REDDIT SENTIMENT ANALYSIS: EXPLORING THE FOOTBALL WORLD THROUGH COMMENTS

ABSTRACT

In this research paper, we embark on an exploration of sentiment analysis and topic modeling within the domain of online discussions centered around two titans of football: Lionel Messi and Cristiano Ronaldo. Our study harnesses a range of data analysis techniques, including sentiment analysis, Latent Dirichlet Allocation (LDA), TF-IDF feature extraction, and named entity recognition (NER), to uncover the sentiments conveyed and the underlying themes prevalent in comments across various online platforms.

By delving into sentiments expressed and themes discussed within comments related to Lionel Messi and Cristiano Ronaldo on Reddit, we aim to shed light on the overall sentiment towards these football icons, pinpoint key topics within the discourse, and offer a quantitative understanding of the vocabulary and sentiment distributions. Through the lens of natural language processing (NLP) techniques such as sentiment analysis, named entity recognition (NER), and topic modeling, our study endeavors to provide a comprehensive analysis of the sentiments expressed and topics discussed by Reddit users regarding Messi and Ronaldo. Furthermore, we utilize feature extraction methods like TF-IDF to gauge the significance of words and phrases in these discussions.

INTRODUCTION

Lionel Messi and Cristiano Ronaldo, towering figures in the football world, inspire profound admiration and ignite passionate debates among fans worldwide. Understanding the sentiments expressed and topics discussed in online conversations about these football icons is essential for comprehending their enduring impact and the intricacies of fan interaction in the digital era.

In this research endeavor, we embark on a data-driven exploration of sentiment and discourse in online discussions revolving around Messi and Ronaldo. Football, renowned as one of the most beloved sports globally, elicits intense emotions and fervent exchanges among enthusiasts. Within this vast football landscape, Lionel Messi and Cristiano Ronaldo emerge as unparalleled luminaries, each commanding a devoted following and leaving an indelible mark on the sport's history.

By delving into the sentiments and conversations surrounding these legendary figures, we aim to glean invaluable insights into fan perceptions, rivalries, and the broader tapestry of football culture.

METHODOLOGY

Our methodology involves several key steps:

1.DATA COLLECTION: Employing the Python Reddit API Wrapper (PRAW) library, we systematically retrieve comments and post titles associated with Lionel Messi and Cristiano Ronaldo from their dedicated subreddits on Reddit. This data acquisition process entails extracting content from discussions relevant to these football legends, with a specific emphasis on posts gaining traction and visibility, indicative of recent and trending conversations within the online community.

By targeting rising posts within Messi and Ronaldo subreddits, we ensure the inclusion of dynamic and current discussions, enabling us to capture the pulse of ongoing discourse surrounding these iconic players. This approach allows for a comprehensive examination of the sentiments expressed and topics discussed by Reddit users, providing valuable insights into the evolving perceptions and interests surrounding Messi and Ronaldo within the digital sphere. Following the web scraping phase, we undertook text preprocessing to enhance the effectiveness of our analysis. To streamline this process, we developed a preprocessing function encompassing several key techniques:

1. Lemmatization: Lemmatization is a linguistic process that involves reducing words to their base or root form, known as the lemma. This helps in standardizing the vocabulary by treating different forms of the same word as identical. For instance, "running" and "ran" would both be lemmatized to "run".

2. Tokenization: Tokenization involves breaking down the text into smaller units, typically words or phrases referred to as tokens. This facilitates the parsing and analysis of the text at a granular level, enabling subsequent processing steps such as sentiment analysis and topic modeling to operate more efficiently.

3. Regular Expression (Regex): Regular expressions are sequences of characters used to define search patterns. In the context of text preprocessing, regex can be employed to identify and manipulate specific patterns within the text, such as removing punctuation marks or filtering out unwanted characters.

4. Converting Letters to Lower Case: Converting all letters to lower case helps in standardizing the text by ensuring uniformity in word representation. This is important for tasks like sentiment analysis and topic modeling, where case sensitivity may lead to inconsistencies in the analysis results.

By incorporating these preprocessing techniques into our workflow, we aim to optimize the quality and consistency of the textual data, thereby improving the accuracy and reliability of subsequent analysis tasks.

2.SENTIMENT ANALYSIS: We conduct sentiment analysis on the comments utilizing two distinct tools: the VADER sentiment analyzer and the Text Blob library. These tools offer quantitative assessments of the sentiment expressed within the comments, allowing us to gauge the prevailing emotional tone towards Lionel Messi and Cristiano Ronaldo.

1. VADER Sentiment Analyzer: VADER (Valence Aware Dictionary and sentiment Reasoner) is a lexicon and rule-based sentiment analysis tool specifically designed for analyzing sentiments in text. It quantifies the sentiment of a piece of text by assigning polarity scores (positive, negative, or neutral) to individual words and then aggregating these scores to derive an overall sentiment score for the entire text.

2. Text Blob Library: Text Blob is a Python library that provides simple API methods for processing textual data, including sentiment analysis. It employs a pattern-based approach to assess the sentiment of text, where it calculates the polarity of the text (positive, negative, or neutral) based on the occurrence of certain keywords and phrases associated with each sentiment category.

By leveraging both the VADER sentiment analyzer and the Text Blob library, we obtain a comprehensive understanding of the sentiment landscape surrounding Messi and Ronaldo, enabling us to quantify and analyze the overall sentiment expressed in the comments with greater accuracy and depth.

3. TOPIC MODELING: Utilizing Latent Dirichlet Allocation (LDA), we uncover latent topics inherent within the comments, thereby elucidating the predominant themes deliberated by online users. LDA is a probabilistic model commonly employed for topic modeling tasks, aiming to discern underlying themes or topics within a collection of documents. By applying LDA to the comment’s dataset, we unveil the underlying structures of discourse, enabling us to identify and analyze the key subjects of discussion prevalent among online users.

4. TF-IDF FEATURE EXTRACTION: TF-IDF (term frequency- inverse document frequency) is a numerical statistic used in natural language processing and information retrieval to evaluate the importance of a word in document relative to a collection of documents.it operates on the premises that words that appear frequently within a document but infrequently across the entire document corpus are likely to be more significant in conveying documents meaning.

5. NAMED ENTITY RECOGNITION (NER): Named Entity Recognition (NER) is a natural language processing (NLP) technique used to identify and classify named entities within text. Named entities are specific objects, people, locations, organizations, dates, numerical expressions, and other named entities that hold significance within a given context.

NER involves the identification of these named entities and the categorization of each entity into predefined classes such as person names, organization names, location names, etc.

6.

RESULTS:

Our analysis yields the following key findings:

1.SENTIMENT ANALYSIS: The sentiment analysis of comments related to Lionel Messi and Cristiano Ronaldo shows an overall positive sentiment for both players, with Messi receiving a higher total sentiment score compared to Ronaldo. The average sentiment score for Messi is 0.333, indicating predominantly positive sentiment, while for Ronaldo it is 0.110, also reflecting a positive sentiment but to a lesser extent. Text Blob sentiment analysis further supports this trend, with Messi having a slightly higher average sentiment score compared to Ronaldo.  
Here are some visual results of sentiment analysis:

PLOT OF SENTIMENT SCORE DISTRIBUTIONS FOR MESSI AND RONALDO

A group of graphs showing different sizes of bars

Description automatically generated with medium confidence

DISTRIBUTION VISUALIZATION

A graph of different sizes and colors

Description automatically generated with medium confidence

A person in a shape of a person

Description automatically generatedA person made out of words

Description automatically generatedWORD CLOUD FOR MESSI WORD CLOUD FOR RONALDO

VISUAL REPRESENTATION FOR TEXT BLOB

A graph of a number of different sizes

Description automatically generated with medium confidence

BARPLOT FOR SENTIMENT SCORES USING TEXTBLOB

A diagram of a football game

Description automatically generated

2.TOPIC MODELING: The topic modeling analysis reveals several themes prevalent in the comments related to Messi and Ronaldo. Topics include discussions about fan rivalry ("ronaldo fan," "pessi," "goat"), participation guidelines on the subreddit ("subreddit," "karma," "moderator"), comparisons between the players ("messi," "ronaldo," "penaldo"), discussions about their performance ("world cup," "goal," "season"), and debates about their status as the greatest of all time ("goat," "best," "true"). These topics reflect the diverse range of opinion and discussions surrounding Messi and Ronaldo in online communities.

A chart of different colored lines

Description automatically generated with medium confidence

WORD CLOUD FOR TOPIC MODELING

A close-up of words

Description automatically generated

3.TF-IDF FEATURE EXTRACTION: Feature extraction techniques such as named entity recognition and vocabulary analysis provide insights into the entities mentioned in comments and the vocabulary used by fans. In Messi comments, entities such as "GPE" (Geopolitical Entity), "PERSON," and "CARDINAL" are frequently mentioned, indicating discussions about places, individuals, and numerical counts. Ronaldo comments also mention similar entities, with a focus on "PERSON," "CARDINAL," and "GPE." The vocabulary analysis reveals common words associated with each topic, such as "ronaldo," "messi," "goal," and "fan," indicating the central themes of discussions. TF-IDF analysis reveals the vocabulary size and highlights the unique terms and phrases used in discussions about Messi and Ronaldo.

4.NAMED ENTITY RECOGNITION:

The named entity recognition (NER) analysis provides insights into the entities mentioned in comments related to Messi and Ronaldo. In Messi comments, entities such as geopolitical entities (GPE), persons (PERSON), and numerical counts (CARDINAL) are frequently referenced. This indicates discussions about locations, individuals, and statistical figures, reflecting the diverse range of topics associated with Messi, including his international performances, personal achievements, and statistical records.

Similarly, in Ronaldo comments, entities such as persons (PERSON), numerical counts (CARDINAL), and geopolitical entities (GPE) are prevalent. This suggests discussions about Ronaldo's personal attributes, statistical achievements, and international engagements, mirroring the multifaceted nature of conversations surrounding Ronaldo's career.

The NER analysis underscores the prominence of Messi and Ronaldo in discussions, with fans often referencing their individual performances, achievements, and interactions with various geopolitical entities such as clubs, national teams, and competitions. Overall, NER analysis provides valuable insights into the key entities driving discussions and narratives surrounding Messi and Ronaldo in online communities.

A graph with blue and orange bars

Description automatically generated

PLOT OF MESSI AND RONALDO'S GOAL CONTRIBUTIONS (2004-2023)

A graph with lines and dots

Description automatically generated

TOTAL GOALS SCORED BY MESSI AND RONALDO

A graph of a graph with numbers and a number of years

Description automatically generated with medium confidence

DISCUSSION

The positive sentiment towards Messi and Ronaldo reflects their enduring popularity and influence in the football community. The diverse range of topics discussed underscores the multifaceted nature of fan engagement, encompassing player performance, team dynamics, and personal narratives. The TF-IDF analysis elucidates the richness of vocabulary and the distinctiveness of discussions surrounding these football icons. Additionally, NER highlights the significant entities shaping the discourse around Messi and Ronaldo.

CONCLUSION

In conclusion, our research offers valuable insights into online discussions about Lionel Messi and Cristiano Ronaldo. By leveraging sentiment analysis, topic modeling, TF-IDF feature extraction, and named entity recognition, we gain a comprehensive understanding of the sentiments expressed and the topics discussed in relation to these football legends. Our findings contribute to a deeper appreciation of the dynamics of fan engagement and the enduring impact of Messi and Ronaldo in the digital landscape of football discourse.