may require further refinement or focus. We haven't encountered many coding issues yet, thanks to online resources and generative AI to address any errors we encounter. However, I recently faced a challenge with generating a word cloud, which required installing and importing unfamiliar packages. Thankfully, a team member assisted me in resolving this, and I adapted their code to the filtered dataset.

My main contribution was on the Instagram dataset, which had a limited number of variables. Tan and I collaborated on the current code committed to the repository, with my focus mainly on data filtering and EDA, while he worked on emoji removal, word cloud generation, and sentiment analysis. I also initiated a messenger chat at the project's outset to maintain constant communication and set up regular Tuesday meetings for progress updates.

To improve my performance, I could engage more with my teammates outside of scheduled meetings and class, fostering better overall communication. Additionally, integrating some of the techniques discussed in class into our analysis might help uncover insights we may have otherwise overlooked. Overall, I believe we are on track to deliver a compelling presentation with intriguing findings for our stakeholders.

Week 7 Reflection Entry

This week has been productive in our project's progress. We successfully separated the hashtags from the content and generated a word cloud of the most common ones. By filtering out the hashtags, it allowed us to further analyse if they have any relevance and importance to the content.

Furthermore, we utilised the textblob package to calculate the polarity and subjectivity of the content provided us with a deeper understanding of the sentiment associated with the text. This analysis could potentially offer valuable context for interpreting user engagement and content perception.

We also experimented with a simple linear regression model to predict comments based on various variables. Although the results were lacking, it helped us understand if there was any relationship between the engagement and other variables.

Our group meeting this week provided an opportunity to discuss the progress made in the code and the preliminary sections of the project report. Although not much headway was made in the report, we allocated tasks for those not actively involved in coding. It also became apparent this week that we need to refine our strategy for addressing the research questions. Although we have a substantial number of questions, we currently lack sufficient analysis to answer them comprehensively, especially for the mock presentation next week.

Moving forward, our focus will be on refining the research questions to align them with the available data and analysis. Hopefully the feedback from the mock will provide us with an idea of how we want to go about this process.

Figures Generated:

Week 8 Reflection Entry

This week was a bit slower for me as most of my focus was directed towards completing an essay. Despite this, I did organise a group meeting where we collectively deliberated on the organisation of our research for the mock presentations scheduled on Friday. We reached a consensus to showcase the analyses and graphs we had generated up until this point.

Following the mock presentation, we received valuable feedback that highlighted the need for more concise research questions. It was suggested that we limit our focus to 3 or 4 key research questions. This advice resonated with our current analysis progress, and we wanted to try and incorporate a predictive model, as suggested. It was evident that this would add depth to our research.

In the upcoming week, our primary goal is to address the feedback we received. We plan to refine our research questions, ensuring they are clear and concise. Additionally, we aim to intensify efforts towards building any form of predictive model to enhance the depth of our analysis.

Week 9 Reflection Entry

For our progress this week, our team successfully narrowed down the research questions to a concise set of four. We believe these questions effectively encompass elements across all three datasets, and can reach meaningful conclusions from them.

Unfortunately, we encountered some hurdles in advancing the linear regression model. As a result, we opted to develop a correlation matrix and examine feature importance. The results showed us that there was a strong relationship between likes and comments, which we thought we could further explore, along with the content.

No meeting this week as I had prior commitments, however, the team unanimously agreed to utilise the break to work on a sentiment-related predictive model. We also received an update regarding the progress of the report, with the inclusion of our finalised research questions and an overview of the data structure.

As we move forward, it is evident that our progress has been steady, but a little disorganised. However, there is still work to be done, particularly in terms of building a model and answering our research questions. With a clear plan and a unified approach, I am confident that we will be able to overcome the challenges and accomplish our objectives efficiently.

Week 10 Reflection Entry

This week, we focused on refining our data for the number of likes and comments, opting to regularise it. This step was crucial to ensure that the data remained consistent and manageable. Regularising the data also helped in minimising the impact of outliers and scaling the values for a more comprehensive analysis.

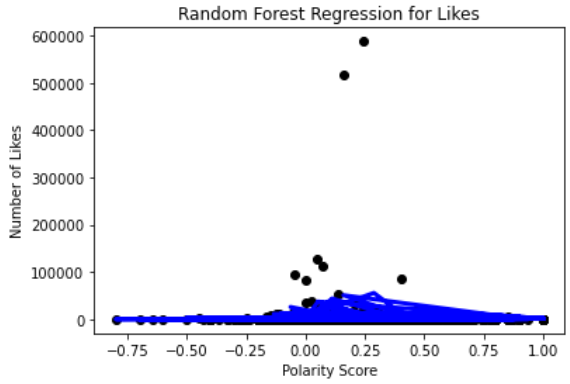
As we ended the break, we utilised the time to experiment with a random forest model in an attempt to predict likes and comments solely using the polarity score. However, this approach did not yield the expected results. The low Mean Squared Error (MSE) and R2 scores indicated a lack of substantial correlation between the variables. Consequently, we collectively decided to abandon this model and shift our focus towards exploring alternative methodologies to assess the relationship between content and sentiment more effectively.

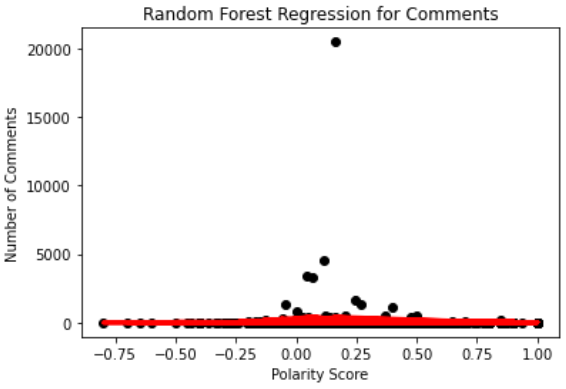
Our group meeting during the week served as a checkpoint to consolidate the progress made by each team member. It became apparent that others had also encountered challenges in progressing with their respective models. As a result, we had to reassess our approach to the third research question.

I also took the initiative to create a new notebook for my code to streamline what was relevant. This helped me keep my conclusions to the research questions in one place and not have to scroll through a long notebook for a certain graph or important information.

Moving forward, we plan to explore different methods to find a relationship between content and sentiment. We know it is a little last minute, but as our main focus is on modelling, it shouldn't be too hard to try and experiment with something.

Figures Generated (not used):





Week 11 Reflection Entry

This week was primarily dedicated to working on our research questions (RQ1, RQ2, and RQ4) and ensuring that our analysis provided meaningful insights into each platform's behaviour. Given that we were well aware of the impending busy week 12, with both the presentation and other assignments on the horizon, we recognized the importance of making progress this week.

During our team meeting, we discussed potential models for RQ3 and how to effectively implement them. We realised that addressing RQ3 would be an ongoing effort, and it's a critical component in completing our project. This is something we need to continue working on in the coming week to ensure a comprehensive answer to that question.

I also requested my team members to update the report with our analysis for the questions we've already tackled. This way, we can focus on filling in the modelling segment later on, minimising any last-minute rush during the project's final stages.

As we move forward, our priorities include dedicating most of our efforts to address RQ3 effectively. As most of our conclusions for the other questions are done, we need to re-evaluate how we want our modelling to answer this question.

Week 12 Reflection Entry

This week was very hectic as we focused on the presentation, which turned out to be a last-minute scramble. To ensure efficiency, I assigned specific sections to each group member for discussion in the PowerPoint and prompted the others to begin updating their corresponding segments in the report.

On the technical side, we managed to successfully implement SVM and Naive Bayes classifiers as our prediction models, achieving decent accuracy scores. I then shared the models with the team to utilise in their datasets. However, after the presentation, I realised that the report was not properly composed and lacked the necessary formatting. This discovery led me to send out a proper template for the report and fill in the required sections, ensuring that the appropriate format was adhered to.

Juggling the presentation, report, and technical aspects proved to be challenging, and it was a frustrating week overall. I found myself constantly following up with the team to ensure that we had ample time for rehearsal and to refine our presentation. It was crucial to guarantee that we could meet the submission deadline promptly.

Despite the challenges, this whole experience reinforced the significance of effective communication and proper planning to prevent last-minute crises.

Semester End Reflection

During this project, I gained a lot of insight into the intricacies of working with text data, which starkly contrasted with my prior experience with numerical data. The extensive process of cleaning and formatting the text data, despite appearing straightforward initially, demanded more time than we had anticipated. I came to realise that tasks like removing copyright information and dealing with punctuation required careful consideration and sometimes didn't yield the expected results.

Our primary challenge centred around formulating our research questions and integrating modelling into our project scope. Reflecting on our process, initiating the modelling phase earlier could have significantly enhanced our overall project execution. Although we managed to effectively narrow down our research questions, more collaborative efforts would have likely improved the quality of our modelling.

My main role in this project was focused on the Instagram dataset analysis for each research question. Additionally, I took on the responsibility of managing the team, ensuring that we remained on track and met our deadlines. This experience highlighted the importance of effective communication and task management in collaborative endeavours.

Working with this team has reinforced the significance of clear communication. I now understand that it's crucial not only to manage my own expectations and instructions but also to encourage others to share their ideas and seek their input. I also should've been more diligent in tracking our deliverables to prevent last-minute rushes.

Overall, this project has not only deepened my understanding of managing complex text data but has also highlighted the critical role of proactive communication and meticulousness, and I look forward to applying these lessons in future collaborations.