Global Echoes of the FIFA World Cup 2022: Sentiment and Theme Analysis via Deep Learning and Machine Learning on Twitter

Jacob Jensen and Muhammad Abusaqer

Department of Mathematics and Computer Science

Minot State University

Minot, North Dakota 58703

jacob.d.jensen@minotstateu.edu and muhammad.abusaqer@minotstateu.edu

Abstract

On December 2, 2010, Qatar was chosen as the host for the 2022 FIFA World Cup, marking a historic moment as the first Middle Eastern nation to organize the tournament. The decision ignited a multitude of opinions across social media, with Twitter becoming a primary venue for global discourse. This research delves into an extensive corpus of 100,000 tweets to analyze the public sentiment and various topics that surfaced during the World Cup in Qatar. Employing RoBERTa for sentiment analysis and BERTopic, as well as Latent Dirichlet Allocation (LDA) for topic modeling, the study captures the rich tapestry of global conversation sparked by the event. The results reveal a predominantly positive sentiment, with over half of the tweets reflecting positivity. Topic modeling further dissects the discussions, uncovering themes from active engagement with the World Cup narratives and anticipation of key matchups to broader socio-cultural dialogues and emergent topics like digital collectibles. This examination of the digital narrative surrounding the FIFA World Cup 2022 emphasizes the event's expansive influence and the diverse global interests it encapsulated, offering a window into the worldwide community's digital interaction with one of the most significant sporting events.

1 Introduction

The FIFA World Cup [6] is more than just a global sporting event; it is a phenomenon that generates widespread conversation and debate. [7] In 2022, the World Cup was held in Qatar, marking the first time a Middle Eastern country hosted the World Cup. [8] This historic occasion sparked a flurry of online discussions, particularly on social media platforms such as Twitter. Recognizing the rich potential of this digital discourse, our study seeks to delve into the public sentiment and thematic trends expressed on Twitter during the 2022 FIFA World Cup. By harnessing the power of Natural Language Processing (NLP), Deep Learning, and Machine Learning, we aim to analyze a dataset comprising 100,000 tweets, which reflect the global opinion spectrum. Through sophisticated sentiment analysis and topic modeling, this research aims to uncover the prevailing attitudes and topics, offering insights into the public's response to this monumental event.

2 Related Work

Luis Fernando Nuñez Franco [3] explored the general sentiment on Twitter regarding the 2022 FIFA World Cup and found predominantly positive responses despite underlying controversies. Franco's study, underscores the significance of social media as a platform for gauging public opinion on global events, though he also cautions about the representativeness of such analyses [3].

Similarly, Syarafina Dewi and Dede Brahma Arianto [2] utilized TextBlob for sentiment analysis of Twitter content related to Qatar’s selection as the 2022 FIFA World Cup host. Their findings indicate a majority positive sentiment but shift depending on the period of analysis, highlighting the dynamic nature of public opinion on social media platforms. [2]

Rafiul Biswas, Sulaiman Khan, and Zubair Shah [4] focused on the human rights controversy in Qatar leading up to the 2022 FIFA World Cup. They employed both machine learning and deep learning techniques for sentiment analysis, revealing predominantly negative sentiments. This contrasts with other studies, indicating the topic-centric nature of sentiment on social media. [4]

James She, Kamilla Swart-Arries, Mohammad Belal, and Simon Wong [1] contributed by analyzing tweets related to the 2022 FIFA World Cup using the VADER algorithm. Their research offers insights into public excitement about the World Cup and popular football stars, reflecting the broader application of sentiment analysis in understanding public interests and emotions. [1]

Pelin Avcı and Gökmen Kılınçarslan's [5] study conducted sentiment analysis on Turkish language tweets, employing the MAXQDA program. Their findings of predominantly positive reviews highlight the regional variations in public sentiment and the importance of considering different languages in social media analysis. [5]

While sentiment analysis remains the cornerstone of such research, additional methods such as topic modeling, word cloud diagrams, and tabular representations of common

words add depth to the understanding of social media narratives. The temporal analysis, such as examining the number of tweets per hour, can further reveal the ebb and flow of public engagement during key event milestones. The interplay of these tools facilitates a more granular understanding of public discourse, which is a gap our current study aims to fill by incorporating a multi-dimensional approach to the analysis of Twitter data during the World Cup period.

The collective findings from these studies, while methodologically diverse, emphasize the multifaceted nature of social media sentiments and the importance of comprehensive analytical frameworks to capture the full spectrum of global public opinion.

3 Methodology

In this section, we will discuss the dataset, the sentiment model RoBERTa, BERTopic, and Latent Dirichlet Allocation (LDA) for topic modeling, and how we used the models to the dataset. We will also be talking about data preprocessing and cleaning before doing sentiment analysis and topic modeling.

3.1 Dataset

For the dataset, we got it from the website Kaggle from the user KUMARI2000. [9] They collected a total of 100,000 tweets using a Twitter tracking platform Trackmyhashtag. The dataset contains 19 different features. These tweets were collected starting on December 14, 2022, at 11:58 IST, which is 12:28 PM CST, and ending on December 15, 2022 at 8:41 AM IST, which is December 14, 2022, 9:11 PM CST. KUMEARI2000 was located in Noida, Uttar Pradesh, India so that is why the data is in India Standard Time. The timeframe for the tweet collection was during the semi-final match between France and Morocco. This knowledge will be useful when looking at the most common words used and the top topics talked about during the data visualization portion of the results.

3.2 Data Preprocessing

Before we can perform topic modeling, or trend identification, we need to clean the data. The hashtags, URLs, mentions (@) and special characters were removed from the Tweet Content and the cleaned Tweet Content was put into a new column. After we had to remove all of the sequence characters (\n, \t, \b, \r, \v), get rid of the stop words (a, an, the, and, it, for, or, but, in), and expand the contractions. [12]. Regarding the sentiment analysis, we used the original text since punctuation marks and emojis influence the sentiment analysis [18] [19].

Many of the tweets were not English, so it is important to remove all of the non-English tweets from the dataset. When doing sentiment analysis, topic modeling, and trend analysis, the data has to be in English or the models might have trouble[17]. After removing the non-English tweets this reduced the size of the dataset to 58,914 from 100,000 tweets.

After, we would have to update the time so it is in the correct format. This will make it easier to graph when visualizing the dataset. Also, we removed unused columns that were not important for our sentiment analysis, topic modeling, and trend analysis. This brought the dataset's feature count to four. These features were "Tweet ID", "Tweet Posted Time", "Tweet Content", and "Cleaned Tweet Content". "Tweet ID" is the unique ID for each tweet [10], "Tweet Posted Time" is the time that the tweet was posted, "Tweet Content" is the original content of the tweet, and "Cleaned Tweet Content" is the preprocessed content of the tweet.

3.3 RoBERTa

RoBERTa (Robustly Optimized BERT Approach) is an advanced natural language processing (NLP) model developed by Facebook AI. [11] It's based on Google's BERT (Bidirectional Encoder Representations from Transformers) model but optimized for improved performance. RoBERTa can be used for sentiment analysis on text, giving it a "Positive", "Negative" and "Neutral" label for each text. The power of RoBERTa in sentiment analysis lies in its deep understanding of language nuances, enabling it to capture complex sentiments in text effectively [11]. The RoBERTa model that we used was finetuned for sentiment analysis and works best on English text. The model was found on the Hugging Face website by Cardiff NLP [17] and the model version we used was twitter-roberta-base-sentiment-latest. As the name depicts, the AI model is pre-trained to find the sentiment for tweets.

We perform sentiment analysis using the finetuned RoBERTa model to our original Tweet content. This gives us insight into how people felt about the semi-final game between France and Morocco. The reason why we do not use the cleaned Tweet content is the emojis, hashtags, mentions, and others get to factor into the sentiment of the Tweet by RoBERTa as indicated earlier.

3.4 BERTopic

BERTopic is an algorithm that leverages advanced NLP techniques to automatically identify and group topics within large collections of text data. [14] It's built on top of BERT (Bidirectional Encoder Representations from Transformers), a powerful language model developed by Google [16]. BERTopic will be used for topic modeling on the cleaned Tweet content because it can show irrelevant topics if the content is full of URLs, mentions, and emojis. Topic modeling can help us identify trends and topics that are being discussed the most in all our tweet content.

3.5 Latent Dirichlet Allocation (LDA)

Similar to BERTopic, Latent Dirichlet Allocation (LDA) is used to discover hidden topics and patterns in a large amount of text. It is a type of statistical model used for discovering different topics within a collection of text. It's widely used in natural language processing

to organize and understand large sets of textual data. [15] LDA will be used on our cleaned Tweet content for the same reason as BERTopic. Uncleaned data can hinder the performance of LDA, making it show not relevant topics. We will see how much LDA differs from BERTopic in choosing topics that it finds in the cleaned Tweet content.

3.6 Data Visualization

After performing sentiment analysis and topic modeling on our Tweet data, we can visualize the data. This gives us a better understanding of the data and gives us insight into the data as shown in the next section.

4 Results and Discussion

After doing the sentiment analysis, topic modeling, and data visualization, we can see the results of the data analysis.

4.1 Sentiment Analysis

The sentiment analysis revealed that 55% of the tweets were positive, 32.7% neutral, and 12.3% negative, as depicted in Figure 1. This suggests a predominantly positive sentiment towards the semi-final match between France and Morocco. The data supports a generally favorable public response, reflecting an overall enthusiasm surrounding the event.

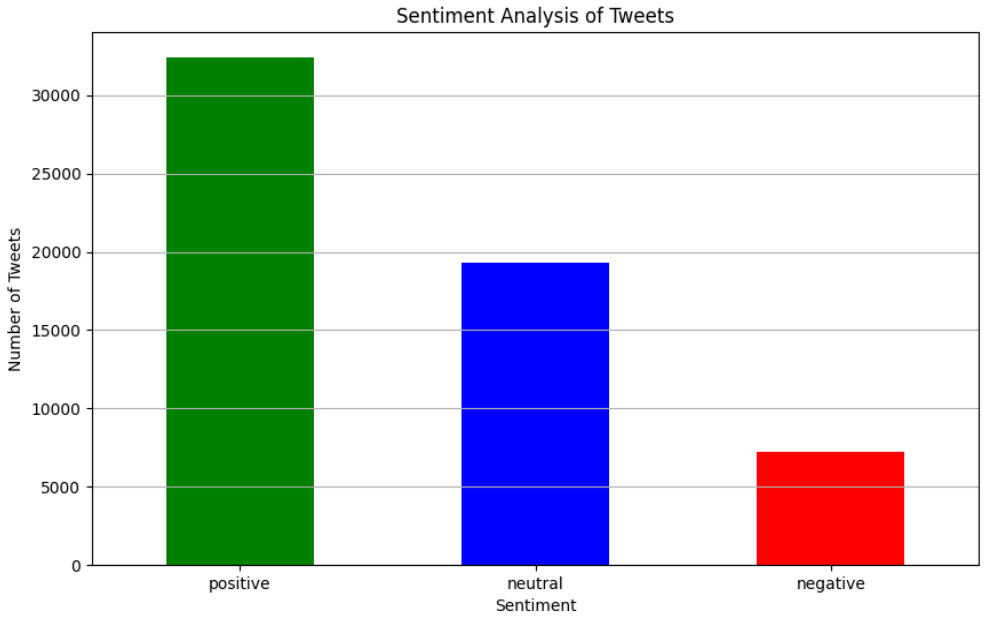

Figure 1: Sentiment distribution among the English-language tweets. Out of the evaluated tweets, 32,423 (55%) exhibited a positive sentiment, 19,283 (32.7%) were neutral, and

7,208 (12.3%) expressed negative sentiment. The sentiment classification was based on the original tweet content.

4.2 Topic Modeling with BERTopic and LDA

Next, we delve into topic modeling to uncover thematic patterns within the same corpus of tweets. The combined use of BERTopic and Latent Dirichlet Allocation (LDA) models provided a comprehensive thematic landscape, identifying dominant topics that captured both the spirit of the FIFA World Cup and ancillary discussions.

4.2.1 BERTopic Analysis

BERTopic identified a range of topics, including discussions not directly related to football, such as digital collectibles (NFTs) and online giveaways. However, it also captured topics that resonate with the World Cup's spirit, such as the performance of teams and players, and the social and cultural celebrations or tensions following the matches. The top five topics, ranked by the number of associated tweets, were as follows:

1. General discussions around Argentina's performance, indicating a global interest in the team's strategy and gameplay.

2. Conversations centered around giveaways and contests, showcasing the commercial aspect of the event's digital engagement.

3. Discussions about NFTs and digital collections related to the World Cup, revealing the intersection of sports and the burgeoning digital collectibles market.

4. A significant volume of tweets focused on the historical and fantastic journey of players like Hakim, highlighting personal stories and achievements within the larger context of the World Cup.

5. Topics related to post-match activities, including celebrations and rioting, providing insights into the social and cultural impacts of match outcomes.

4.2.2 LDA Analysis

Conversely, LDA, focusing strictly on the frequency of terms, yielded topics more directly tied to the World Cup context:

- The emergence of NFTs and digital collections, reflecting the modern intersection between sports and digital memorabilia.

- A strong sense of pride and achievement related to Morocco's performance, emphasizing the team's impact on the World Cup narrative.

- Online interactions related to giveaways, indicating a vibrant digital engagement among the fanbase.

- Anticipation and discussions surrounding the final match, highlighting the global excitement for the showdown between Messi and Mbappe.

- Social and cultural responses to match outcomes, including celebrations and riots, underscoring the deep emotional investment of fans.

4.3 Comparing BERTopic and LDA Outcomes

The juxtaposition of BERTopic and LDA results offers a nuanced understanding of public discourse during the FIFA World Cup 2022. Both methods spotlight the diverse interests of the global audience, from football-related discussions to engagement in digital and social activities.

4.3.1 Insights and Implications

- **Digital Engagement:** The prominence of topics related to NFTs and giveaways across both methods underscores the evolving landscape of fan engagement, where digital platforms offer new avenues for interaction and commemoration.

- **Cultural and Emotional Resonance:** LDA's highlighting of Morocco's pride and the social responses to match outcomes, combined with BERTopic's findings, reflect the significant emotional and cultural weight of the World Cup, transcending mere sport.

- **Anticipation for the Final:** The anticipation surrounding the final match between Argentina and France, accentuated by the prospect of Messi and Mbappe facing off, illustrates the narrative power of major sporting events, fueled by the personal stories and rivalries of star athletes.

4.3.1 Methodological Complementarity

- The differences in the topics identified by BERTopic and LDA demonstrate the *complementary nature* of these approaches. BERTopic's utilization of embeddings allows for a nuanced detection of themes, including less obvious ones like digital collectibles, while LDA's reliance on term frequency surfaces the most dominant discussions within the dataset.

- Employing both methods provides a more comprehensive view of the dataset, capturing both the depth and breadth of the public discourse.

4.3 Word Clouds Diagram & Most Common Words

We can also find out more about what was discussed during the semi-final game between France and Morocco by using word clouds and finding the most used hashtags and words. Word clouds are visual representations of text data where the size of each word indicates its frequency or importance in the text. [13] Figure 2 is a word cloud of the cleaned Tweet

content and Figure 3 is a word cloud of the original Tweet content. The most common hashtags can be shown in Table 1 and the most common words in Table 2.

Figure 2: This is a word cloud of the cleaned Tweet content. The bigger the word, the more frequent and important the word is for a large text.

Figure 3: This is a word cloud of the original Tweet content. The bigger the word, the more frequent and important the word is for a large text.

Hashtag	Occurrences
#FIFAWorldCup	44,794
#FIFA22	7,218
#Qatar2022	6,879
#FIFA	5,021
#Messi	4,659
#GOAT 🐐	4,084
#LeoMessi	3,585
#Morocco	3,115
#MoroccoVsFrance	2,606
#FIFAWorldCupQatar2022	2,239

Table 1: The most common hashtags used for all of the Tweets.

Original Tweet Content		Cleaned Tweet Content	
Word	**Occurrences**	**Word**	**Occurrences**
fifaworldcup	47,807	france	12,376
fifa	16,219	rt	11,211
france	16,052	world	10,062
morocco	14,535	morocco	8,549
rt	11,257	amp	8,405
world	10,103	final	7,741
messi	9,744	cup	7,513
qatar	9,219	nfts	7,179
amp	8,424	collection	7,137
final	8,371	like	6,394

Table 2: The most common words used in all of the Tweets. Both original Tweet content and cleaned tweet content were used.

We also found how many tweets were tweeted per hour from the start of the tweet collection to the end of the tweet collection. The number of tweets per hour was consistent where it peaked at 7,753 tweets from 4:00 PM to 5:00 PM, which can be seen in Figure 2. The 12:00 PM column is low because the Tweet collection started at 12:28 PM so there was only 28 minutes' worth of Tweets for that hour. The same thing goes from 9:00 PM, where the Tweet collection ended at 9:11 PM.

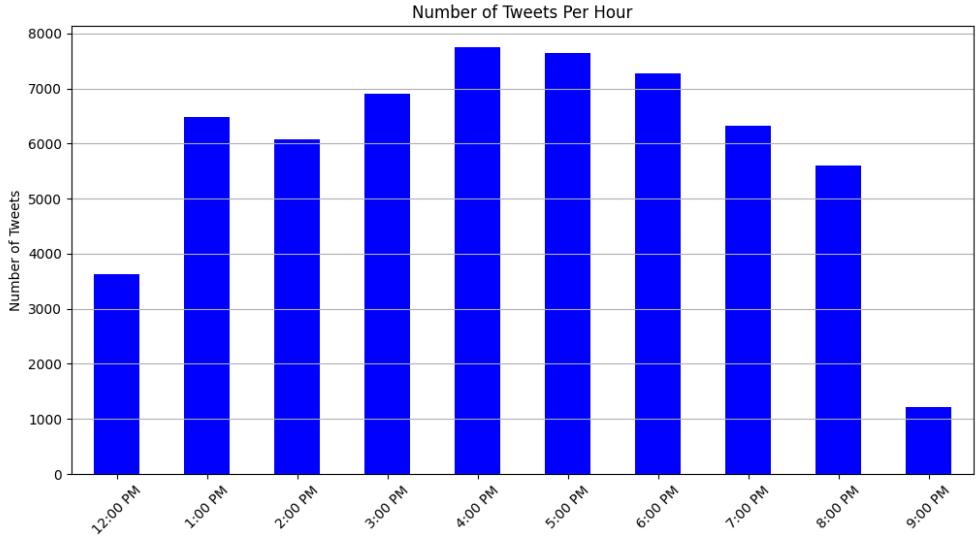

Figure 4: Number of Tweets per hour. The Tweet collection started at 12:28 PM CST and ended at 9:11 AM CST.

5 Conclusion

Our analysis of Twitter data during the 2022 FIFA World Cup has offered key insights into global sentiment and thematic discourse, showcasing a positive outlook amidst a breadth of topics. Advanced NLP models—RoBERTa for sentiment analysis and BERTopic and LDA for topic modeling—were instrumental in parsing this extensive dataset. While the overall sentiment skewed towards neutrality with a positive bias, the topic models highlighted a rich tapestry of discussions that transcended the sports arena, touching on socio-political and cultural narratives.

This research underlines the effectiveness of leveraging Deep Learning and Machine Learning to decipher complex public conversations around major events. The insights obtained point to a global audience interconnected not only by sport but by a host of shared experiences and dialogues. Future research will aim to incorporate a more extensive,

temporally diverse dataset covering the World Cup's entire span, analyzing sentiments around individual matches and incorporating multilingual analysis to capture the full spectrum of global commentary.

References

[1] J. She, K. Swart-Arries, M. Belal, and S. Wong, “What Sentiment and Fun Facts We Learnt Before FIFA World Cup Qatar 2022 Using Twitter and AI,” *arXiv preprint arXiv:2306.16049*, 2023. doi: 10.48550/arXiv.2306.16049. [Accessed Mar. 7, 2024].

[2] S. Dewi, and D. B. Arianto, “Twitter Sentiment Analysis Towards Qatar as Host of the 2022 World Cup Using TextBlob,” *Journal of Social Research*, 2023, vol. 2, no. 2, pp. 443-455, doi: 10.55324/josr.v2i2.615. [Accessed Mar. 7, 2024].

[3] L. F. Nuñez Franco, “On Sentiment Analysis of Twitter Content Related to the FIFA World CUP 2022 in Qatar,” B.S. thesis, Technische Hochschule Ingolstadt, Ingolstadt, Germany, 2022.

[4] M. R. Biswas, S. Khan and Z. Shah, "Twitter Users Discussions About Migrants Rights and Facilities During FIFA World Cup 2022," *2023 IEEE International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings)*, 2023, pp. 1-5, doi: 10.1109/AIBThings58340.2023.10292466. [Accessed: Mar. 7, 2024].

[5] P. Avcı, and G. Kılınçarslan, “THE EXCITEMENT OF FOOTBALL KICKS OFF IN QATAR: A SAMPLE OF TWITTER DATA ANALYSIS,” *The Online Journal of Recreation and Sports*, 2023, vol. 12, no. 4, pp. 678-686, doi: 10.22282/tojras.1340305. [Accessed: Mar. 7, 2024].

[6] Fédération Internationale de Football Association. “About FIFA.” *Fédération Internationale de Football Association*. [Online]. Available: https://inside.fifa.com/about-fifa. [Accessed: Mar. 24, 2024].

[7] J. W. Kim, K. Dongwoo, K. Brian, H. K. Joon, K. Suin, and O. Alice, "Social media dynamics of global co-presence during the 2014 FIFA World Cup," *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 2623-2632, doi: 10.1145/2702123.2702317. [Accessed: Mar. 29, 2024].

[8] Georgetown University Qatar. “The First World Cup in the Middle East.” *Georgetown University Qatar*. [Online]. Available: https://cirs.qatar.georgetown.edu/research/research-initiatives/building-legacy-qatar-fifa-world-cup-2022/1-first-world-cup-middle/. [Accessed: Mar. 29, 2024].

[9] KUMARI2000, Dec. 2022, "FIFA World Cup Twitter Dataset 2022," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/kumari2000/fifa-world-cup-twitter-dataset-2022. [Accessed: Jan. 15, 2024].

[10] X. "Twitter IDs." *X Development Platform*. [Online]. Available: https://developer.twitter.com/en/docs/twitter-ids. [Accessed: Mar. 29, 2024].

[11] Meta. "RoBERTa: An optimized method for pretraining self-supervised NLP systems." *Meta*. [Online]. Available: https://ai.meta.com/blog/roberta-an-optimized-method-for-pretraining-self-supervised-nlp-systems/. [Accessed: Mar. 29, 2024].

[12] A. Meddeb, and L. B. Romdhane, "Using topic modeling and word embedding for topic extraction in Twitter," *Procedia Computer Science*, 2022, 207, pp. 790-799, doi: 10.1016/j.procs.2022.09.134. [Accessed: Mar. 30, 2024].

[13] Alida. "The pros and cons of word clouds as visualizations." *Alida*. [Online]. Available: https://www.alida.com/the-alida-journal/the-pros-and-cons-of-word-clouds-as-visualizations. [Accessed: Mar. 30, 2024].

[14] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv preprint arXiv:2203.05794*, 2022, doi: 10.48550/arXiv.2203.05794. [Accessed: Mar. 19, 2024].

[15] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of machine Learning research 3*, 2023, pp. 993-1022. [Accessed: Mar. 19, 2024].

[16] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018, doi: 10.48550/arXiv.1810.04805. [Accessed: Mar. 19, 2024].

[17] Cardiff NLP. "cardiffnlp/twitter-roberta-base-sentiment-latest". *Hugging Face*. [Online]. Available: https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest. [Accessed: Mar. 9, 2024].

[18] Shiha, M., & Ayvaz, S. (2017). The effects of emoji in sentiment analysis. Int. J. Comput. Electr. Eng.(IJCEE.), 9(1), 360-369.

[19] Lou, Y., Zhang, Y., Li, F., Qian, T., & Ji, D. (2020). Emoji-based sentiment analysis using attention networks. ACM Transactions on asian and low-resource language information processing (TALLIP), 19(5), 1-13.