University of Warsaw
Faculty of Economic Sciences

Jakub Bandurski Nr albumu: 417911

Adam Janczyszyn Nr albumu: 419350

World Cup 2022 in various media outlets

Final project
Topic Modelling 2022Z

Warsaw, February 2023

Abstract

This project aims to compare the coverage of the 2022 World Cup by the Qatari government media and the global media giant, CNN. We sought to examine the differences in the presence of non-sports related topics in relation to the World Cup. Methods used in this analysis cover topic modelling techniques, LDA model and a novel out-of-the-box Large Language Model BERT. The results support the predictions and show significant differences in the topics covered by the two media outlets. The Qatari media overlooked crucial topics such as human rights and Iranian protests, while CNN published extensive information on these sensitive issues. We want to remind about importance of human rights in future global events.

Key words

topic modelling, LDA, BERT, Qatar, World Cup 2022, human rights, media

Contents

1 Introduction

Motivation

- The main motivation behind the work was the desire to compare how the World Cup was perceived by the Qatari government media and the recognised global media, CNN.
- We also wanted to check the difference in the occurrence of non-sports topics in relation to the 2022 World Cup.
- Lastly, as football fans, we wanted to check and summarise the key events of the World Cup in Qatar.

2 Prerequisites

Data Overview

News articles were scrapped, which related to the World Cup 2022 events. For western recognised global media we've chosen CNN cnn.com. The time span ranges from April 2022 to January 2023 which includes information from way before the World Cup and several events after the finals. The total number of articles is 370.
For qatari-aligned media we've chosen the official World Cup 2022 media outlet qatar2022.qa/en/news. The time span ranges from April 2022 to January 2023 which includes information from way before the World Cup and several events after the finals. The total number of articles is 157,

Preprocessing

We used the standard process of preprocessing used in NLP methods. Firstly, we cleaned data from any missing values etc. Secondly, we tokenized all the articles, lemmatization is not necessary as we are working with English data. Next, we stemmed the whole corpus. After that issue of removing stopwords has to be considered. Apart from the regular stop word we removed the following widespread words: world, cup, FIFA, has, was, were etc. We used 1-grams, 2-grams and 3-grams to incorporate n-grams in our considerations. In order to specify an approach to the available corpora we used both TF and TF-IDF.

3 Workflow

Our analysis is spread out between many python notebooks and folders. This was a deliberate choice which made the cooperation easier and we hope that it is more readable for you. First of all we divided all files into two folders cnn and qatar which include files relating only to those media outlets. Next inside each folder, we have a separate folder for data and scraping notebook. For each dataset, we created separate data preprocessing notebook which handles tokenization, stopwords etc. Both TF and TF-IDF approaches were used. Finally, the main analysis was performed in two notebooks one with the implementation of LDA and the other with the implementation of BERT.

4 CNN

All of the optimization steps were discussed during the presentation and the rest can be found in the notebooks. What's left is a clear interpretation of the topics generated by both algorithms and contextualising with ongoing events.

LDA - Topics interpretation

Topic 3: ['fifa', 'ronaldo', 'rights', 'human rights', 'ten hag', 'match', 'saudi', 'host', 'commit', 'does get']
- At the start of 2023, Cristiano Ronaldo was the hero of a high-profile transfer to Saudi Arabia's Al-Nassr club. With the World Cup in Qatar taking place recently beforehand and all the events associated with it, the subject of human rights continued to stay on the lips of sports fans and beyond. Amnesty International asked Cristiano Ronaldo to 'draw attention to human rights issues' in Saudi Arabia. Hence this link between the Portuguese football star, the subject of human rights, Saudi Arabia and the Qatar World Cup.
Topic 4: ['workers', 'rights', 'migrant', 'migrant workers', 'human rights', 'human', 'tournament', 'fifa', 'qatars', 'qatari']
- The appearance of the topic of human rights as part of the Qatar World Cup story was expected in our analysis. The numerous deaths of foreign workers in the construction of the stadiums, the inhumane working hours and the numerous health and safety failures meant that the World Cup in Qatar would not only be associated with a sporting event, but also with a crowning example of human rights violations.

Topic 9: ['todays match', 'iran', 'human rights', 'iranian', 'iranian football', 'cavallo', 'us armband', 'news', 'portugal', 'shock grief']
- During the ongoing World Cup in Qatar, people continued demonstrations in Iran over the death of Mahsa Amini in September 2022. Mahsa Amini fell into a coma and died after being arrested in Tehran by the morality police for allegedly violating hijab laws. Fans show support for 'Women, Life, Freedom' Iran human rights protests at Qatar World Cup.
Topic 18: ['jakob jensen', 'happy', 'danilo', 'qatar also', 'wage', 'played', 'advocates', 'football governing body', 'earlier year', 'return home']
- The above topic most likely refers to FIFE's banning of the Danish men's national soccer team's ability to train in jerseys bearing inscriptions promoting human rights during the then upcoming World Cup in Qatar. Jakob Jensen is The Danish Football Federation's (DBU) CEO. The DBU planned for the jerseys to read "Human Rights for All".
Topic 19: ['appearing shadow domestic', 'took control', 'agency tass', 'xherdan shaqiri', 'vladimir putin', 'wanted changeofficial, evening, recently told, lifting trophy, us soccer federation']
- In the 1/8th match of the World Cup, the team of Switzerland and Serbia faced each other. This soccer clash is also a political clash, which is revived when these teams play each other. The Serbs and the Swiss faced each other in a group during the 2018 World Cup in Russia. After the goals for the Helvetians, Granit Xhaka and Xherdan Shaqiri showed a two-headed eagle - the symbol of Albania. Both players have Albanian roots. And ethnic tensions between Serbs and Albanians have been ongoing for many years. The salt of the earth is the issue of Kosovo, a region that Serbs consider an integral part of the country. The snag is that the Kosovars declared independence in 2008. To this day, many countries do not recognize the country's independence. Among them is Russia, which indirectly explains the positive attitude of Serbs toward the country. Hence probably the combination of these names 'xherdan shaqiri', 'vladimir putin' and the subject of the World Cup in Qatar.

BERT - Topics interpretation

Topic 1: ['qatar', 'said', 'rights', 'tournament', 'fifa', 'cnn', 'workers', 'people', 'human', 'human rights']
- What pleases us is that in the case of analysis with BERT, we again received a human

rights topic.
Topic 2: ['said', 'russian', 'ukraine', 'ukrainian', 'russia', 'us', 'city', 'forces', 'military', 'war']
- In our analysis of CNN articles, we also expected some coverage of the most important and yet saddest topic of the past year. Even in the context of arithmetic on the soccer event, the topic of Russia's military attack on Ukraine could not be missed. The connections could be several. The first and probably the most likely is the banning of the Russian Federation from participation in almost all world events including sports. Hence, by FIFA's decision, Russia's national team was excluded from participating in the World Cup.
Topic 3: ['iran', 'iranian', 'irans', 'team', 'said', 'protests', 'national', 'us', 'players', 'people']
- Again, the topic referring to the problem of human rights in Iran and the demonstrations of fans during the World Cup in Qatar.

5 Qatar

All of the optimization steps were discussed during the presentation and the rest can be found in the notebooks. What's left is a clear interpretation of the topics generated by both algorithms and contextualising with ongoing events.

LDA - Topics interpretation

Topic 3: ['workers', 'rights', 'security', 'continued', 'ensure', 'government', 'contractors', 'human', 'dedicated', 'testament']
- We expected that the topics obtained from Qatari government media articles would strongly deviate from Western ones. The first topic looks like the Qatari media reassuring readers that all human rights during the preparations for and during the championship have been ensured. We see words like government, ensured, security, workers. We guess that this is the narrative that accompanied the articles on qatar2022.qa.
Topic 12: ['during year^a', 'ali stadium', 're', 'welcomed', 'multiple', 'mega event', 'historic', 'middle east', 'the art stadiums', 'it']
- Another topic that gives the overtones of great success. The first mega event in the Middle East. Everyone is invited and welcome. You can see that the Qataris were/are proud of the preparations made and the stadiums built as art pieces. What's more, they define this World Cup as historic.

Topic 18: ['pm', 'fan', 'transport', 'bus', 'festival^a', 'fan festival^a', 'metro', 'doha', 'corniche', 'public']
- The third theme also remains in a similar vein. Attention is drawn to the infrastructure that Qatar has prepared specifically for the World Cup. Built a multi-line modern metro, providing free transportation and hundreds of buses. In a word, perfectly connected capital Doha especially for visiting soccer fans. In addition, planned festivals and the Corniche, a seaside promenade stretching seven kilometers along the Gulf of Doha. The conclusion is one, the Qataris are proud of the result of their preparations and are eager to boast about it. There is no denying the great scale of the preparations and development of the Qatari capital, however, at a huge cost.

BERT - Topics interpretation

Topic 0: ['fans', 'first', 'said', 'arab', 'players', 'cup', 'middle', 'region', 'middle east', 'east']
- Again, recurring topics of what's in the LDA. First such major event in the Middle East. A big deal for the entire Arab region.
Topic 1: ['generation', 'amazing', 'legacy', 'generation amazing', 'volunteer', 'sc', 'people', 'work', 'training', 'volunteers']
- Volunteers to support the course of the championship in Qatar appear in topic 1. The government has invited 20000 volunteers from around the world to participate and help organize the championship. Of course the overtones are positive, the words amazing, legacy, generation amazing etc. appear.
Topic 2: ['stadium', 'group', 'lusail', 'fans', 'venue', 'city', 'matches', 'december', 'during', 'metro']
- And so again the topic referring to the massive infrastructure built specifically for the FIFA World Cup. One of the largest stadiums Lusail or the previously mentioned subway.

6 Conclusions

The analysis was very time-consuming but the results are aligned with our predictions. We can clearly see differences in topics presented by West media and Qatar media. Topics such as human rights or Iranian protests were completely omitted in the Qatari media. At the same time, CNN published lots of information that focused on these sensitive yet very

important topics. Overall the World Cup itself was a holiday of football as always, but many fans disregarded conditions in which people who build this event had to live. We should make sure that human rights topics are not neglected in favour of pure image investment of the Middle East.