**Social Media and Network Analysis****(COSC 2671)**

**Assignment 2**

**Exploring Divergent Public Sentiments and Community Patterns around**

***David Beckham’s***

**Football Legacy**

**On**

**Reddit and YouTube**

**Submitted By – Group 7**

**Preetpal Singh (s4062787)**

**Saurabh Sharma (s4038560)**

**Shivam Shinde (s3994666)**

Contents

[**1.** **Introduction** - 3 -](#_Toc180124436)

[**2.** **Data Description** - 3 -](#_Toc180124437)

[**3.** **Methodology** - 4 -](#_Toc180124438)

[**3.1 Data** - 4 -](#_Toc180124439)

[**3.2 Handling Missing and Duplicate Values** - 4 -](#_Toc180124440)

[**3.3 Data Pre-Processing and Exploration** - 5 -](#_Toc180124441)

[**3.4 Term Frequency** - 6 -](#_Toc180124442)

[**Reddit Data Term Frequency** - 6 -](#_Toc180124443)

[**Alternative Perspective for Stakeholders – Unique Words** - 7 -](#_Toc180124444)

[**YouTube Data Term Frequency** - 7 -](#_Toc180124445)

[**Alternative Perspective for Stakeholders – Unique Words** - 8 -](#_Toc180124446)

[**Overall Conclusion of Sentiments Frequency on Reddit and YouTube** - 9 -](#_Toc180124447)

[**4.** **Data Modelling** - 9 -](#_Toc180124448)

[**4.1 TF-IDF Vectorization** - 9 -](#_Toc180124449)

[**4.2 Sentiment Analysis (VADER)** - 9 -](#_Toc180124450)

[**4.3 Latent Dirichlet Allocation LDA Modelling** - 11 -](#_Toc180124451)

[**4.4 Topic Labelling** - 12 -](#_Toc180124452)

[**4.5 Word Cloud Visualization for Topics** - 14 -](#_Toc180124453)

# **Introduction**

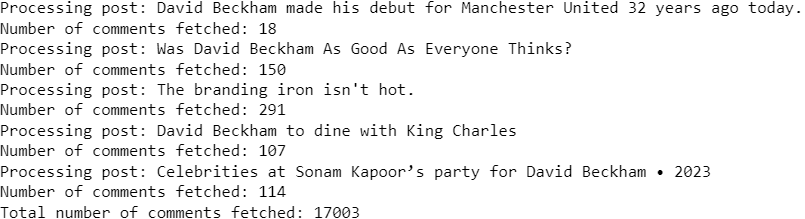
**David Beckham** is not just a footballer; he is a global icon who presence extends far beyond the pitch. From leading England’s national team to his ventures in fashion and philanthropy, Beckham’s life continues to inspire and intrigue people worldwide. But how do fans and public talk about him today? By analyzing discussions on two major social media platforms – **Reddit and YouTube** – we explore not only what people say about Beckham but also what these conversations reveal about his impact on culture, sports and beyond.

# **Data Description**

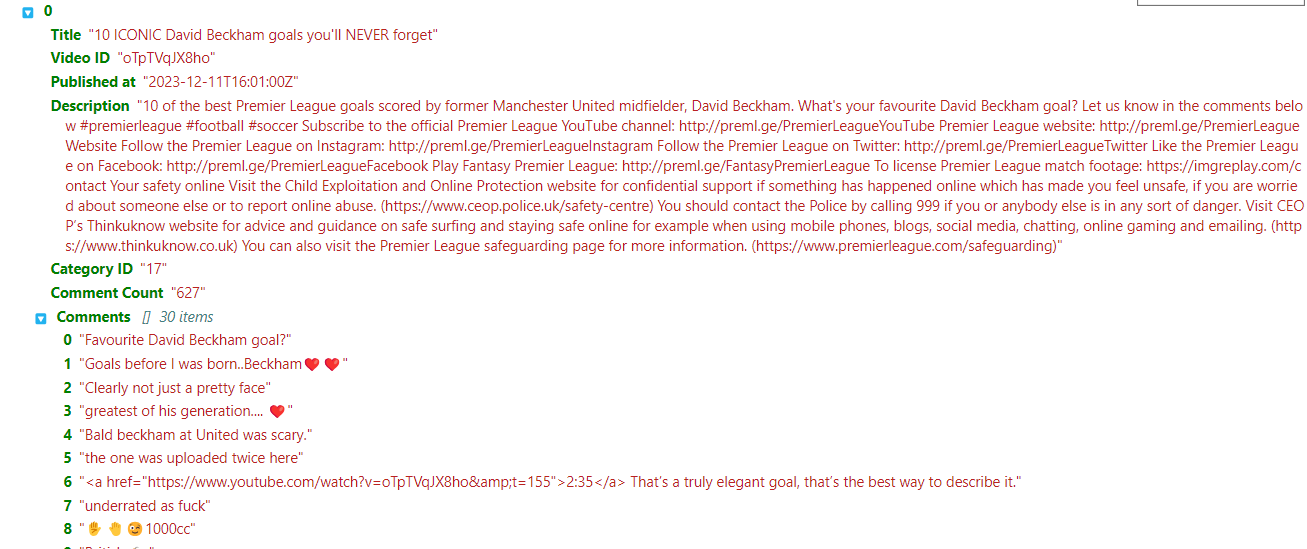
Imagine two arenas of conversation: Reddit, where debates rage about everything from sports statistics to celebrity gossip, and Youtube, where fans relive Beckham’s iconic moments through highlight reels and documentaries. We tapped into these platforms to gather thousands of user comments related to David Beckham. Reddit gave us the inside track of fan debates, while Youtube provided raw, emotional responses to Beckham’s on-screen appearances and football highlights.

Python Reddit API Wrapper **(PRAW)** was used which allowed the extraction of relevant posts and comments. **100** posts were extracted containing **17,003** comments over a period of one yearto get a broader view of public sentiment. **YouTube Data API v3** was used to extract **30 videos** containing a total of **83,825** comments.

David Beckham’s influence extends far beyond the football field, with discussions across both YouTube and Reddit reflecting his multifaceted legacy. On YouTube, fans celebrate his iconic moments, particularly his legendary free kicks and leadership on the pitch, while his media appearances, playful pranks and interviews showcase his charm and sense of humour. His personal life draws admiration, with many praising his supportive relationship with his wife, Victoria (Formerly member of Iconic Pop band group **Spice Girls**) and his role as a family man. On Reddit, conversations delve deeper into his family dynamics, often focusing on humorous and touching moments, such as Beckham’s 25th wedding anniversary and his wife Victoria’s comments about her background. His **football legacy** remains a prominent topic on both the social media platforms, with fans reminiscing about his free kicks and influence on the sport, and comparing him to contemporary players like **Messi** and **Neymar.** Additionally, Reddit comments highlight Beckham’s continued relevance in pop culture and media, with users reflecting on his **Netfilx documentary “Beckham”**, endorsements and legal disputes. **Overall**, comments across both platforms reveal a deep admiration for **Beckham as a football legend, a global celebrity, and a cultural icon.**



***Fig. 2.1 sample discussion titles about Beckham as extracted from Reddit***



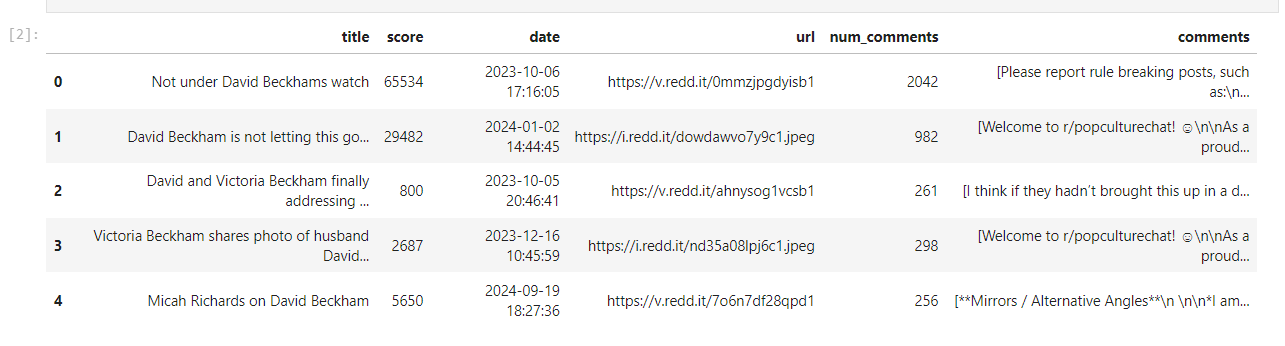
***Fig. 2.2 sample discussion data about Beckham as extracted from YouTube***

These comments and descriptions were subjected to text analysis techniques to extract meaningful insights regarding trending topics and public sentiments.

# **Methodology**

**3.1 DataLoading**

The first step involved loading the dataset in JSON format into a panda DataFrame. This was done for both the YouTube and Reddit datasets to facilitate further analysis.



***Fig: 3.1 Raw Reddit Data into Dataframe***



***Fig: 3.2 Raw YouTube Data into Dataframe***

## **3.2 Handling Missing and Duplicate Values**

Before proceeding with data preprocessing, it is essential to ensure that the dataset is clean and free of issues that could negatively impact the analysis, such as missing or duplicate values. To this end, the following checks were performed:

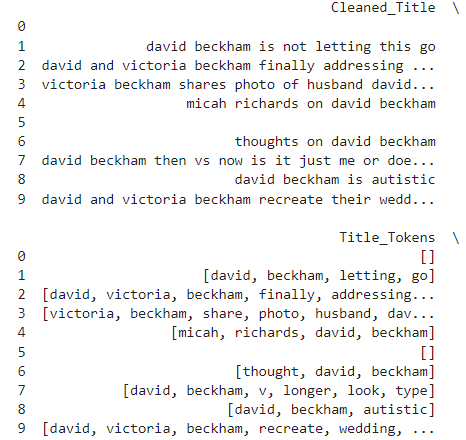
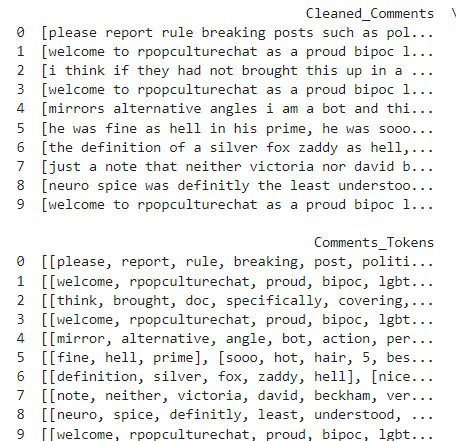
* **Checking for Missing Values**: A summary of missing values was generated for all columns in the dataset using the **isnull()** function. The check revealed that there were no missing values in the dataset, as indicated by the absence of any null values across all columns.
* **Checking for Duplicate Values**: Duplicate entries were checked using the **duplicated ()** function. The check showed that no duplicate rows were present in the dataset based on the unique identifier (Video ID for YouTube data and url for Reddit data). Thus, no further action was necessary to handle duplicates.

Although no missing or duplicate values were found, performing these checks was crucial to ensure the integrity of the dataset before conducting any further analysis.

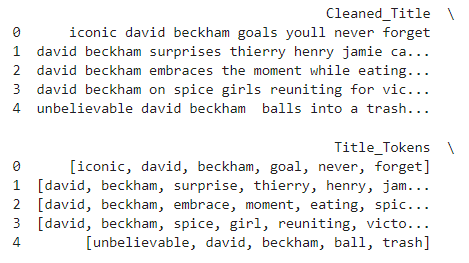
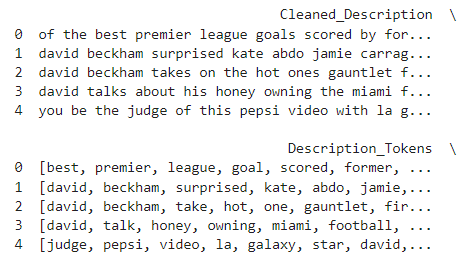
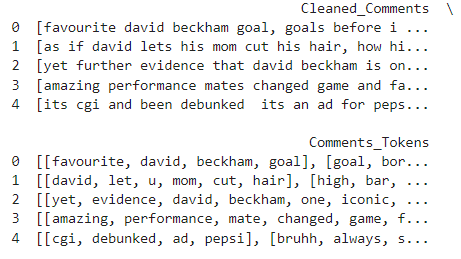
## **3.3 Data Pre-Processing and Exploration**

The data collected from Reddit and YouTube underwent several preprocessing steps to ensure it was suitable for analysis. These steps included: -

* **Removing URLs and Special Characters**: All URLs and special characters were removed from the text using regular expressions, and the text was converted to lowercase.
* **Handling Contractions**: Contractions were expanded using the contractions package to ensure consistency in the text.
* **Language Detection**: To ensure that only English text was processed, the langdetect library was used to filter out non-English text.
* **Tokenization** – Breaking down text into individual words, allowing for detailed analysis of word frequency and usage.
* **Stopword removal** – Removing common words that do not contribute significantly to the analysis (eg “and, “the”, etc.)
* **Lemmatization** – Converting words to their base forms to standardize the text and reduce dimensionality.

***Fig 3.3: Processed Data (Reddit)***

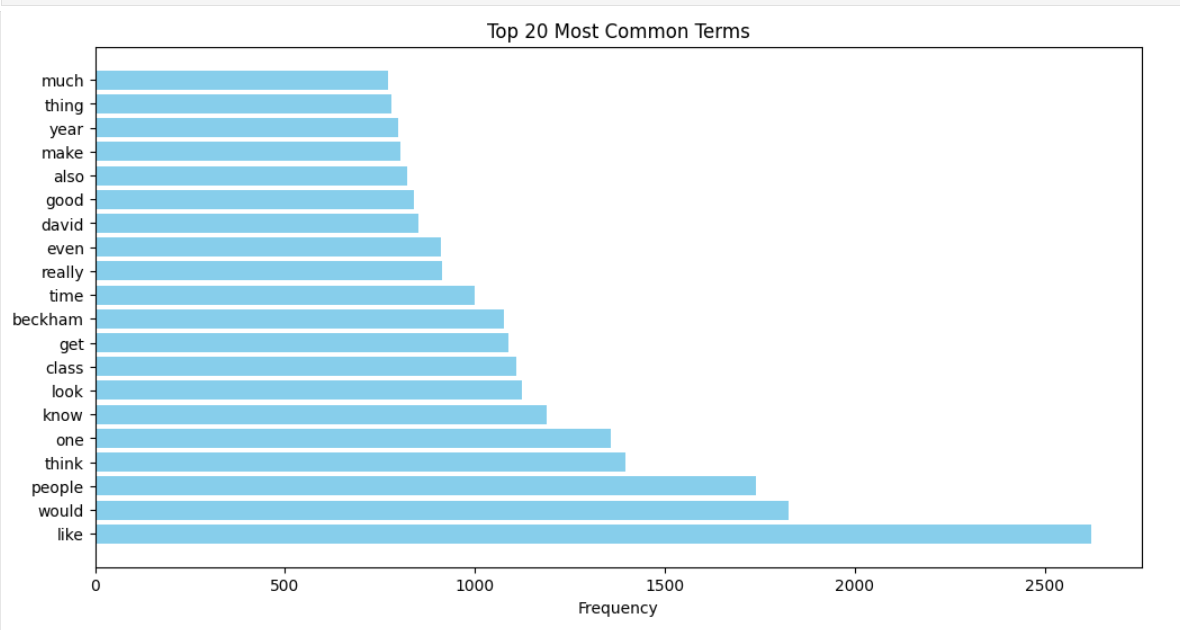
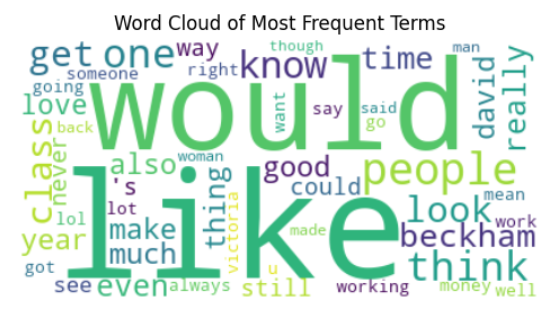
  

***Fig 3.4: Processed Data (YouTube)***

## **3.4 Term Frequency**

After cleaning the textual data (removing stopwords, tokenizing, and lemmatizing), we calculated the **term frequency** for both the title, description, and comments sections of each dataset (YouTube and Reddit). This was done to identify the most commonly occurring words, which provide insights into the key themes or topics within the data. The textual data, including the cleaned titles, descriptions, and comments, was first tokenized and lemmatized using the **WordNetLemmatizer** from the NLTK library. Stopwords, which are commonly used words that do not contribute much meaning (e.g., "and", "the"), were removed from the text. This ensured that the terms contributing to the frequency count were relevant.

### **Reddit Data Term Frequency**

**** ****

***Fig 3.5 Top 20 most common terms Fig 3.6 – Word Cloud***

The bar chart and word cloud together of **Reddit data** provide a comprehensive view of the **sentiment-heavy words** frequently used in the discussions about David Beckham, **focusing on words that carry significant emotional weight**. This approach is particularly effective for sentiment analysis because: -

* **Relevance to Sentiment (Capturing Emotional Intensity)** – Visualizations highlight the emotional tone and opinions expressed, showcasing words like **“like” (2624), “good” (840), “love” (719)** dominates, showing strong admiration for his career, personality, and family life.
* **Actionable Insights for Strategic Decisions**

1. **Leverage Positive Sentiment** – The overall positivity can be harnessed for branding opportunities, endorsements and partnerships, particularly with products that align with Beckham’s values and family-oriented image.
2. **Maintain Relevance** – Words like **“still”** and **“even”** indicate Beckham’s ongoing relevance. Strategic efforts would focus on keeping him visible through media and public appearances.
3. **Family Image** – The focus on his wife “Victoria” suggests joined branding opportunities featuring the Beckham family could resonate well with the audiences.
4. **Address Criticism** – Although minimal, potential criticism can be addressed through proactive PR strategies that reinforce positive aspects of Beckham’s legacy.

**By focusing on sentiment heavy words rather than just unique words**, both the visualizations provide a richer understanding of public sentiments surrounding David Beckham. Key topics of discussion are revealed along with emotions that are driving those discussions. This approach **ensures that analysis is both relevant and practical, offering stakeholders clear insights into public opinions and guiding effective strategic communication efforts.**

### **Alternative Perspective for Stakeholders – Unique Words**

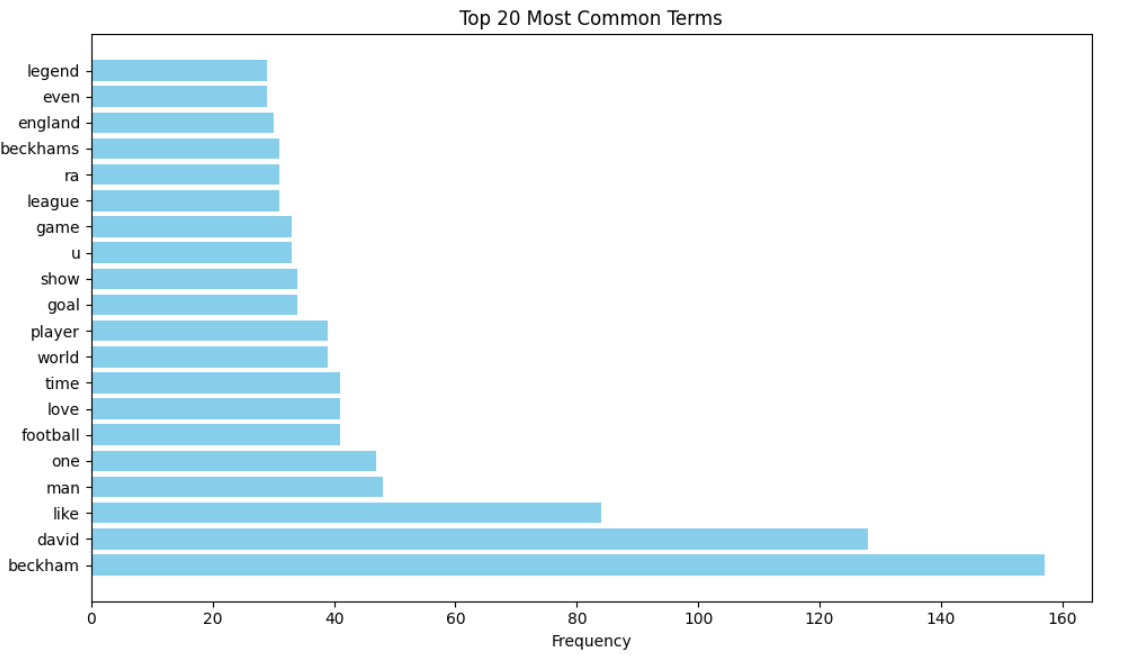
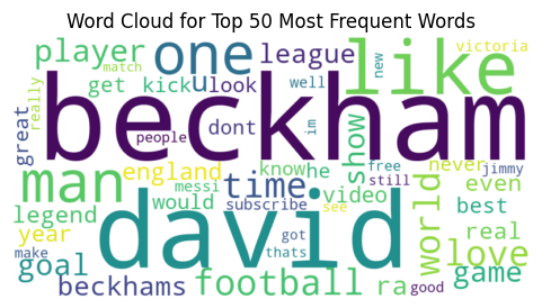
While sentiment heavy words provide critical insights into the emotional tone and intensity of discussions, **identifying unique words offers a complementary view that highlights specific entities, topics or concepts being discussed.**

***Unique words – based on context of discussions and not on frequency***

* **Victoria** (Specific Person) - Frequent mentions of Victoria, highlight the public’s interest in his relationship with his wife, adding to his image as a family man.
* **David** (Specific Person) – Represents a more personal reference to Beckham himself, indicating intimate or individual focused conversations.
* **Right** (Affirming statements or comparison) – Used in debate or affirmations about Beckham’s abilities or decisions.
* **Well** (Performance or outcome) – Reflects commentary on how Beckham has performed or handled situations.
* **Could** (Speculative discussions) – Suggests hypothetical discussions about Beckham’s career, life choices, or potential future roles.
* **Really** (Emphasis or intensity) – Used to amplify or emphasize sentiments often reinforcing admiration or posing questions about his legacy.
* **Mean** (Understanding context) – Suggests where people seek clarity or questions aspects of Beckham’s life.

This approach provides a holistic understanding of both **what is being talked about (such as Beckham’s personal and professional life) and how it is being talked about (through speculative, affirming, or critical discussions)** on **Reddit** enabling more strategic and targeted actions in response to public disclosure.

### **YouTube Data Term Frequency**

 ****

***Fig 3.7 Top 20 most common terms Fig 3.8 – Word Cloud***

The bar chart and word cloud together of **YouTube data** provide a comprehensive view of the **sentiment-heavy words** frequently used in the discussions about David Beckham, **focusing on words that carry significant emotional weight**. This approach is particularly effective for sentiment analysis because: -

* **Relevance to Sentiment (Capturing Emotional Intensity)** – Visualizations highlight the emotional tone and opinions expressed, showcasing words like **“beckham” (157), “david” (128), “like” (84), “player” (39), “love” (41)** dominates, showing strong admiration for his football career, personality, and legacy. Words like **“legend” (29), “good” (21) and “best” (26),** further reinforce the positive sentiment surrounding his accomplishments as a footballer.
* **Actionable Insights for Strategic Decisions**

1. **Leverage Positive Sentiment** – The overall positivity can be harnessed for branding opportunities, endorsements, and partnerships, particularly with products that align with Beckham’s image as a football legend and sports icon.
2. **Maintain Relevance** – Words like **“still”** and **“ever”** indicate Beckham’s ongoing relevance in football. Strategic efforts would focus on keeping him visible through media and public appearances.
3. **Football Legacy** – The focus on terms like **“football” (41), “goal” (34), “match” (19) and “game” (33)** suggests endorsements or partnerships with sports brands would strongly resonate with the audiences on YouTube platform.
4. **Address Criticism** – Although minimal, words like **“never” (15) and “dont” (70),** suggest some level of criticism, likely comparisons with other footballers. These can be addressed through proactive PR strategies that reinforce positive aspects of Beckham’s legacy and achievements.

**By focusing on sentiment heavy words rather than just unique words**, both the visualizations provide a richer understanding of public sentiments surrounding David Beckham. Key topics of discussion are revealed along with emotions that are driving those discussions. This approach **ensures that analysis is both relevant and practical, offering stakeholders clear insights into public opinions and guiding effective strategic communication efforts.**

### **Alternative Perspective for Stakeholders – Unique Words**

While sentiment heavy words provide critical insights into the emotional tone and intensity of discussions, **identifying unique words offers a complementary view that highlights specific entities, topics or concepts being discussed.**

***Unique words – based on context of discussions and not on frequency***

* **Football** (Specific Sport) - Frequent mentions of **“football”**, highlight the public’s interest in Beckham’s legacy as a legendary football player.
* **Player** (Football Role) – Represents Beckham’s professional role, reinforcing admiration for his on-filed performance and legendary status.
* **Goal** (Achievements) – Words like **“goal”** emphasize Beckham’s contributions and memorable moments in football, showcasing his excellence in the sport.
* **England** (National Identity) – Discussions around **“England”** refer Beckham’s role as an English football star, often linking his career to national pride.
* **Team** (Collaborative Role) – The word **“team”** reflects Beckham’s association with multiple successful football clubs and highlights his role in a team context.
* **Messi** (Comparison) – Mention of **“Messi”** suggests comparison between Beckham and other iconic players which can represent opportunities to discuss Beckham’s unique legacy.
* **Legend** (Status) – Beckham’s status as a **“legend”** reflects his enduring influence in football and pop culture, highlighting his global impact.

This approach provides a holistic understanding of both **what is being talked about (such as Beckham’s professional life as a footballer) and how it is being talked about (through discussions of his role as a player, his legendary status and comparison with other players)** on **YouTube** enabling more strategic and targeted actions in response to public disclosure.

### **Overall Conclusion of Sentiments Frequency on Reddit and YouTube**

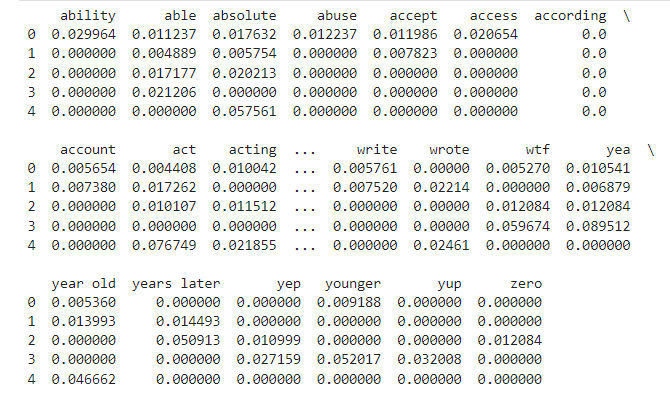
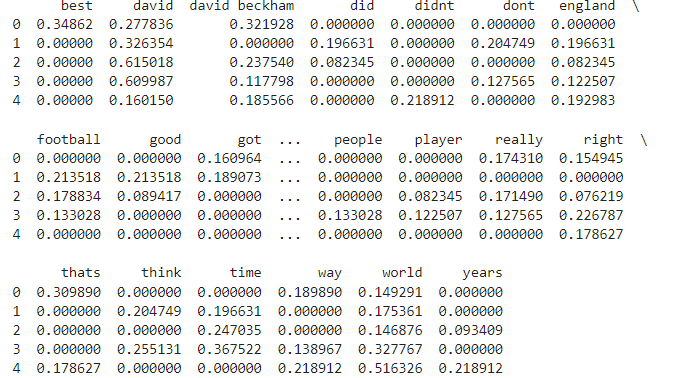
On **Reddit**, Beckham’s **family image** and **celebrity status** are discussed alongside his career, reflecting a **more diverse admiration** for him as a public figure.

On **YouTube**, the focus is primarily on Beckham’s **football career**, with a strong emphasis on his **athletic accomplishments** and **legendary status** in the sport. Both platforms show **positive sentiments** but then nature of discussions reflect the **specific interests** of each audience: personal life on Reddit and football achievements on YouTube.

# **Data Modelling**

## **4.1 TF-IDF Vectorization**

The TF-IDF vectorization process transforms the textual data into a numerical representation that highlights the importance of specific words in a comment, providing stakeholders with a focused view of the most relevant terms in discussion about David Beckham. TF-IDF assigns a score to each word, balancing how often it appears in a single comment (term frequency) and how rare it is across all the comments (inverse document frequency). Words that occur frequently in a specific comment but are rare across the dataset receive higher scores. By analyzing the TF-IDF scores, stakeholders can prioritize content, tailor communication strategies, and gain deeper insights into the specific subjects that resonate within the conversations.

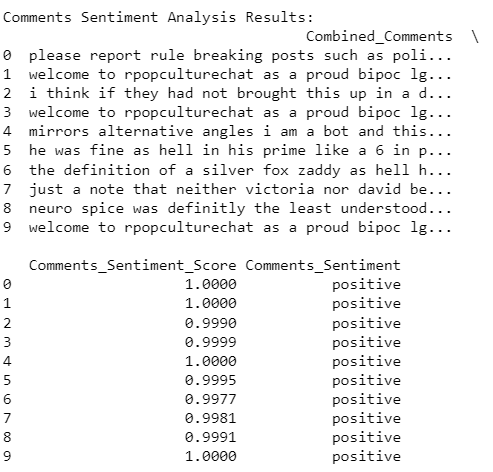
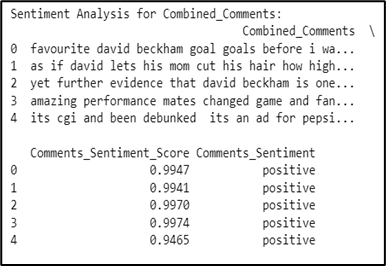
***Table 4.1 – TF-IDF Vectors – (Reddit) Table 4.2 – TF-IDF Vectors – (YouTube)***

**Reddit** has a broader range of discussions (1000 columns), including both personal life and casual comments.

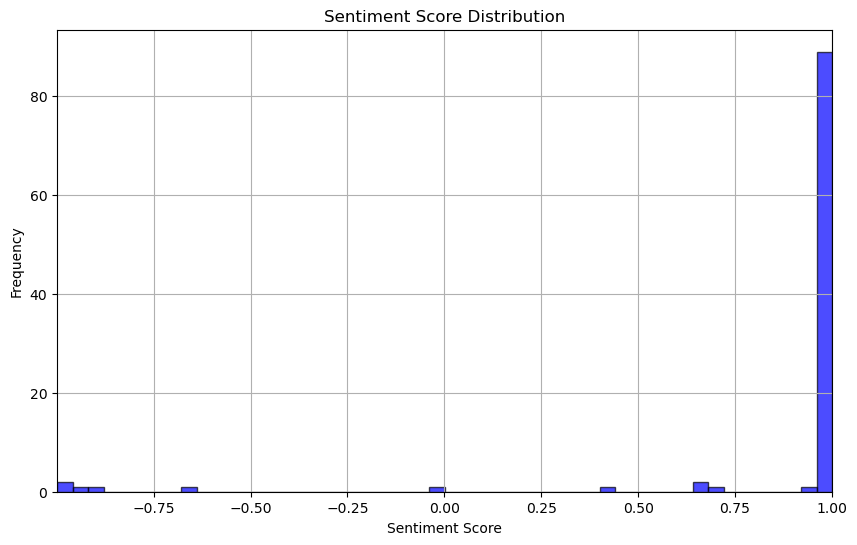
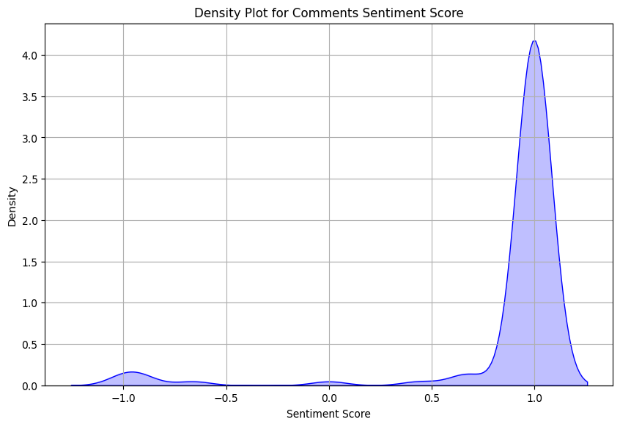
**YouTube** has a more focused set of terms (327 columns), concentrating on Beckham’s football legacy, goals, and achievements.

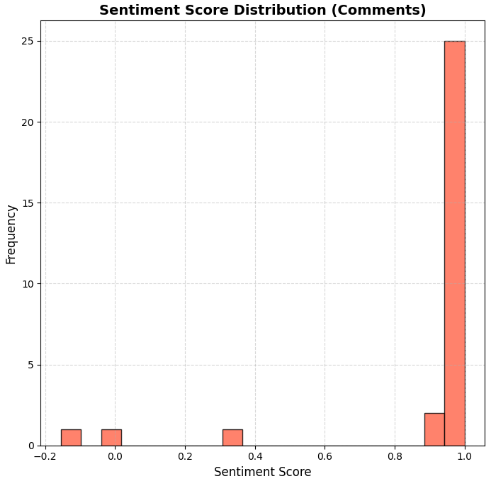
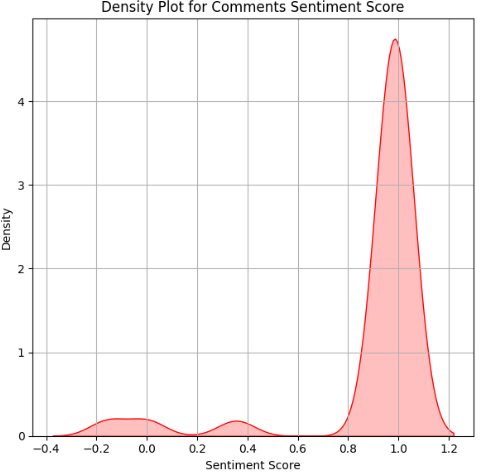
## **4.2 Sentiment Analysis (VADER)**

The VADER sentiment analyzer was used to assess the sentiment of each comment. VADER is particularly suited for social media text as it accounts for the context of the words, including slang and emoticons, providing a compound score that reflects the overall sentiment.

** **

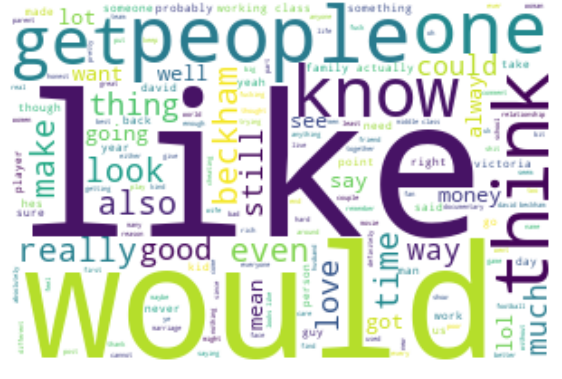
***Fig 4.3 – VADER Analysis – (Reddit) Fig 4.4 – VADER Analysis – (YouTube)***

Each comment has been assigned a **sentiment score ranging from -1 to 1**, where negative value indicates a negative sentiment, positive value indicates positive sentiment and scores close to zero (0) indicates neutral sentiment. These scores are then categorized into **three sentiment categories of negative, positive, and neutral.** This gives a clear understanding of how different comments reflect varying degrees of sentiment towards David Beckham. **This categorization is crucial for stakeholders to identify areas of concern, positivity, or neutrality in public discussions.**

***Fig.4.5–Histogram-Sentiment Score-All Comments (Reddit) Fig.4.6–Density Plot-Sentiment Score-All Comments (Reddit)***

***Fig.4.7–Histogram-Sentiment Score-All Comments (YouTube) Fig.4.8–Density Plot-Sentiment Score-All Comments (YouTube)***

On **Reddit**, the discussions are broader but with a clear skew towards **highly positive sentiment**, as indicated by most sentiment scores being near 1.0 in the histogram and density plots. There is still a presence of both **neutral and positive sentiment** but positive sentiment overwhelmingly dominates, reflects stronger admiration for Beckham. However, the diversity of discussions still highlights Reddit’s tendency to engage in broader debates and discussions that cover Beckham’s **personal life, public image, and career highlights**. Meanwhile, **YouTube** shows a more focused discussion around Beckham’s **football legacy** with a **clear positive sentiment bias**. Both platforms reflect admiration for his career achievements, but the sentiment on Reddit suggests a much more positive perception than initially assumed, with relatively few critiques. **Overall**, YouTube portrays Beckham as a revered football figure, while Reddit highlights a similarly positive sentiment towards Beckham, primarily focusing on his personal and professional life. This comparative sentiment analysis suggests that Beckham’s football career resonates more positively on both platforms, with Reddit showing slightly more varied sentiment scores but still leaning heavily towards more on positive side.

*** Fig 4.7 – Word Cloud of All Comments (Reddit) Fig 4.8 – Word Cloud of All Comments (YouTube)***

**On Reddit**, conversations are more **opinion-driven,** with prominent words like **“people”, “think” and “know”,** indicating a focus on sharing thoughts and reflecting on public perceptions. Reddit users also delve into Beckham’s personal life, with mentions of his wife **Victoria, the personal aspects are not as dominant as the reflective discussions on broader topics**. In contrast, YouTube discussions focus heavily on Beckham’s football career, with words like **“love”, “football”, “player” and “legend”** dominating. YouTube users admire Beckham’s achievements, highlighting his football legacy and individual accomplishments. **Overall**, Reddit offers more personal discussions, while YouTube centres on Beckham’s **sports legacy** and **public admiration**.

Above approach gives stakeholders a comprehensive understanding of both the content and tone of the discussions, enabling more informed decision making and strategic communication.

## **4.3 Latent Dirichlet Allocation LDA Modelling**

To provide a data driven response to **what people are discussing about David Beckham and how they perceive him, a detailed analysis using topic modelling and word clouds was conducted. The Latent Dirichlet Allocation (LDA)** Model with both Sckit-Learn and CountVectorizer were used to uncover key topics being discussed in relation to David Beckham.

To uncover hidden topics within the user-generated comments from both YouTube and Reddit datasets, we employed **Latent Dirichlet Allocation (LDA)** using the scikit-learn implementation. The following methodology was followed:

* For each dataset, the cleaned comments were combined into a single string for each row, allowing the model to consider all comments associated with a particular post or video. This step facilitated efficient processing for the LDA model.
* To convert the textual data into a format suitable for topic modelling, we used the **CountVectorizer** from scikit-learn. The vectorizer transforms the combined comments into a **document-term matrix (DTM)**, where each document is represented as a vector of term frequencies.
* With the document-term matrix ready, we applied the LDA model to discover the underlying topics within the comments. We set the number of topics to 10, as a balance between detail and interpretability.
* Once the model was trained, we extracted and displayed the top words associated with each topic. These words help interpret the content of each topic and identify the general themes present in the comments.
* The LDA model generated 10 distinct topics for each dataset, with the top words providing insights into the recurring themes within the comments. Below is an example of the output for both the YouTube and Reddit datasets:



***Fig 4.9 LDA Topics Generated - Reddit***



***Fig 4.9 LDA Topics Generated – YouTube***

## **4.4 Topic Labelling**

After generating the topics through Latent Dirichlet Allocation (LDA), we assigned descriptive labels to each topic based on the most frequent terms. This process helped to contextualize and better interpret the topics for both the Reddit and YouTube datasets.

* **Reddit Topic Labels**

The LDA model identified 10 distinct topics from the Reddit comments. Based on the top words for each topic, we manually assigned the following labels to describe the content of each topic:

|  |  |
| --- | --- |
| **Topic Number** | **Assigned Label** |
| Topic 0 | David Beckham's Football Career and Qatar Influence |
| Topic 1 | Public Perception of Harry and Victoria Beckham |
| Topic 2 | David and Victoria Beckham's Relationship and Family Life |
| Topic 3 | Football Players and Their Performance on the Field |
| Topic 4 | Automated Actions and Moderation in Online Discussions |
| Topic 5 | Public Opinions on David and Victoria Beckham's Public Image |
| Topic 6 | Class and Economic Status in Society |
| Topic 7 | David Beckham's Style and Public Image Over the Years |
| Topic 8 | Family and Parenting in the Beckham Household |
| Topic 9 | David Beckham's Media and Public Representation |

These topic labels provide insight into the range of discussions on Reddit, from Beckham's football career to public perception and family life.

* **YouTube Topic Labels**

Similarly, 10 topics were identified from the YouTube comments and descriptions combined, and the following labels were assigned:

|  |  |
| --- | --- |
| **Topic Number** | **Assigned Label** |
| Topic 0 | David Beckham’s History and Achievements |
| Topic 1 | Football Matches and Game Highlights |
| Topic 2 | Match Insights and David Beckham’s Strategy |
| Topic 3 | Victoria Beckham and David’s Family Life |
| Topic 4 | David Beckham’s Club and Fans |
| Topic 5 | Football Career and Legacy of Ronaldo and Beckham |
| Topic 6 | David Beckham’s Football Skills and World Cup Performance |
| Topic 7 | Real Madrid and Beckham’s Performance in Football |
| Topic 8 | Football Achievements and Messi vs Beckham |
| Topic 9 | Real Madrid and Beckham’s Role in Football |

These topics highlight the focus on Beckham's football career, his relationship with other football legends like Ronaldo and Messi, and his influence on the Real Madrid football team.

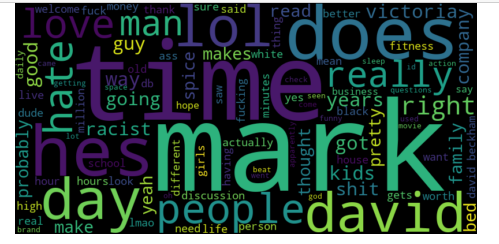
By assigning these labels, we were able to categorize the primary themes and patterns in user discussions across both platforms. The Reddit dataset reflected a broader range of topics, including public opinions, family life, and media representation, while the YouTube dataset focused heavily on football-related content and Beckham’s career achievements.

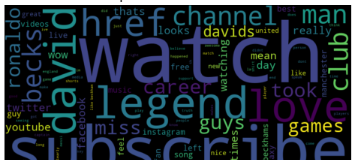
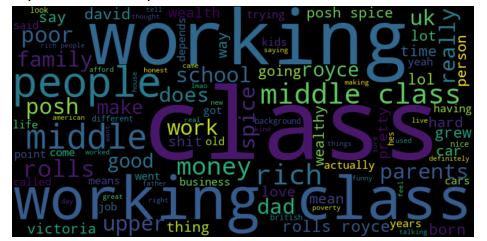
## **4.5 Word Cloud Visualization for Topics**

To further enhance the interpretation of the topics generated by the LDA model, we created word clouds for each topic. Word clouds provide a visual representation of the most frequently occurring terms in each topic, with larger words indicating higher frequency. For each identified topic, a word cloud was generated using the top words extracted from the LDA model. The **WordCloud** library in Python was used to create the visualizations. For each dataset, the word clouds were generated using the top terms for each topic identified by the LDA model. The word clouds visually highlight the most important words in the topics, aiding in the interpretation of the themes.

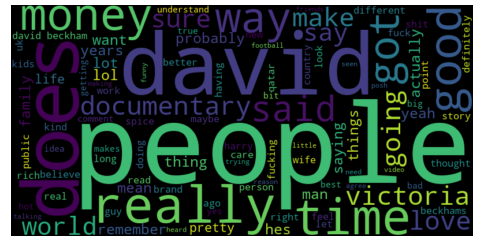
**Trending Topics associated with David Beckham on Reddit and YouTube**

**Reddit YouTube**

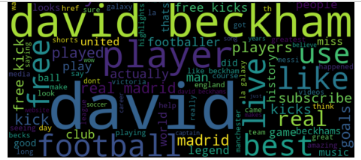
 

***Topic - David and Victoria Beckham's Relationship and Family Life Topic – Match Insights and David Beckham’s Strategy***

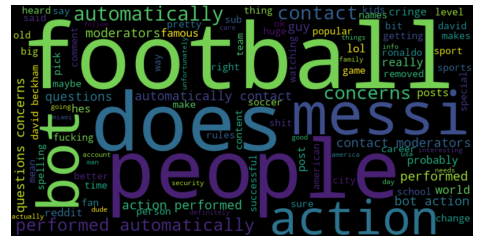
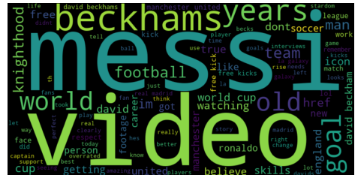
***Topic - Public Opinions on David and Victoria Beckham's Public Image Topic – David Beckham’s Clubs and Fans***

****

***Topic - David Beckham's Media and Public Representation Topic - Football Career and Legacy of Ronaldo and Beckham***



***Topic - David Beckham's Football Career and Qatar Influence Topic - Real Madrid and Beckham’s Performance in Football***



***Topic - Family and Parenting in the Beckham Household Topic - Football Achievements and Messi vs Beckham***

**Reddit** discussions reflect a more balanced view, incorporating both personal critiques and professional admiration, often diving into socio-political topics, Beckham’s public perception and his media portrayal.

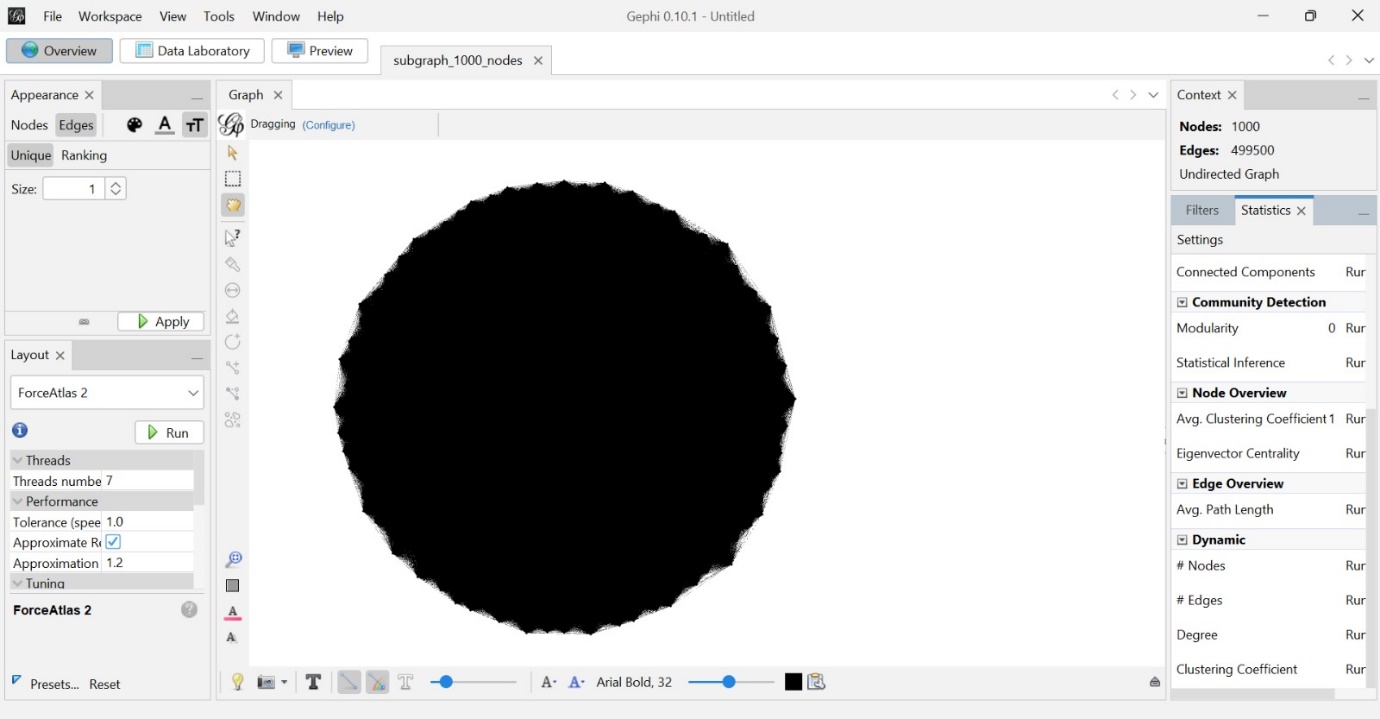
**YouTube,** reveals a more football-centric conversations. Terms like “match” and “watch” dominate, reflecting, platform’s strong focus on Beckham’s football career and his on-field achievements. Terms like “legend”, “love, “goal”, reflects overwhelming positive sentiment towards his career accomplishments. Comparison with football icons “ronaldo” and “messi” indicating debates over his football legacy in comparison to other greats.

**Technical Insights –** LDA topic modelling successfully identifies distinct clusters of discussions across both platforms, which are reflected in the word clouds. This validates the effectiveness of the LDA model in uncovering meaningful insights, as each platform’s discussions have been accurately captured. The model’s application through Gensim and CountVectorizer ensures that high frequency terms are mapped to specific topics. Allowing for clearer identification of themes. Visualizing these topics through word clouds further enhances stakeholder’s ability to quickly gasp the nature of discussions.

**Stakeholder Implications** – For stakeholders, these insights are crucial in understanding the distinct narratives around David Beckham across different social media platforms. On Reddit, the discussions are broader and often more critical. Stakeholders can focus on managing Beckham’s personal and public image, ensuring that any critiques or debated are addressed through strategic communication. On YouTube, where the conversations centres more around admiration for his football career, opportunities exist to further enhance Beckham’s legacy through engagement strategies focussed on celebrating his sports achievements.

# **Community Detection and Community Analysis**

Based on initial network analysis, we have found that for reddit data, there are around 6109 nodes and around 18.66 million edges. It is computationally very expensive to visualise these large number of nodes and edges. After randomly selecting 1000 nodes and 499500 edges from the dataset, it created a completely black hole type visualisation from which it is very difficult to get any meaningful inference.

. 

## **Approach to get meaningful results from Community Analysis**

To do meaningful community analysis, we need dataset related to communities which are talking about some particular topic. The topics which are discussed among users can be successfully obtained with topic modelling. In the Heading 4, we already discussed about topic modelling.

In the topic modelling, we got different topics related to Beckham. We carefully analysed the topics and add them to suitable group. We have formed 3 groups related to different aspects of David Beckham’s life. These groups are,

* **Beckham’s Personal Life**
* **Beckham’s Football Career**
* **Beckham’s Media and Public Perception**

After forming the groups, we have extracted the documents associated with these groups. In Community Analysis we have utilized those extracted documents.

## Why Combining Topics

We have combined the multiple topics because for individual topics, the number of nodes and edges generated were very less in number which does not produce very meaningful visualisations. After combining multiple topics, the results were easy to visualise and meaningful inference can be obtained from them.

## 5.3 **Community Analysis for Reddit Data**

After carefully studying different topics associated with reddit data of David Beckham, We have grouped the topics into these groups.

**Group 1: Personal Life**

* **Topic 0**: *"Beckham's Marriage, Cheating Rumors, and Family"*
  + Focuses on Beckham's personal life, especially rumors of infidelity, his marriage to Victoria, and family matters.
* **Topic 4**: *"Relationships and Marriage: Love and Life"*
  + Discusses relationships, love, marriage, and life, with general mentions of Beckham.
* **Topic 7**: *"David Beckham's Wedding and Appearance"*
  + Talks about Beckham’s wedding, his appearance, and related aspects such as his looks and personal grooming.
* **Topic 10**: *"Beckham’s Looks and Plastic Surgery Rumors"*
  + Focuses on Beckham’s physical appearance, mentions of potential plastic surgery, and how he is perceived.

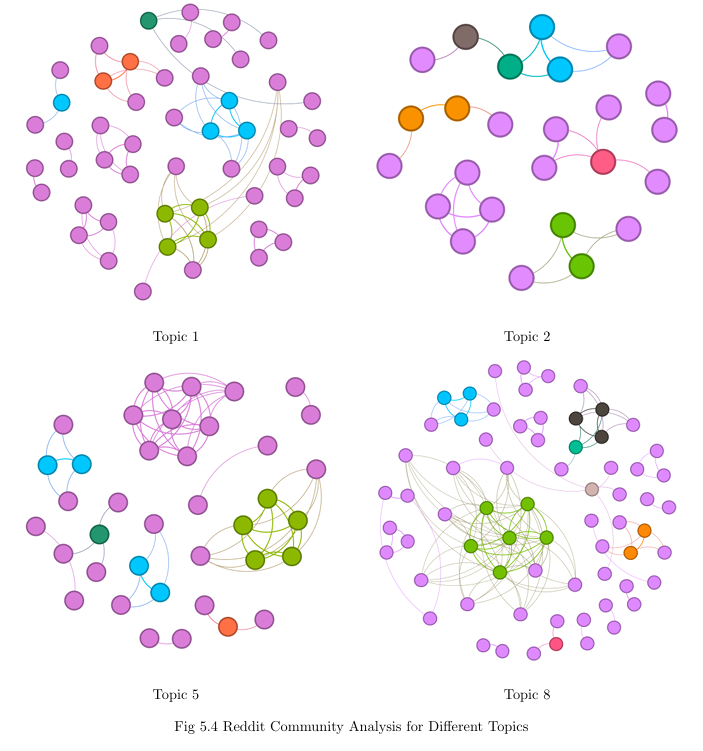
**Group 2: Football Career**

* **Topic 6**: *"David Beckham’s Football Career: Team and Players"*
  + Discusses Beckham’s football career, his team, his role in England, Manchester United, and his playing style.
* **Topic 14**: *"Beckham’s Football Career and Success in the League"*
  + Focuses on Beckham’s career in major football leagues like the Premier League, including mentions of United and Real Madrid.
* **Topic 17**: *"Beckham's Football Achievements: Goals, Free Kicks, and England"*
  + Talks about Beckham’s football achievements, goals, and free kicks, especially his time with England.

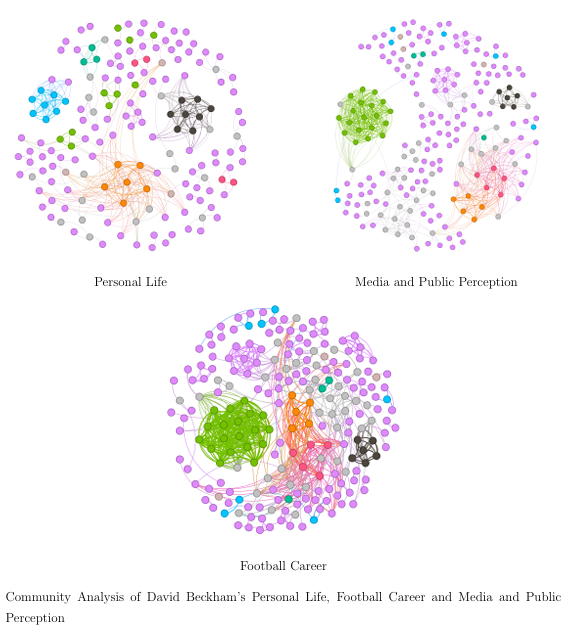
**Group 3: Media and Public Perception**

* **Topic 1**: *"Beckham's Image in Movies and Public Perception"*
  + Discusses Beckham’s public image, particularly his involvement in movies and media appearances.
* **Topic 2**: *"Beckham’s Working-Class Roots and Social Status"*
  + Talks about Beckham’s social status, working-class background, and his rise to fame, with mentions of “posh” and middle class.
* **Topic 5**: *"Public Banter About Beckham and Rooney"*
  + Reflects public banter and discussion about Beckham, Rooney, and their respective careers, focusing on looks and football.
* **Topic 8**: *"Beckham’s Physical Appearance and Aging"*
  + Focuses on Beckham’s looks, hair, and aging, including how his appearance has changed over time.

In the below graph, we have done the community analysis of topics related to David Beckham’s Media and Public Perception. We have plotted the communities formed by documents of topics 1, 2, 5 and 8. **The pink color nodes are the nodes with 0 betweenness centrality.** **Betweenness centrality measures the extent to which the node acts as a bridge in a network**. **We can see from the below graph that large number of nodes which have 0 betweenness centrality which means there are very few nodes which works as bridges between communities**.



Below we have combined the topics for different groups and generated the community graphs for David Beckham’s personal life, Football Career and media and public perception.



Here, we will explain more about the community graph of Football Career of David Beckham. In the Football career graph around 59.91% have zero betweenness centrality(pink nodes) which is comparatively less than individual topic graphs. Around 30.88% of nodes have closeness centrality 1 which means 30.88% of nodes can reach all other nodes in the network with just one step. 8.76% of nodes have closeness centrality of 0.31%.

## **5.4 Community Analysis of YouTube Data**

**Group 1: Beckham’s Personal Life**

* **Topic 0: *"Beckham's Family, Wife, and Life Beyond Football"***
  + **Focuses on Beckham’s family, his wife (Victoria), and mentions of his personal life.**
* **Topic 3: *"Beckham’s Personal Life with Victoria and Public Persona"***
  + **Discusses Beckham’s relationship with Victoria (Posh Spice) and how his personal life is perceived.**
* **Topic 10: *"Beckham’s Family Life and Victoria Beckham"***
  + **Mentions Beckham’s family, his life with Victoria, and his class.**
* **Topic 6: *"David Beckham’s Legacy and Respect as a Person"***
  + **Highlights Beckham’s respect as an individual, focusing on his global legacy beyond just football.**

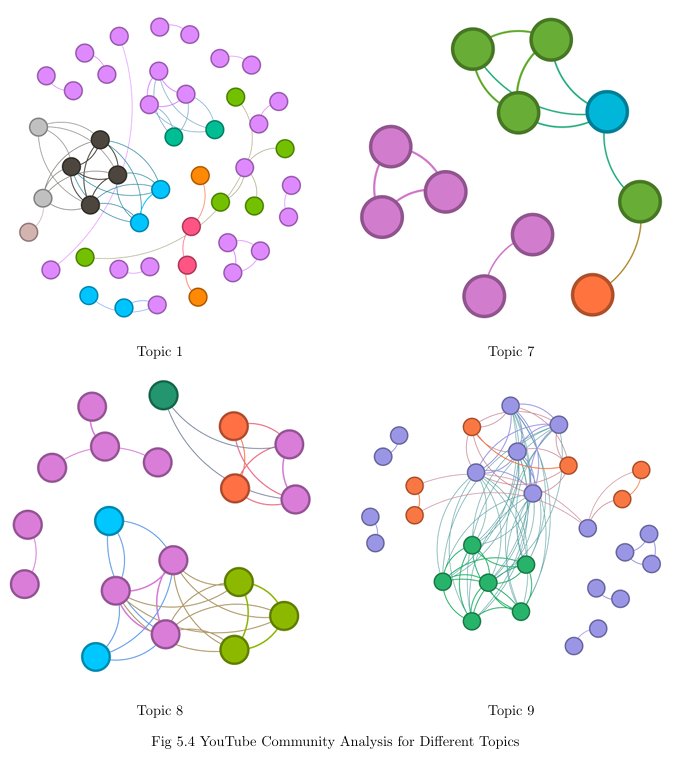
**Group 2: Beckham’s Football Career**

* **Topic 1: *"Beckham’s Football Career and Success in the Premier League"***
  + **Discusses Beckham’s football career, with a focus on his time in England and his skills as a player.**
* **Topic 7: *"Beckham’s Football Skills and Real Madrid Career"***
  + **Focuses on Beckham’s football skills, his time at Real Madrid, and his performance as a football player.**
* **Topic 8: *"Beckham’s Football Career: World Cup and Team Impact"***
  + **Mentions Beckham’s performance in World Cups and his role in various football teams.**
* **Topic 9: *"Beckham as a Football Legend: Real Madrid and Beyond"***
  + **Discusses Beckham’s football legacy, comparing him to other football legends like Messi, and his time at Real Madrid.**

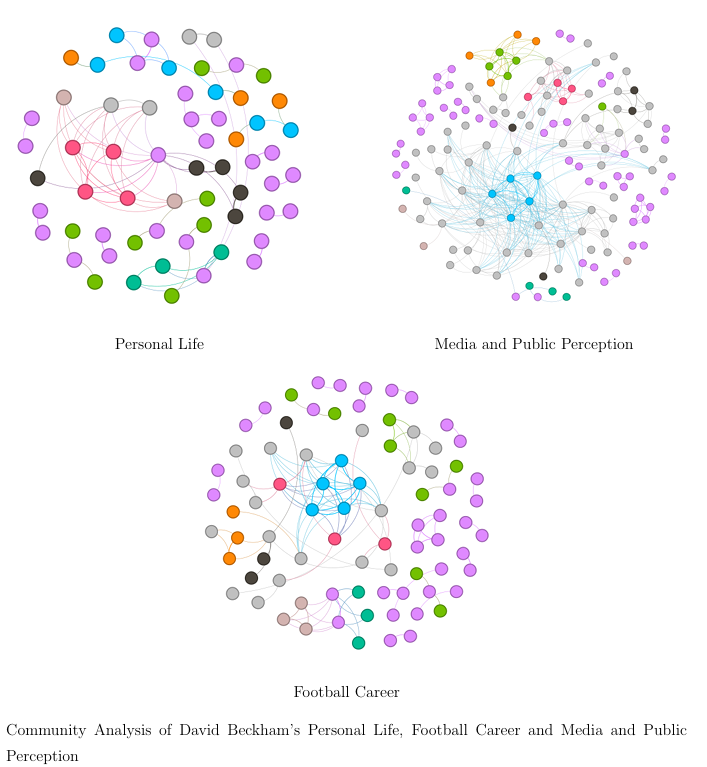
**Group 3: Beckham’s Media and Public Perception**

* **Topic 4: *"Beckham as a Media Icon: YouTube, Subscriptions, and Popularity"***
  + **Discusses Beckham’s presence in media, particularly his online influence, YouTube, and social media.**
* **Topic 5: *"Beckham’s Media Presence and Popularity Among Fans"***
  + **Reflects on Beckham’s popularity, his media presence, and how fans perceive him.**
* **Topic 11: *"Beckham’s Heroic Public Image and Media Appeal"***
  + **Talks about Beckham as a hero figure and his appeal to a broad audience through media.**
* **Topic 15: *"Beckham’s Public Persona and Leadership"***
  + **Mentions Beckham’s public leadership, his persona in the media, and his career highlights.**

In the below graph, there is collage of topics 1,7, 8 and 9 which comes under Group 1 which discusses Football career of David Beckham.



In the above graph, For example if we just consider topic 1 to analyse. In topic 1 around 73.92% of nodes have 0.0 betweenness centrality which means there are very few nodes which makes connections between two or more nodes. Besides this there are also 52.17% nodes have 1.0 closeness centrality. These nodes play an important role in the network’s overall functioning and information flow. If these nodes are removed then it could significantly impact the network’s connectivity.



In the above diagram, we show the graphs for David’s personal life, media and perception and football career. In this graph around 65.32% nodes have 0 betweenness centrality. In the graph, 34.48% nodes have closeness centrality of 1.

# Meeting With Imesh

In the meeting, we discussed the problems we faced in the project, received suggestions, and showcased our work distribution. These are the screenshots attached of our meetings with Imesh.

