Text mining of Twitter data for Qatar World Cup

Text Mining and Search Project

Bruni Lorenzo, 886721
Farallo Simone, 889719
Puccinelli Niccolò, 881395

January 19, 2023

Abstract

In this research we performed two different text mining tasks on a novel dataset: a collection of approximately 100000 tweets concerning the 2022 World Cup held in Qatar. For the first task, we executed a cluster analysis by grouping different tweets into a discrete number of clusters with similar properties. A summary was also extracted from these clusters using the BART model. For the second task, we performed text summarization on tweets united by the main hashtags. Specifically, we focused on hashtags regarding the event itself and the 4 semifinalists of the tournament. Finally, we evaluated the results, also considering the limitations due to the inherent complexity of the dataset.

Contents

1 Introduction

The football World Cup held in Qatar in 2022 was one of the most incredible and controversial that has ever been played. The tournament elicited many reactions from spectators, both on the sports side and the organizational side. In particular, the scandals due to the location chosen and the deaths of several workers during the construction of the stadiums

have been the subject of much controversy. Nevertheless, the extraordinary matches held, along with World Cup stars such as Lionel Messi and Kylian Mbappé, partially overshadowed the criticism.

In this work, we attempted to analyze a set of about 100.000 tweets with a twofold purpose: to group the tweets into a discrete number of clusters with similar properties and to generate summaries regarding the main hashtags.

First of all, we extracted tweets from Twitter by scraping (section 2). After applying some specific preprocessing techniques in order to clean up the tweets (section 3), we superficially explored (section 4) the distribution of the data. Then, in sections 5 and 6, we describe the methodology and application of clustering and summarization techniques. Finally, section 7 contains the final thoughts about our work and section 8 includes several suggestions for future research.

The computing environment on which the software was developed is Google Colab Pro, with 16 GB of RAM.

2 Data acquisition

The data were obtained from Twitter by scraping, via the *snscrape.modules.twitter* [1] library. The query we ran included all English tweets with a minimum of 5 likes up to December 19, 2022. We scraped the tweets on the basis of the most popular hashtags. The search has been truncated immediately after the final match because otherwise there would have been too many tweets about Argentina, winner of the tournament. Furthermore, our goal is to conduct an analysis on the tweets posted during and before the tournament.

Each tweet was entered into a dataframe, which included tweet text, user name, number of likes, number of retweets and timestamp. The dataset thus composed contained 151012 tweets.

Moreover, we saved another dataset containing the most popular hashtags and their occurrences among the tweets. This was helpful in deciding which hashtags to focus our research on.

3 Preprocessing

Tweets are difficult textual data to deal with, due to numerous grammatical errors, overuse of emojis and the large presence of insignificant texts. Tweets as they are cannot even be explored without first performing a preprocessing step. For all these reasons, we executed the following operations, in order:

1. We extracted hashtags and mentions and put them in two separate columns.
2. Extra-spaces removal.
3. Case folding.
4. Repeated characters removal.
5. We checked the word *stream* in order to remove those tweets that provide the streaming of the matches.
6. We kept only those tweets with more than 10 likes. These are probably the most informative, and thus we reduced the computational complexity.
7. We removed only those symbols which are not used in natural language (i.e. we kept punctuations).
8. Even if we specified english language at query time, we performed another control through the *langdetect* [2] library.
9. We cleaned our tweets by means of the *tweet-preprocessor* [3] library. Specifically, we set the function to remove URLs, mentions and hashtags.
10. We removed tweets with less than 10 words, usually less informative.
11. Duplicated tweets removal.

In the end, our dataset was composed by 89261 tweets. At this stage, we did not remove punctuation and emoji, nor did we apply tokenization or lemmatization. In fact, tweets were kept in natural language in order to be suitable for natural language processing.

However, for those embeddings such as TF-IDF or Doc2Vec [4], for which more ad-

vanced preprocessing is needed, additional steps were applied:

1. Demojizing. Emoji have been transformed into the respective text, using '_' as delimiter, through the *emoji* ([5]) library.
2. Numbers removal.
3. Contractions were fixed through the *contractions* [6] library (e.g. from "you're" to "you are").
4. Punctuation removal. However, we kept the '_' symbol because it's the delimiter for the emoji.
5. Tokenization. We used the *nltk TweetTokenizer*, which is specific for tweets.
6. Lemmatization. We decided to apply lemmatization instead of stemming for a more meaningful representation of tweets, which by themselves are difficult to treat due to their high level of insignificance.
7. Stop words removal.

4 Data exploration

At this stage we explored the dataset through wordclouds and temporal distributions. We report some basic features, three wordclouds, regarding hashtags (Fig. 1), mentions (Fig. 2) and words (Fig. 3), and a time distribution plot (Fig. 4).

- Total number of tweets: 89261.
- Total number of sentences: 180170.
- Average number of sentences for each tweet: 2.018.
- Total number of words: 2256636.
- Average number of words for each tweet: 25.281.
- Total number of unique words: 105242.
- Ratio between unique words and words: 0.047.

These statistics are in line with the type of textual data analyzed. In addition, the number of unique words is very low since this is a set of tweets about a rather polarized event. We might uncover some insights when we select subsets of the dataset at the summarization phase.

Figure 1: *Hashtag-cloud*

Hashtags In addition to the obvious presence of more generic hashtags such as #Qatar2022 and #FIFAWorldCup, we note #Argentina and #Messi (winners of the competition), #Morocco (the tournament's revelation team) and #BoycottQatar (this World Cup was one of the most contested ever).

Figure 2: *Mention-cloud*

Mentions The most mentioned users are journalists, football players and TV channels. Specifically, the football players most cited are Neymar, Cristiano Ronaldo, Lionel Messi and Kylian Mbappe, the 4-stars of this tournament. Moreover, we can notice the particular presence of @GhanaBlackstars (Ghana almost qualified to the round of 16 after a spectacular match against South Corea) and @Socceroos (the Australian team passed the first turn with the same points as France).

Figure 3: *Word-cloud*

Words Leaving aside the more common words such as world, cup or qatar, we can see many words regarding the winner team (Argentina) and the African teams, another sign of their extraordinary performances in this World Cup.

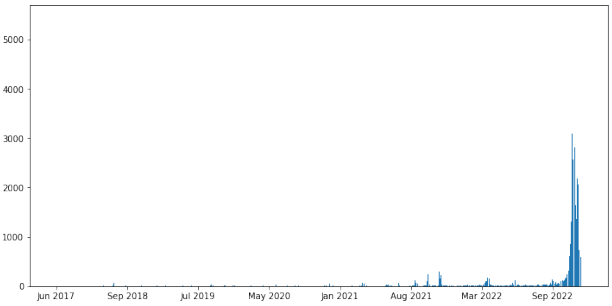

Figure 4: *Time distribution*

Temporal distribution The number of tweets is obviously higher in the period close to the World Cup. However, we notice some peaks on March 2022 and September 2021. Let's zoom in with the figures 5 and 6.

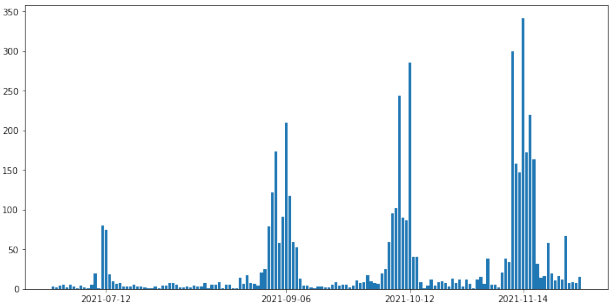

Figure 5: *Time distribution - Sep 2021*

These multiple spikes are simply due to the different tournament qualifying matches played at that time.

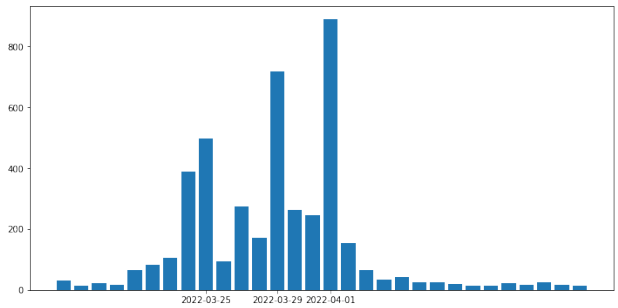

Figure 6: *Time distribution - Mar 2022*

This chart is far more interesting. In fact, in March 2022, Italy was suffering a historic elimination from the World Cup (the second in a row after 2018), which prompted multiple reactions on the web. In addition, on April 1, 2022, the teams were drawn along with their respective groupings.

5 Text clustering

After the phase of preprocessing and exploration, a text clustering analysis has been conducted in order to find groups of tweets that share similar properties. We implemented two different kind of clustering approaches: K-means and hierarchical clustering. Firstly, before applying the clustering algorithms, since we are dealing with text data, we used 3 different representation techniques: TF-IDF, Doc2Vec [4] and BERTweet [7].

For each approach, we evaluated the results through the silhouette coefficient (computed with both the standard measure, i.e. the euclidean distance, and the cosine distance) and a summary for each cluster, obtained with the BART pre-trained model [8].

The differences between the techniques and the results obtained are shown in the following subsections.

5.1 TF-IDF

As an initial approach we used the TF-IDF measure that takes into account not only the term frequency, but also the inverse document frequency. In the function used to obtain the TF-IDF matrix we took uni-grams, bi-grams

and 3-grams. Moreover, we specified a minimum and a maximum frequency for the words, respectively 10 words and 80% frequency, also in order to reduce the computational cost. In the end, we obtained a matrix with 89261 rows and 23750 features.

K-Means First of all, since in the k-means algorithm we have to specify the number of clusters as a model parameter, we applied the elbow method.

The result obtained, as shown in fig. 7, did not reveal a clear solution for this kind of representation. Nevertheless, we chose a number of cluster equal to 40, in order to possibly reveal some interesting clusters.

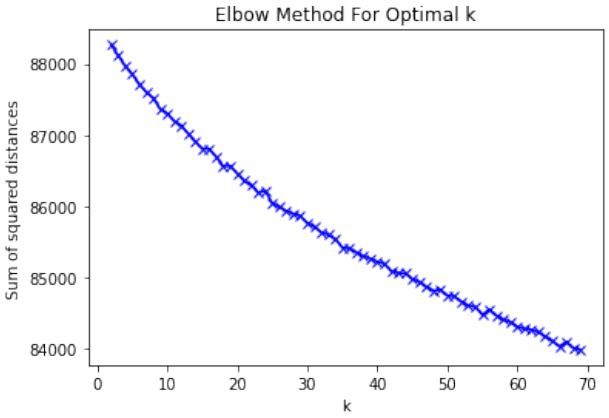

Figure 7: *Elbow Method for TF-IDF representation*

The main well-known drawback of TF-IDF approach is that it does not take into account the context of the words in a document. For this reason, the value of the standard silhouette coefficient calculated with the euclidean distance and the one computed by means of the cosine similarity were very low, respectively equal to 0.0093 and 0.0176.

Hierarchical clustering A completely different clustering approach was performed using the same representation technique. Due to limited computational resources, we couldn't perform hierarchical clustering among all the tweets. So, we chose a subset containing the 20000 most liked tweets, obtaining a matrix with 20000 rows and 6090 features.

Afterwards, we computed the cosine distance and then we performed Ward's linkage. This linkage function specifies the distance between two clusters as the increase in the "error sum of squares" (ESS) after fusing two clusters into a single cluster. Ward´s method seeks to choose the successive clustering steps so as to minimize the increase in ESS at each step.

The dendrogram obtained, shown in fig. 8, suggests that a good point to cut the tree might be the one relative to 11 clusters.

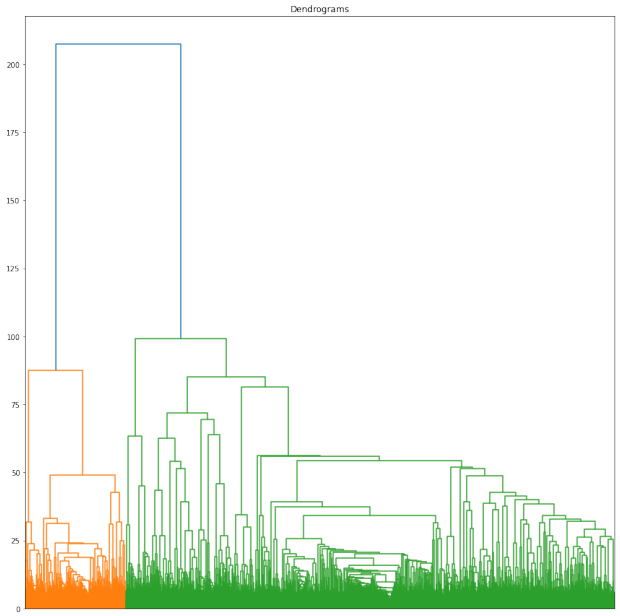

Figure 8: *Dendrogram for TF-IDF representation*

In this case the silhouette coefficients obtained were equal to 0.0039 and 0.0070.

5.2 Doc2Vec

Another embedding approach to improve the previous results was carried out. We chose the Doc2Vec embedding, that is an extension of Word2vec which encodes entire documents as opposed to individual words. In order to initialize the Doc2Vec function, we performed a grid search to find the best combination of the model parameters, i.e. able to minimize the inertia (within-cluster sum-of-squares criterion, a measure of how internally coherent clusters are) for k-means. In the end, we

found that the best combination of parameters was:

- *dm* = 0 (distributed bag of words is employed instead of distributed memory).
- *vector size* = 256 (it specifies the dimensionality of the feature vectors).
- *hs* = 0 (i.e. the model is trained with negative sampling).

K-Means As for the TF-IDF embedding technique, we used the elbow method in order to find the optimal number of clusters to initialize the k-means function.

The chart that we derived is far easier to interpret than the previous one, as we can see in fig. 9, and the optimal number of clusters chosen is 4.

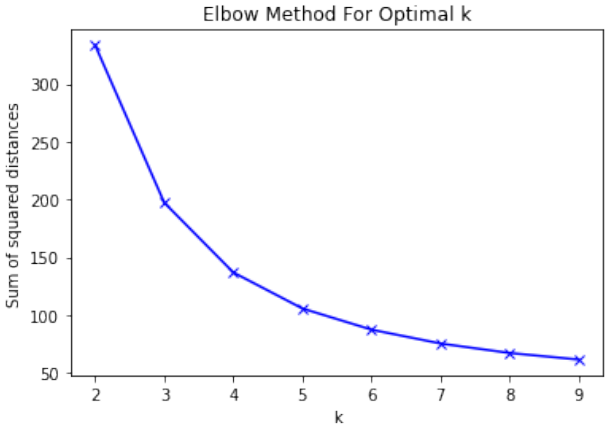

Figure 9: *Elbow method for Doc2Vec embedding*

The silhouette score based on the euclidean distance was equal to 0.4097, much higher than the TF-IDF representation. However, the silhouette computed through cosine similarity was very low, even lower than 0 (-0.0841), probably because of the k-means algorithm, which seeks to minimize the euclidean distance.

Hierarchical clustering In the hierarchical clustering we chose a subset of the tweets, for the same reasons explained before. The dendrogram obtained is shown in fig. 10.

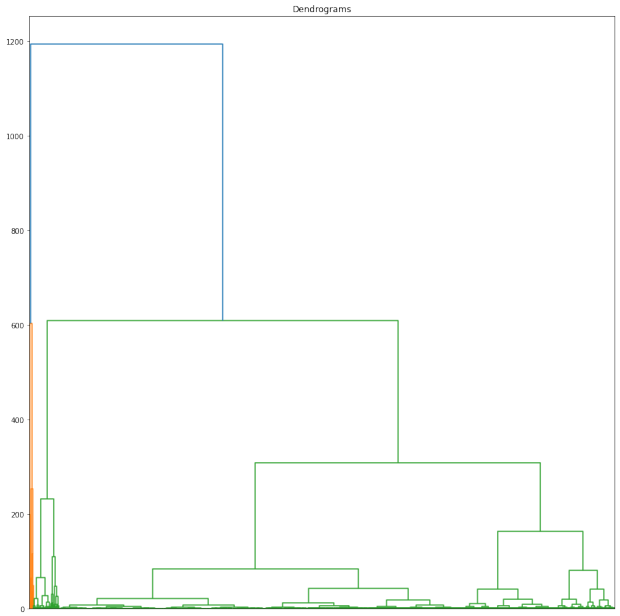

Figure 10: *Dendrogram for Doc2Vec embedding*

In this case we found a good point in which to cut the dendrogram in correspondence to 6 clusters. With respect to TF-IDF approach, we achieved better results in terms of both the silhouette coefficients (0.2019 and 0.3115). Moreover, the score based on the cosine similarity was much better than the one computed after the k-means algorithm, as opposed to the standard score (i.e. the euclidean one).

5.3 BERTweet

The last model used was BERTweet, a sentence-transformer specifically pre-trained on tweets, which embeds each document in a vector composed of 768 features. So, we obtained a matrix with 89261 rows and 768 columns.

K-Means As for the other approaches, we used the elbow method to find the optimal number of clusters and we chose 5 as the reasonable optimal number of clusters (Fig. 11. However, it is important to note that, compared to the previous approach, this graph has less interpretability.

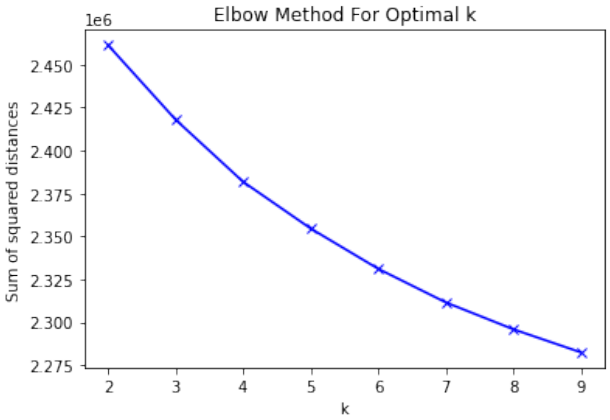

Figure 11: *Elbow method for BERTweet embedding*

In this case, the silhouette scores obtained from the model were very low, respectively equal to 0.0229 for the standard measure and 0.0374 for the cosine distance.

Hierarchical clustering We used only a subset of the data as before. The corresponding dendrogram is shown in figure 12, from which we decided to use a number of clusters equal to 12.

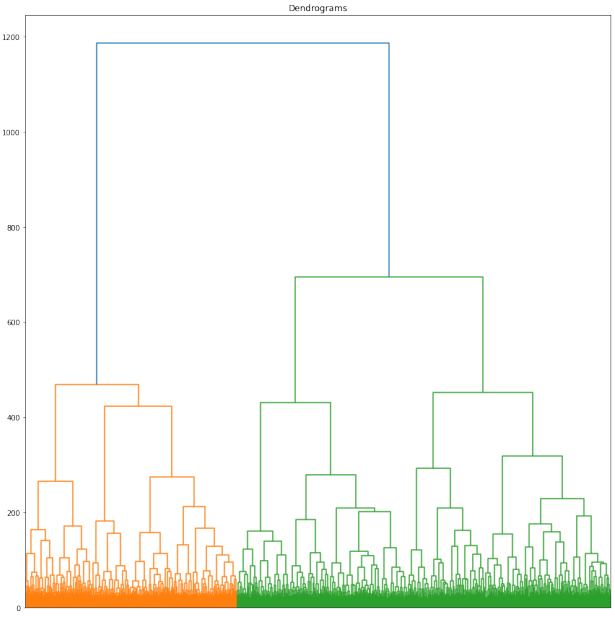

Figure 12: *Dendrogram for BERTweet embedding*

The final silhouette coefficients were equal to -0.0117 for euclidean distance and -0.0206 for cosine similarity, indicating that with this approach we didn't obtain good results.

5.4 Evaluation results

As mentioned in the previous paragraphs, the principal evaluation metrics used to compare the clustering models were the silhouette coefficient based on euclidean distance and the silhouette coefficient based on cosine similarity. A summary of the scores for each model is reported in table 1 and 2.

	Clustering approaches	
Embedding	**K-means**	**Hierarchical**
TF-IDF	0.0093	0.0039
Doc2Vec	**0.4097**	**0.2019**
BERTweet	0.0229	-0.0117

Table 1: *Evaluation results with silhouette coefficients based on euclidean distance*

	Clustering approaches	
Embedding	**K-means**	**Hierarchical**
TF-IDF	0.0176	0.0070
Doc2Vec	-0.0841	**0.3115**
BERTweet	**0.0374**	-0.0206

Table 2: *Evaluation results with silhouette coefficients based on cosine distance*

Overall, based on the silhouette coefficients, we can say that the best clustering results are obtained with Doc2Vec embedding. In terms of the two different clustering approaches, on the other hand, the k-means algorithm outperforms the hierarchical method for euclidean distance, while it is not suitable for cosine similarity, for which hierarchical clustering seems to perform better. This, as mentioned before, is probably due to the fact that k-means tends to minimize the euclidean distance instead of the cosine one.

5.4.1 BART summaries

In order to evaluate more empirically the usefulness of the various clustering approaches, we derived a summary from each cluster using the BART pre-trained model. We report,

for each method, the most interesting summaries.

K-means

- **TF-IDF**
  - Cluster 0: *A 14-year-old girl in tehran was identified by cameras after removing her hijab and died in the hospital after being arrested by government forces due to a severe rupture of her vagina. 41 years ago, iranian women were active participants in the islamic revolution of iran. 43 years of forced religion has made iranians allergic to the concept of islam.*
  - Cluster 20: *This is the third final to go for penalty shootout following 1994 and 2006. leo messi has scored in each of the stages at the qatar 2022 fifa world cup. this will be the fourth world cup meeting between argentina and france. only once before in the men's world cup have both finalists been defeated in the group stage of the tournament.*
  - Cluster 28: *More than 6.500 workers died building the stadiums people are now cheering from. Many of the migrant workers who survived after making possible haven't been paid what they were promised. Host-nation qatar, which has spent a record $220 billion to make this wc possible has not coughed up the money.*
- **Doc2Vec**
  - Cluster 0: *qatar hosted one of the greatest world cup's ever! safe, entertaining, welcoming and showcasing islam and arab culture superbly. western media are angry about the garment the emir put on messi. it's the same garment worn by kings and princes in the gulf nations.* [...]
  - Cluster 3: *qatar has made football worldwide! now even the americans are following it! this is one of the greatest of all time! excellent job for hosting such a memorable event! kudos to the team and security! success indeed. this is surely is definitely the best ever. this world cup has been incredible! big shoutout to the best multilingual sports platform in the world. we're just getting started this is not the end!*
- **BERTweet**
  - Cluster 3: *argentina are only the 2nd team in history to win the world cup after losing their 1st game of the tournament.* [...] *what a great fight, even without benzima, pogba and kanye. first hat-trick in world cup finals since 1966. let's welcome the newest goat as the old exits.*

Hierarchical clustering

- **TF-IDF**
  - Cluster 2: [...] *karim benzema will miss the world cup. sadio mané will not be able to be part of the squad as he's not recovering from his injury. christopher nkunku will miss world cup due to injury in today's training session.*
  - Cluster 11: *iranian team refused to sing the official iranian anthem at as a sign of support for protesters in their homeland. In the last 63 days, more than 53 children have been killed by the iranian regime in the streets. Football has no meaning without people's support, while one of the famous players, voria ghafouri was arrested yesterday.*
- **Doc2Vec**
  - Cluster 5: *i can't express my gratitude and happiness for my participation in the biggest event of all times the world cup in my country with the talented jung kook, thank you for everything. great to see such a fairytale is still possible in modern football - this will give so many people so much power.* [...]
- **BERTweet**
  - Cluster 3: *iranian team refused to sing the official iranian anthem at as a sign of support for protesters in their homeland.* [...]
  - Cluster 11: *been told france president emmanuel macron has visited morocco's dressing room after the game and told sofyan amrabat that he has been "the best midfielfer of the tournament" in front of all the squad.* [...]

The reported clusters can be divided as follows:

- **Negative opinions**:
  - K-means TF-IDF - cluster 0, 28.
  - Hierarchical TF-IDF - cluster 11.
  - Hierarchical BERTweet - cluster 3.
- **Positive opinions**:
  - K-means Doc2Vec - cluster 0, 3.
  - Hierarchical Doc2Vec - cluster 5.
- **General statistics and episodes**:
  - K-means TF-IDF - cluster 20.

  - Hierarchical TF-IDF - cluster 2.
  - Hierarchical BERTweet - cluster 11.

For the purpose of our research regarding the different interpretations of this controversial World Cup, it is therefore interesting to note the polarization of the tweets. On the one hand, many users are satisfied with the organization and the spectacularity of the tournament, while on the other hand there are several complaints and controversies, mostly regarding the conditions of those who worked on the construction of the stadiums and Iran's participation in the tournament. This World Cup is meant to sharply divide the various opinions of the spectators.

There are also several tweets that focus on interesting statistics (e.g. hierarchical TF-IDF - cluster 2 focuses on the absence of several important players due to injuries) or on singular events that occurred during the course of the tournament, such as the visit of the French president Emmanuel Macron inside the Moroccan dressing room (hierarchical BERTweet - cluster 11).

6 Text summarization

Text summarization is a task in Natural Language Processing (NLP) that aims to condense a text document or passage into a shorter version that still contains the most important information. This can be done using a variety of techniques, including extractive and abstractive methods.

Our goal was to summarize the most important hashtags and those hashtags related to the teams that made it to the semifinals.

Taking into account to avoid very similar hashtags (e.g. #FIFAWorldCup and #WorldCup2022), we chose and grouped the tweets based on the following 8 hashtags:

- #Qatar2022 (i.e. the most general hashtag regarding the World Cup).
- #SayTheirNames.
- #BoycottQatar.
- #Messi.
- #Argentina.
- #France.
- #Croatia.
- #Morocco.

Firstly, for each sub-dataset (except the most general one, i.e. #Qatar2022) we explored the following features, which are reported in table 3 and 4:

- Number of tweets (N).
- Number of sentences (S).
- Average number of sentences for each tweet (Avg-S).
- Number of words (W).
- Average number of words for each tweet (Avg-W).
- Number of unique words (U).
- Ratio between unique words and words (U/W).

	Hashtags		
Statistics	**#SayTheirNames**	**#BoycottQatar**	**#Messi**
N	1567	779	2556
S	3245	1562	5179
Avg-S	2.070	2.005	2.026
W	50751	20942	59633
Avg-W	32.38	26.88	23.33
U	5140	5183	8612
U/W	0.101	0.247	0.144

Table 3: *Features for relevant hashtags*

	Hashtags			
Statistics	**#Argentina**	**#France**	**#Croatia**	**#Morocco**
N	3392	2342	1063	2079
S	6876	4770	2181	4283
Avg-S	2.027	2.036	2.051	2.060
W	77502	53981	24995	50253
Avg-W	22.848	23.041	23.513	24.171
U	10449	8447	4950	7737
U/W	0.134	0.156	0.198	0.153

Table 4: *Features for semifinalists hashtags*

Among all these statistics, it is interesting to note, also from the graph below (Fig. 13) that controversial tweets, i.e. those containing the hashtags #SayTheirNames and #BoycottQatar, have the highest average number of words per tweet.

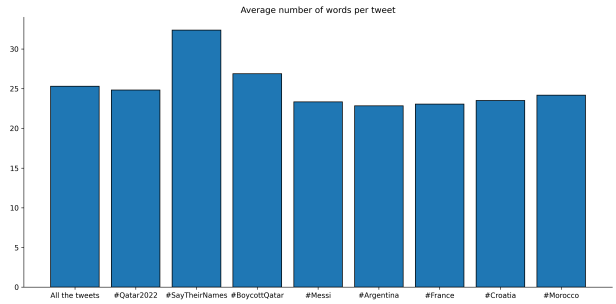

Figure 13: *Average number of words per tweet*

Since we are working on a domain without having a benchmark and metrics to evaluate our summaries, we decided to use different approaches and manually compare the results.

Moreover, BART and BERTweet are capable of processing emoji, but the other embedding models and methods are not, so each sub-dataset was created with a dual mode: with emoji in their natural format and with emoji transformed into text.

6.1 Abstractive summarization

Abstractive summarization is a technique for text summarization that involves generating new phrases and sentences that convey the most important information from the original text. This technique is more complex than extractive summarization, as it requires the model to understand the meaning of the text and then generate a summary that captures that meaning.

We used this technique with 2 pre-trained models: T5 and BART. These models were fine-tuned on our tweet sub-datasets.

6.1.1 T5 model

The T5 (Text-to-Text Transfer Transformer) model [9] is a pre-trained model that can be fine-tuned for abstractive summarization. The T5 model was introduced by Google in 2020, it is based on the transformer architecture and has been pre-trained on a large corpus of text.

It's worth noting that T5 model is a large model, with billions of parameters, and requires a significant amount of computational resources to fine-tune and run. Additionally, like other abstractive summarization models, the generated summaries may not always be completely accurate or faithful to the original text, and require human evaluation to assess their quality.

6.1.2 BART model

BART (Denoising Autoencoder for Pre-training of Transformer-based Language Models) [8] is a pre-trained model developed by Facebook AI Research that can be fine-tuned for text summarization. The BART model is based on the transformer architecture and has been pre-trained on a large corpus of text, similar to BERT [10] and GPT-2 [11].

The BART model can be fine-tuned for text summarization by providing it with the text to summarize. The model is trained to reconstruct the summary from the input text by learning to selectively copy, delete and generate new words.

6.2 Extractive summarization

Extractive summarization is a technique for text summarization that involves selecting important sentences or phrases from the original text and combining them to create a summary. This method aims to identify the most informative parts of the text and present them in a condensed format. The summary generated by extractive summarization is a subset of the original text and it preserves the original order of sentences.

Extractive summarization is relatively easy to implement and can achieve good performance with a small amount of labeled data, but the quality of the summary may depend on the specific text and the technique used.

Furthermore, the summary generated by extractive summarization may lack fluency and coherence, as the sentences are often taken out of context.

We used LexRank [12] and TextRank [13] because these algorithms have been shown to be effective for text summarization, achieving good performance on several benchmark datasets.

6.2.1 LexRank

LexRank is an extractive summarization algorithm that is based on the idea of graph-based centrality. It is a graph-based method for identifying the most important sentences in a text. It works by constructing a graph of sentences, where each sentence is a node, and edges are drawn between sentences that are similar to one another.

The algorithm uses cosine similarity to measure the similarity between sentences and thus determine the edges in the graph. We used a *ranked_sentence* method to rank sentences in a text based on their importance. This method returns a list of scores, where each score represents the importance of a sentence in the text.

Specifically, we took as output the first 4 ranked sentences.

6.2.2 TextRank

TextRank is another extractive summarization algorithm, very similar to the LexRank algorithm described above. The difference is that the former uses cosine similarity to measure the similarity between sentences and thus determine the edges in the graph, while TextRank uses a word-overlap based similarity measure. It is a simple technique that counts the number of common words between two sentences, and use that as a similarity measure.

We combined the TextRank algorithm with different text representations, the same performed in the clustering task: TF-IDF, Doc2Vec and BERTweet.

As in the case of LexRank, for each summary we chose the first 4 scores but, instead of sentences, we considered tweets. This is because LexRank was applied directly to raw texts through the proper library, whereas with TextRank we first applied embedding techniques. In particular, Doc2Vec and BERTweet are designed specifically for, respectively, documents and tweets, which in our case correspond. Therefore, embedding via these two methods outputs features related to tweets, not sentences.

TF-IDF The algorithm employs TF-IDF to weight the words in each tweet before constructing the graph. This means that the edges between tweets are drawn based on the similarity of the weighted words rather than the raw frequency of words.

Doc2Vec We used Doc2Vec to obtain vector representations of the tweets, then we applied the TextRank algorithm to these vectors in order to identify the most important tweets in the text.

Specifically, the model was fine-tuned with the same parameter combination that we found before during clustering.

BERTweet BERTweet encodes each tweet as a vector with 768 features. So, in this case we applied the TextRank algorithm to the BERTweet output to identify the most important tweets.

6.3 Evaluation results

The summaries obtained were evaluated by means of human assessments, as it is not possible to use metrics (such as ROUGE) without references or labeled summaries. However, one possible method of evaluation concerns the scores of sentences (for LexRank) and tweets (for TextRank), depending on the different embedding methodology applied.

6.3.1 Abstractive summarization

Some summaries obtained with this methodology were reasonable, while others were inconsistent. In general, the most sensible results came from the hashtag #BoycottQatar. These tweets (and tweets in general) are a difficult textual data to treat: they are short, informal and very domain-specific. However, the tweets about controversial aspects of the World Cup (the same tweets with the highest average number of words per tweet) seems more suitable for this task. Probably, the reason lies in the fact that other tweets are often simple externalizations regarding the matches, whereas these tweets often consist of more articulate sentences, used to express negative opinions about the event. Regarding the two different models, BART seems to produce more consistent summaries.

We report the summaries for T5 and BART, highlighting in red the parts that do not make sense or are wrong and in yellow the repetitions and those parts not relevant to the specific hashtag.

T5

- #**Qatar2022**: *argentina beat france 1-0 in the world cup final in rio de janeiro . lionel scaloni wore the same shirt he wore in the 1997 world cup final. qatar 2022 is the biggest footballing nation in the world. argentina will be represented by the 'pirate flag' of qatar .*
- #**SayTheirNames**: *iranians are celebrating islamic republic national team's loss against the united states. iranians are celebrating elimination of islamic republic from world cup. iranians are celebrating the death of children killed by the islamic regime.*
- #**BoycottQatar**: *i've boycotted the whole men's world cup in as more than 6.500 workers died. i'm not watching the final today as i'm going to watch the documentary. i'm not sure i'm going to watch the final but i'm going to watch the documentary .*
- #**Messi**: *argentina beat france 4-2 on penalties to win their third world cup. lionel messi was the _goat_ who won the _world_cup_ in 1986 . argentina will travel on an open bus on a victory parade through lusail boulevard. argentina will play england in the final on june 14 in brazil .*
- #**Argentina**: *argentina beat france 4-2 on penalties to win the world cup for the third time. lionel messi's once-in-a-generation career is complete: he is a world cup champion. argentina fans will be able to celebrate in the streets of buenos aires.*
- #**France**: *argentina beat france 4-2 on penalties in world cup final in qatar. lionel messi wore traditional arabic robe to celebrate win. argentina fans clapped and cheered as they won the gold cup .*
- #**Croatia**: *croatia airlines brought back the country's national football team home after a secured third-place victory in the fifa world cup 2022. "nothing can stop us, we are always here. these people have united once again," said luka.*
- #**Morocco**: *croatia defeated morocco 2-1 to claim third place in the 2022 qatar world cup. danish television channel came under fire after its presenter compared the moroccan players and their mothers with monkeys. a danish television channel said: "ya allah" — — was sounded during the opening ceremony of the world cup .*

BART

- #**Qatar2022**: *Argentina beat France 2-0 on penalties to win the 2022 fifa world cup in Qatar. Lionel Scaloni celebrated the win in the same shirt he wore when he won the u-20 world Cup in 1997. "It is a dress for an official occasion, worn for celebrations. this was a celebration of messi," ceo hassan al thawadi said.*
- #**SayTheirNames**: *Ilyas: "Shameless" falls short to explain your behavior! up to 5% of iran consists of people and yet they make for 30% of hanged prisoners in while we celebrate the irans losing against usa, you should .. not so fast, even before you meet. he got murdered last night by security forces of islamic regime in iran. why? simply because he was happy that islamIC regime football team lost a game to usa and got eliminated from, and honk his car horn. i wish he would've cried like this for all the children who were murdered by islamics regime. i cannot watch soccer while kids are being raped and murdered.*
- #**BoycottQatar**: *More than 6.500 workers died building the stadiums people are now cheering from. i've boycotted the whole men's world cup in as more than 6,000 workers died. i think the only people worse than political pundits are sports pundits. we choose to get together and watch the documentary*

"the worker's cup".

- #**Messi**: *12/18/2022 - times square, nyc argentina beats france on penalty kicks, winning world cup for third time. still can't sleep. this magical moment is still on our minds. no messi fan will pass without giving it a like. messi is the goat. i wanted this to happen so much. i cheer for since 1986. my first experience seeing on tv with my grandparents & learning about football. to see argentine in doing that again is so magical.*
- #**Argentina**: *12/18/2022 - times square, nyc argentina beats france on penalty kicks, winning world cup for third time. i celebrate with a quick drawing sorry for the delay, i'm preparing several illustrations based on the world cup and commissions. still can't sleep. this magical moment is still on our minds. no messi fan will pass without giving it a like. messi is the goat. where is de paul. "all's well that ends well" congrats .*
- #**France**: *12/18/2022 - times square, nyc argentina beats france on penalty kicks, winning world cup for third time. when you live in little buenos aires and argentINA wins the bravo argentine!. look at this picture very and understand one thing, follow who know road because this picture represent something huge. follow your destiny don't give up until you get it. leo messi, he really do made his dream come true. he did it! best game ever best tournament evar congratulation argentinas.*
- #**Croatia**: *Croatia airlines brought back the country's national football team home after a secured third-place victory in the fifa world cup 2022. beat morocco 2-1 in 3rd-place playoff at khalifa international stadium. after becoming the first african team to reach the world cup semifinals, morocCO's tournament ended in two losses. watch match highlights .*
- #**Morocco**: *Danish television channel came under fire after its presenter compared the moroccan players and their mothers at the 2022 world cup in qatar with monkeys. "ya allah" was sounded during the opening ceremony of the . the most viewed event in the world. regragui became the first african & arab team to reach the last four.*

6.3.2 Extractive summarization

For the same reasons explained before, the most consistent results came from the hashtags #SayTheirNames and #BoycottQatar. We report in the following table (table 5) the results obtained through LexRank and TextRank, using TF-IDF, Doc2Vec, and BERTweet as representations. In particular, we employed the mean of the top-4 scores produced by the models as a performance evaluation method.

		TextRank		
Hashtags	**LexRank**	**TF-IDF**	**Doc2Vec**	**BERTweet**
#Qatar2022	3.47	0.0008	0.0003	0.0005
#SayTheirNames	1.89	0.0015	0.0006	0.0009
#BoycottQatar	3.33	0.0034	0.0013	0.0019
#Messi	2.60	0.0011	0.0004	0.0006
#Argentina	2.52	0.0009	0.0003	0.0004
#France	2.67	0.0014	0.0004	0.0006
#Croatia	2.66	0.0027	0.0009	0.0013
#Morocco	2.38	0.0014	0.0005	0.0007

Table 5: *Top-4 scores for extractive methods*

In general, it is difficult to say with certainty which model is better for this summarization task using human-in-the-loop assessment.

However, looking at the relative scores for the models, we can notice that, for LexRank, the highest values were achieved from the general hashtag #Qatar2022 (3.47) and #BoycottQatar (3.33). The latter hashtag seems, again, to be the one that clusters more informative and meaningful tweets.

With regard to TextRank, TF-IDF representation achieved the best performances, while the embedding technique that returned the best results was BERTweet. The worst performances were achieved by Doc2Vec. The main reason behind this concerns the large amount of data required from both TextRank and Doc2Vec in order to obtain good performances. Moreover, Doc2Vec is a model pre-trained on a large collection of texts, but unlike BERTweet it has not been specifically trained on tweets.

Looking specifically at the hashtags among the various TextRank embeddings, those composed of the most significant tweets are #BoycottQatar, #Croatia and #SayTheirNames.

7 Conclusions

We extracted, by scraping, the tweets regarding the 2022 World Cup held in Qatar (section 2). We initially preprocessed the data according to the model and embedding used (section 3), and then proceeded with a cursory exploratory analysis aimed at highlighting the basic characteristics of our dataset (section 4). Next, the two tasks of text clustering and text summarization were performed. For the former (section 5), two types of algorithms (k-means and hierarchical clustering) were compared, with 3 different text representations (TF-IDF and Doc2Vec and BERTweet embeddings). Each model was evaluated with silhouette coefficients and summaries generated from BART pre-trained model. For the text summarization (section 6), we conducted an analysis based on abstractive and extractive methodologies. Regarding abstractive summarization, we performed fine-tuning of two pre-trained models (T5 and BART), reporting the results by means of human-in-the-loop assessments. LexRank and TextRank algorithms were applied for the extractive methodology. For the latter, 3 different text representations were compared, obtained through TF-IDF, Doc2Vec and BERTweet. These results were compared through the mean of top-4 scores achieved from each model for each subset.

In conclusion, the application of these two tasks showed the extreme bias of the event. Supporters of the World Cup make forceful claims about the spectacular nature of the event and the outstanding matches played, while detractors express polemical claims about the construction of the stadiums and other various scandals that have dotted the unfolding of the competition.

8 Future work

Despite the various insights uncovered, the complexity of the dataset and the novelty of the approach developed inevitably influenced the results. Therefore, it might be interesting to further explore and expand such research. Some possibilities may be:

- Increase the number of tweets.
- Through access to more powerful computational resources, all tweets could be used for hierarchical clustering or other methods such as DBSCAN or spectral clustering may be exploited.
- Use of Named Entity Recognition (NER) and Sentiment Analysis (SA) techniques, in order to identify important entities such as people, organizations, and locations in tweets, as well as to understand the sentiment expressed in them. These informations can be used to help prioritize certain tweets in the summarization process.
- Perform topic modeling task in order to further exploit the highlights of the event.

Bibliography

[1] URL: https://github.com/MartinBeckUT/TwitterScraper/tree/master/snscrape.

[2] URL: https://pypi.org/project/langdetect/.

[3] URL: https://pypi.org/project/tweet-preprocessor/.

[4] URL: https://radimrehurek.com/gensim/models/doc2vec.html.

[5] URL: https://pypi.org/project/emoji/.

[6] URL: https://pypi.org/project/contractions/.

[7] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. *BERTweet: A pre-trained language model for English Tweets*. 2020. DOI: 10.48550/ARXIV.2005.10200. URL: https://arxiv.org/abs/2005.10200.

[8] Mike Lewis et al. *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*. 2019. DOI: 10.48550/ARXIV.1910.13461. URL: https://arxiv.org/abs/1910.13461.

[9] Colin Raffel et al. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. 2019. DOI: 10.48550/ARXIV.1910.10683. URL: https://arxiv.org/abs/1910.10683.

[10] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2018. DOI: 10.48550/ARXIV.1810.04805. URL: https://arxiv.org/abs/1810.04805.

[11] Alec Radford et al. *Language Models are Unsupervised Multitask Learners*. 2019. URL: https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

[12] URL: https://github.com/crabcamp/lexrank/tree/dev/lexrank.

[13] URL: https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.link_analysis.pagerank_alg.pagerank.html.