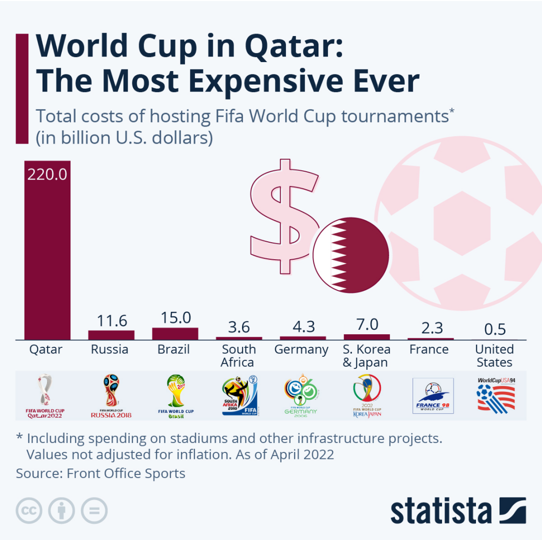
2022 FIFA World Cup Qatar Twitter Analysis

The final project for course: Social Media Analytics
At university: Università degli Studi di Milano-Bicocca
Authors: Konrad Pawlik, Jan Fiszer

Introduction

World cup events have been attracting people from all over the globe for years. Emotions related to the most popular sport spiced up by supporting your national team make the Mundial a real fest for a huge amount of people. Therefore it has been a source of terrifyingly big money.

This year's World Cup was organised by Qatar, which, for many people, exceeded the allowed level of controversy. When they were "chosen" as the host of Mundial there were no stadiums in the whole country and that was just the beginning...


World Cup in Qatar: The Most Expensive Ever

Total costs of hosting Fifa World Cup tournaments* (in billion U.S. dollars)

* Including spending on stadiums and other infrastructure projects. Values not adjusted for inflation. As of April 2022

Source: Front Office Sports

The Guardian published an article[1] that mentions the deaths of 6,500 migrant workers. The way they were manipulated to get the job and used there is just the tip of the iceberg of Qatar's cruelty.

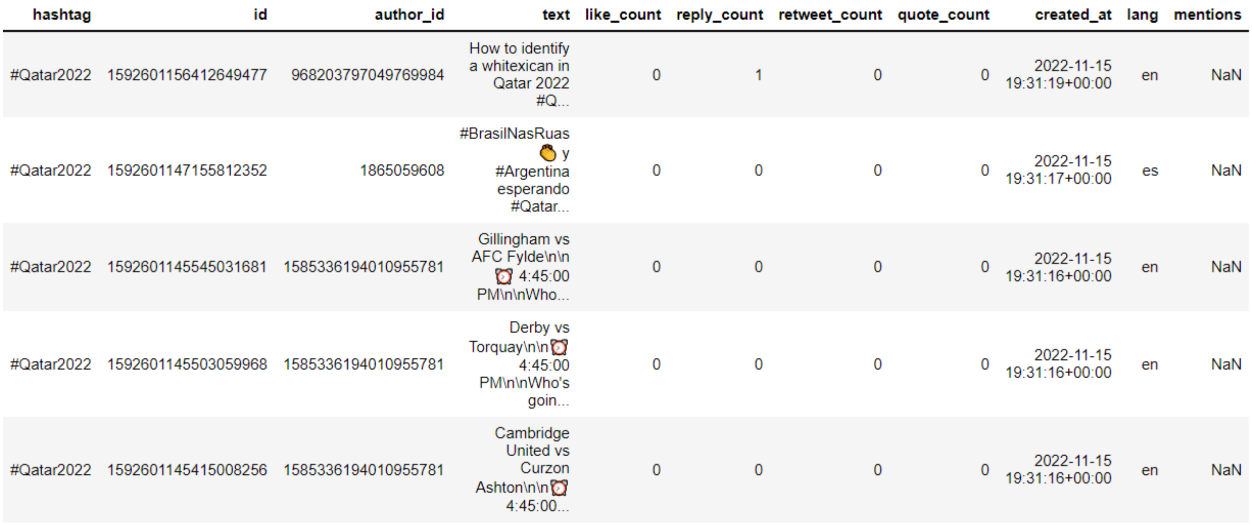
The aim of our project is to analyse public opinion on the World Cup and to investigate the use of publicity of the event by other groups for their own purposes.

[1] link to the article: https://www.theguardian.com/football/2022/nov/27/qatar-deaths-how-many-migrant-workers-died-world-cup-number-toll

Data collection

The first step in such a project is to collect data. After analysing the first World Cup related tweets we decided to collect those which contain hashtags: *#QatarWorldCup2022, #Qatar2022* and *#FIFAWorldCup*.

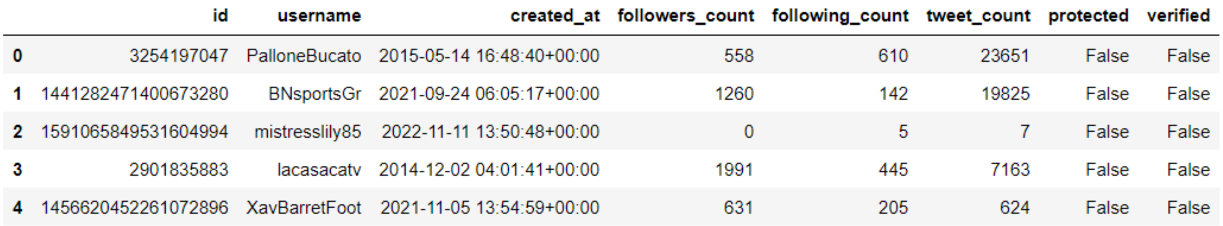
By using Twitter API v2 included in the tweepy python library, over a month we have managed to collect data of 332589 tweets and 142507 users (without duplicates).



hashtag	id	author_id	text	like_count	reply_count	retweet_count	quote_count	created_at	lang	mentions
#Qatar2022	1592601156412649477	968203797049769984	How to identify a whitexican in Qatar 2022 #Q...	0	1	0	0	2022-11-15 19:31:19+00:00	en	NaN
#Qatar2022	1592601147155812352	1865059608	#BrasilNasRuas 👏 y #Argentina esperando #Qatar...	0	0	0	0	2022-11-15 19:31:17+00:00	es	NaN
#Qatar2022	1592601145545031681	1585336194010955781	Gillingham vs AFC Fylde\n\n ⏰ 4:45:00 PM\n\nWho...	0	0	0	0	2022-11-15 19:31:16+00:00	en	NaN
#Qatar2022	1592601145503059968	1585336194010955781	Derby vs Torquay\n\n⏰ 4:45:00 PM\n\nWho's goin...	0	0	0	0	2022-11-15 19:31:16+00:00	en	NaN
#Qatar2022	1592601145415008256	1585336194010955781	Cambridge United vs Curzon Ashton\n\n⏰ 4:45:00...	0	0	0	0	2022-11-15 19:31:16+00:00	en	NaN

Since the API allows us to download information about tweets and tweet-connected users (there might be more of them than one) we decided to divide the collected data into two sets:

- tweets:
    - id - unique tweet identifier, allows distinguishing tweets
    - author_id - unique author identifier, connects a tweet to the user (public_key between two entities)
    - text - tweet content, textual analysis
    - like_count, reply_count, retweet_count, quote_count - public metrics, might be used to judge tweet reliability
    - created_at - date of the tweets creation, allows for analysis through time
    - lang - the language of a tweet, for textual analysis it will be good to work on just one language
    - mentions - ids of mentioned users in a given tweet



	id	username	created_at	followers_count	following_count	tweet_count	protected	verified
0	3254197047	PalloneBucato	2015-05-14 16:48:40+00:00	558	610	23651	False	False
1	1441282471400673280	BNsportsGr	2021-09-24 06:05:17+00:00	1260	142	19825	False	False
2	1591065849531604994	mistresslily85	2022-11-11 13:50:48+00:00	0	5	7	False	False
3	2901835883	lacasacatv	2014-12-02 04:01:41+00:00	1991	445	7163	False	False
4	1456620452261072896	XavBarretFoot	2021-11-05 13:54:59+00:00	631	205	624	False	False

- users:
    - id - same as author_id in tweets (public key)
    - username - allows us to investigate the user by Twitter app
    - created_at - date of account creation, might allow verifying users' reliability
    - followers_count following_count, tweet_count, protected, verified - same, users' reliability

Data Preprocessing

First step in our data preprocessing was to remove duplicates, which may have occurred because of tweets containing more than one searched hashtags while being collected.

We have decided to distinguish tweets in English as they contain the most useful data we intend to use for the following analyses. Moreover, various NLP tools achieve the best performance when it comes to that language.

In order to obtain clean data, we applied following steps:

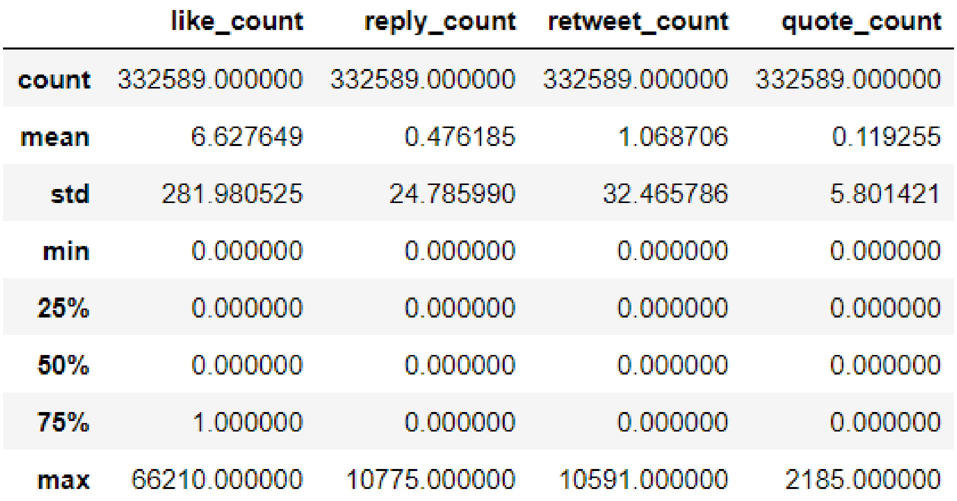
- Links and mentions removal (note: photos are presented as links)
- Hashtags removal (saved in a new column as a further step)
- Whitespace cleaning
- Numbers removal
- Lowering-casing text
- Punctuation removal
- Emoji removal

Further steps included lemmatizing and tokenizing involving WordNetLemmatizer and TweetTokenizer provided by NLTK library. The mentioned package allowed us to get rid of stopwords.

Results of each particular step were contained in distinct columns as various text representations were needed by thereafter used models.

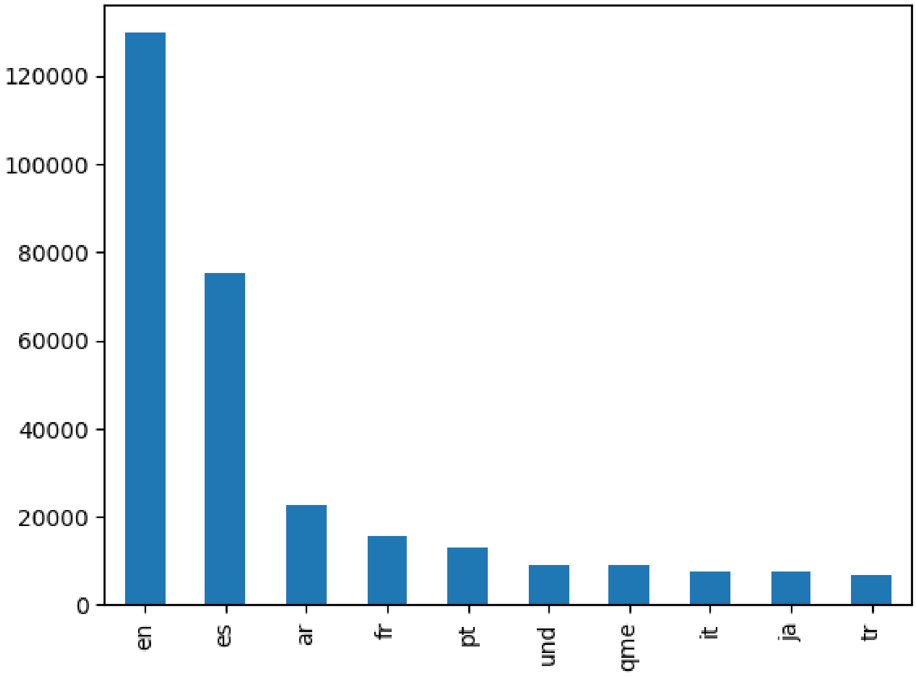
Exploratory Data Analysis

Before getting deeper into social network and content analysis, we’ve decided to briefly explore our data in order to better understand the audience and what their motives are.



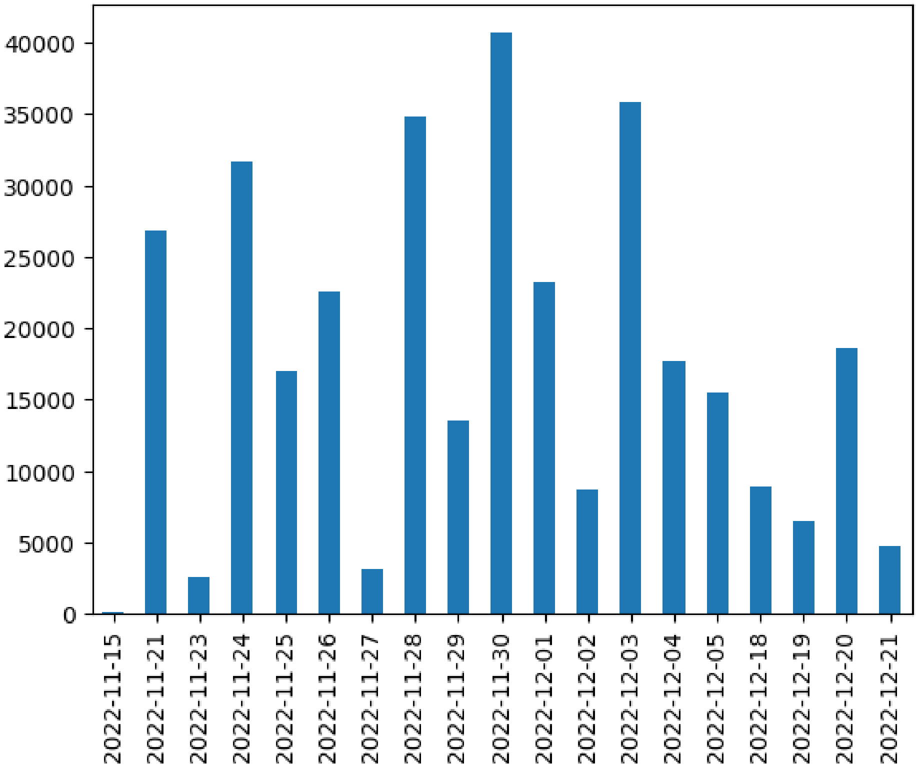
	like_count	reply_count	retweet_count	quote_count
count	332589.000000	332589.000000	332589.000000	332589.000000
mean	6.627649	0.476185	1.068706	0.119255
std	281.980525	24.785990	32.465786	5.801421
min	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000
50%	0.000000	0.000000	0.000000	0.000000
75%	1.000000	0.000000	0.000000	0.000000
max	66210.000000	10775.000000	10591.000000	2185.000000

As seen above, most of the tweets collected did not score highly for public metrics (e.g. like_count, reply_count, …). We can assume it might be due to either irregular data collecting or the presence of a large number of bots and advertisements. The later case we will try to examine in the following sections.



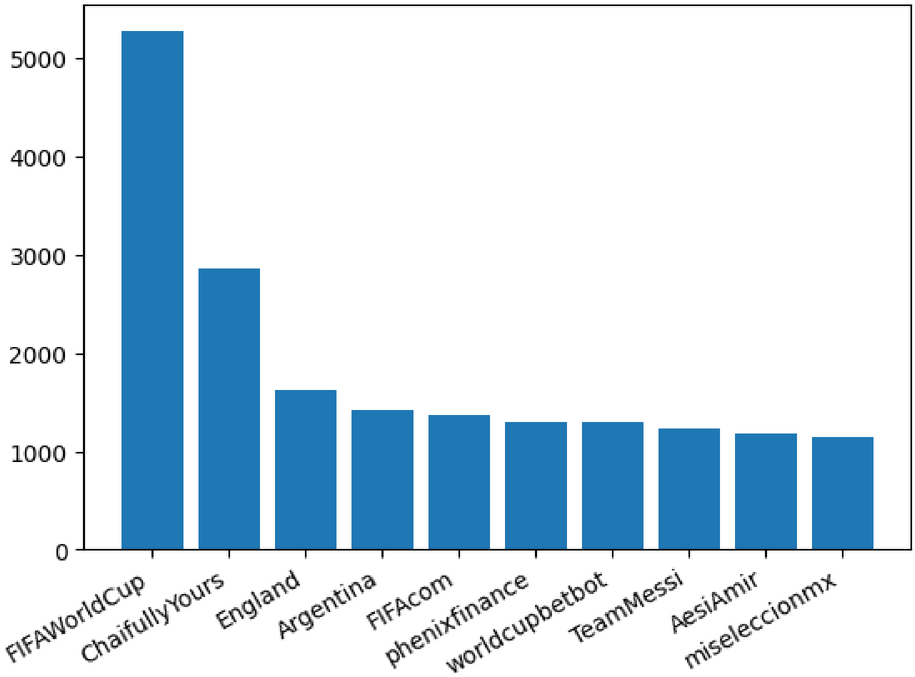
the most commonly used languages

As seen in the plot, English is the most frequently used language in our data, therefore we have decided to conduct a content analysis only on those tweets.



The number of data collected per day

We can observe an imbalance in the number of downloads per day. For this reason, we have abandoned an idea of thorough analysis of the data over a period of days. However, in selected cases we have discarded counts in favour of percentages.

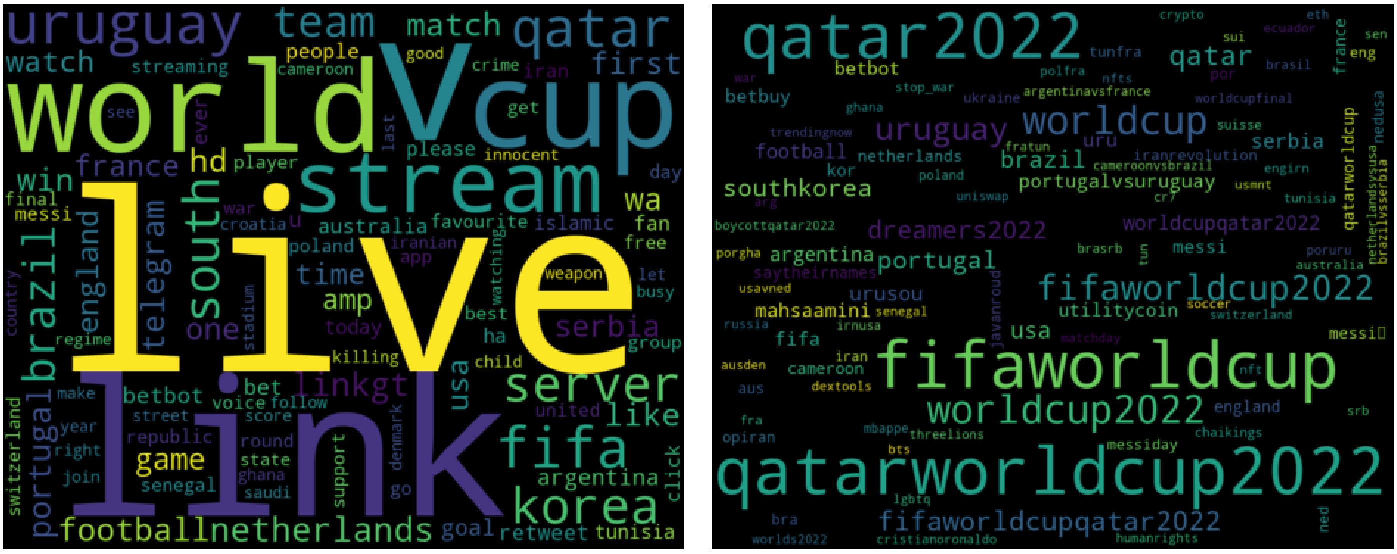


The most mentioned users

Our attention was drawn to a user named “ChaifullyYours”. After displaying a sample of the tweets posted, we have received a list of posts of the nature of the competition.

“#ContestAlert Guess the winner of the match and win goodies from Chai Kings… “

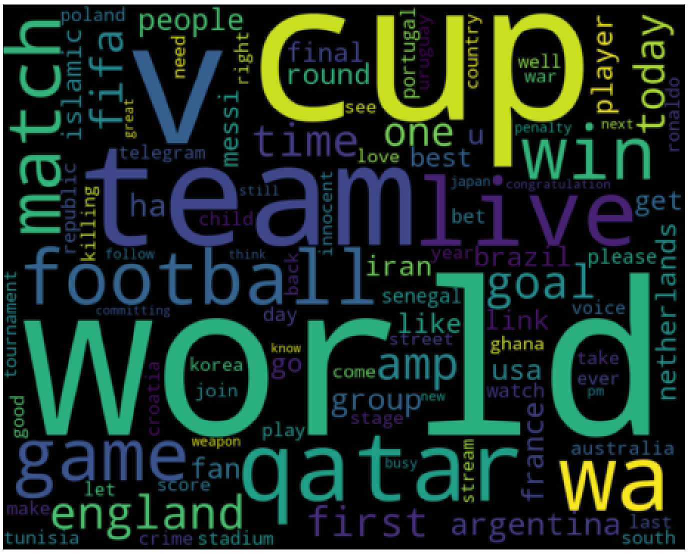
The obtained results confirm our concerns about the existence of numerous adverts and bots in our data

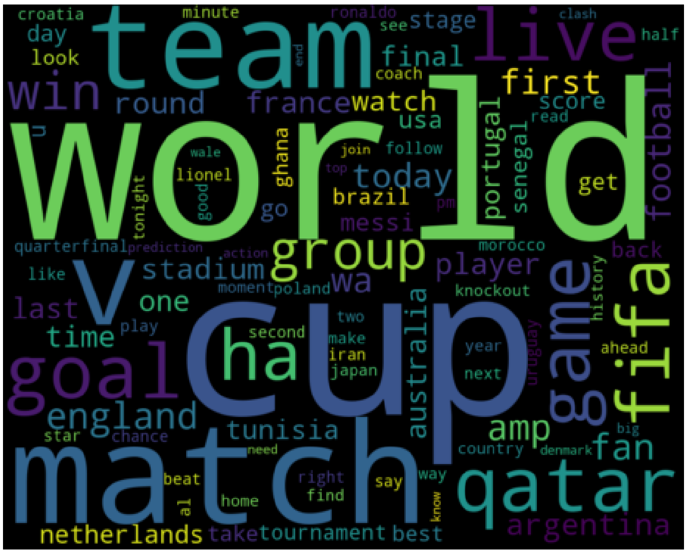


Wordclouds of text and hashtags from tweets in English

Apart from the most common words and hashtags, which are definitely related to competitions and live-streaming services, we observe key words that are not necessarily related to the World Cup or football, e.g. regime, crime, innocent, war… These cases will be the subject of further analysis.

In order to sift out unnecessary data, we have distinguished tweets created by users with a certain followers count and the verified status. The results are following:





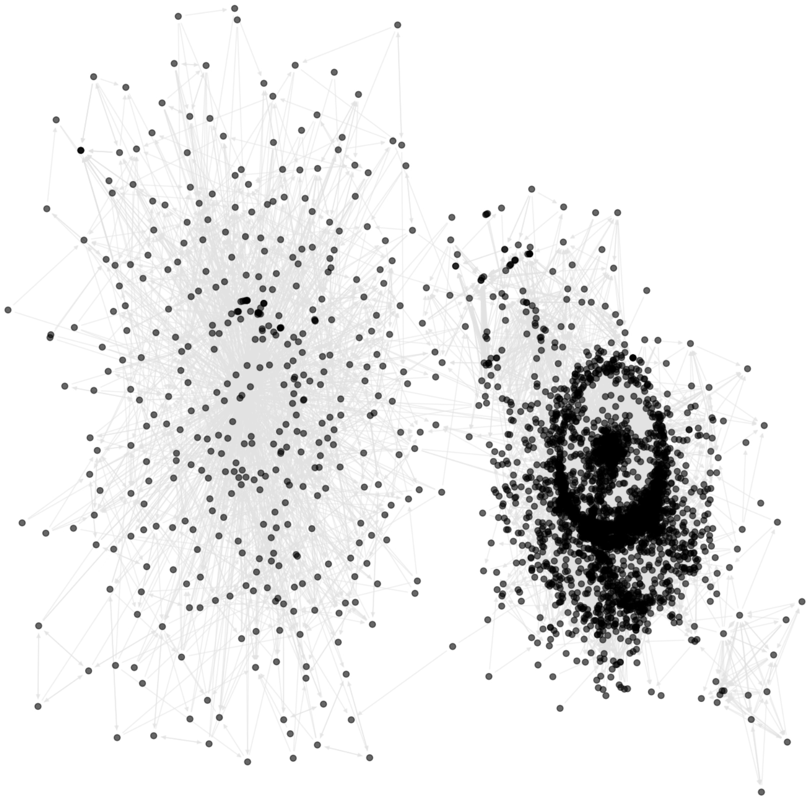
Both of the two representations look remarkably similar, however, what distinguishes them from the previous cases is the lack of words linked to advertising. Among the newly highlighted words are: republic, islamic, republic and busy.

Despite the success with getting rid of bots, in further analysis, we will abandon this approach so as not to skip ordinary users

Social network analysis

Since the profile ChaifullyYours drew our attention, we thought that mentions of the users might be an interesting topic to deepen into. When the second most mentioned id is a competition profile, we may observe much more from a brief analysis.

We build a directed graph where nodes are the users and edges are the mentions. That resulted in a huge social network (60360 nodes and 82915 edges), therefore to perform any investigation it had to be reduced.
First we looked at graph components. The main one has around 128 times more nodes than the second biggest. Moreover, the others were mostly about cryptocurrencies, so they were skipped in the further social network studies. Then by removal of nodes that had a low number(the threshold was 5, set by experimenting), a graph with 3800 nodes and 15954 was obtained which is pretty dense.



A graph plotted using networx python library

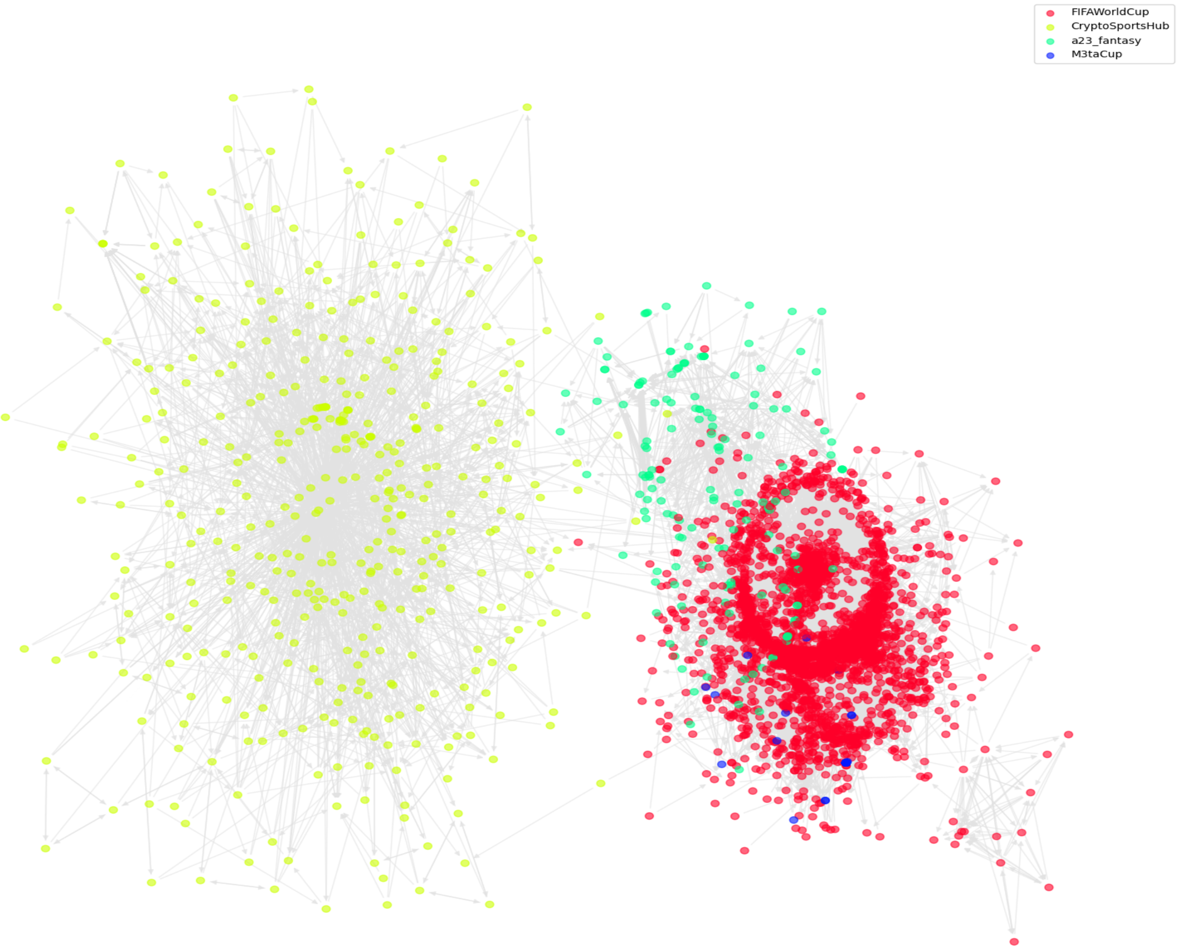
Community detection

Having applied greedy modularity detection with resolution equal to 0.04 established by trial-and-error method, we got 4 communities with 0.31 value of modularity which is a pretty acceptable result.

For each we:

- Found most tagged user and biggest tagger, which was simply finding node which highest in and out degree (degree centrality)

  Note: each of community on the legend is named after highest in degree node
- Computed the betweenness centrality and have a close look at users with the highest values
- Analise emotions (look at Social Content Analysis)

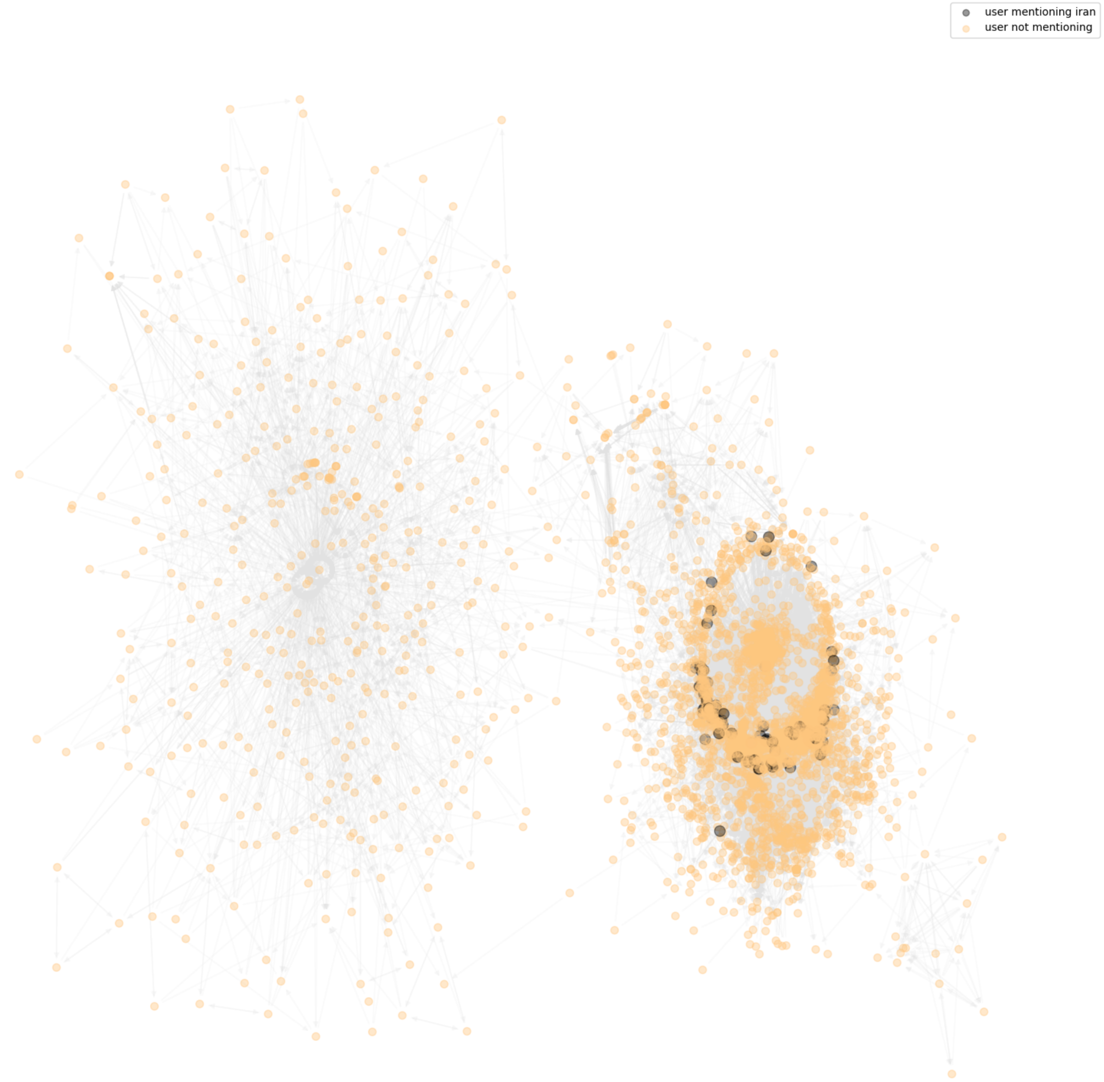


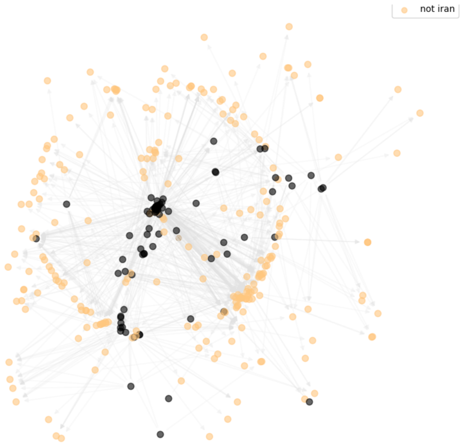
A graph representing detected communities and their most tagged users

INFO: A briefer description and examination of detected communities in Social Content Analysis.

Users mentioning iran

The next thing which we plotted in the graph was users using Iran-related hashtags (how they were collected explained in Social Content Analysis).





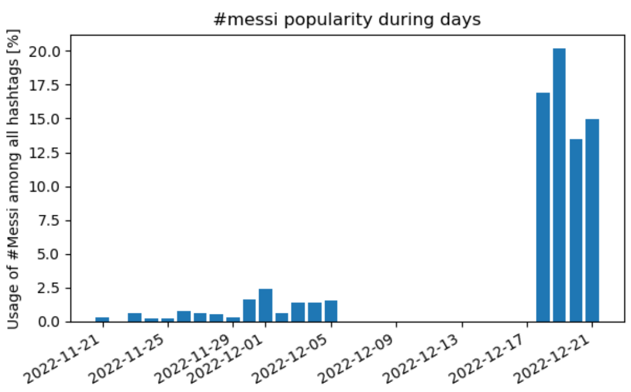
At the first moment, nothing special is visible since there are just 85 of them, but in the second plot, we can observe a very interesting property. It still shows Iran concerned nodes (in black) but the creme ones are those which on the big graph were their neighbours. An important thing to notice is that it has just one component, which means it is a very gated community. It means that there is a certain group of users who were involved in this topic, and all of them are closely connected. If not directly just one user separates them in the social network. In other words, if it were an undirected graph the shortest path between every two Iran nodes would be either one or two.

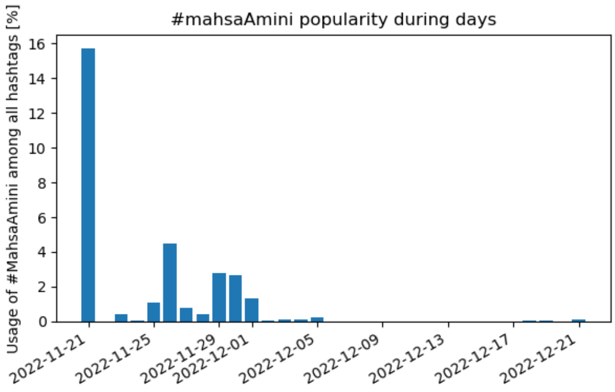
All the graph work was done by use of the networkx library.

Social Content Analysis

Hashtag usage through time

As mentioned before, because of irregular data collection through time analysis might be a bit tricky, but we still managed to observe some interesting properties.

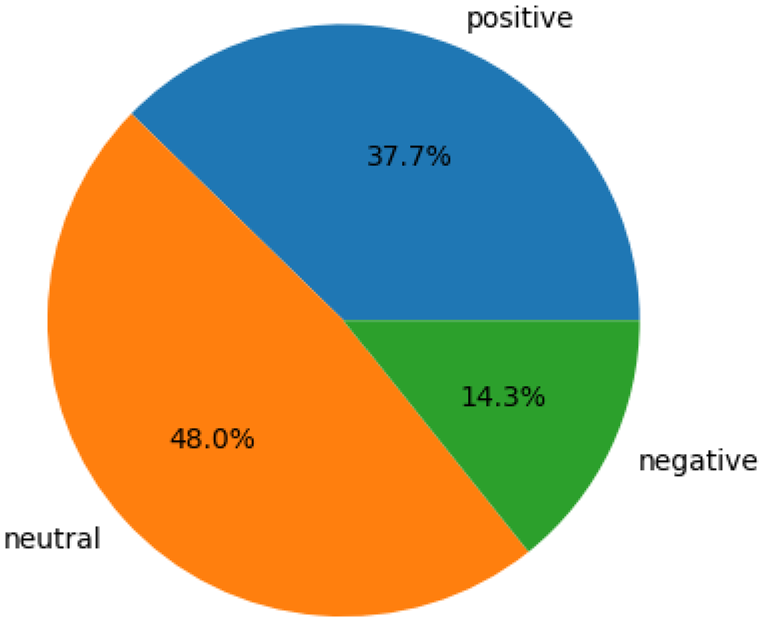




We can see the daily percentage of usage of two popular hashtags have changed through time. When World Cup started a lot of people were trying to discourage football fan by reminding Iran situation using inter alia #mahsaAmini[2] to increase the publicity of the problem. When Mundial was coming to end it became less popular and Messi became the main hashtag.

[2] Mahsa Amini was arrested in Iran for not wearing hijab and then beaten to death in a police station. NOr

Sentiment Analysis



The sentiment score was counted by using VADER lexicon and rule-based analysis tool. As suggested by the authors, we have used the compound score for our analysis with following thresholds:

1. Positive - compound score >= 0.1
2. Negative - compound score <= -0.1
3. Neutral - compound score < 0.1 and compound score > -0.1

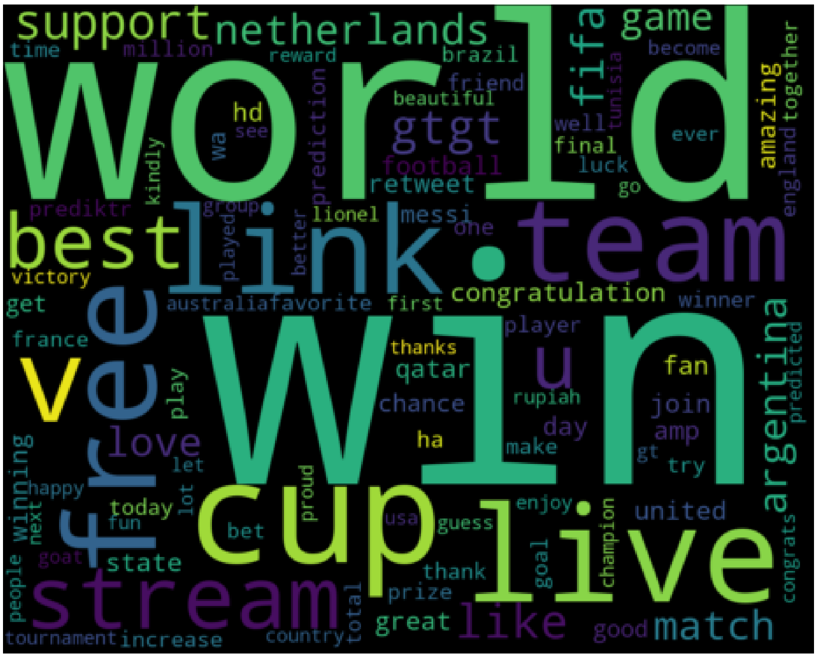
As our aim was to find out the audience's opinion, we decided not to analyse neutrally labelled tweets as they mostly consist of facts and news about upcoming matches.

In order to achieve the best results, we decided to further reduce the number of positive tweets by using a 0.8 compound score threshold for positive tweets and -0.6 for negative tweets.

Positively labelled tweets

Examples of tweets

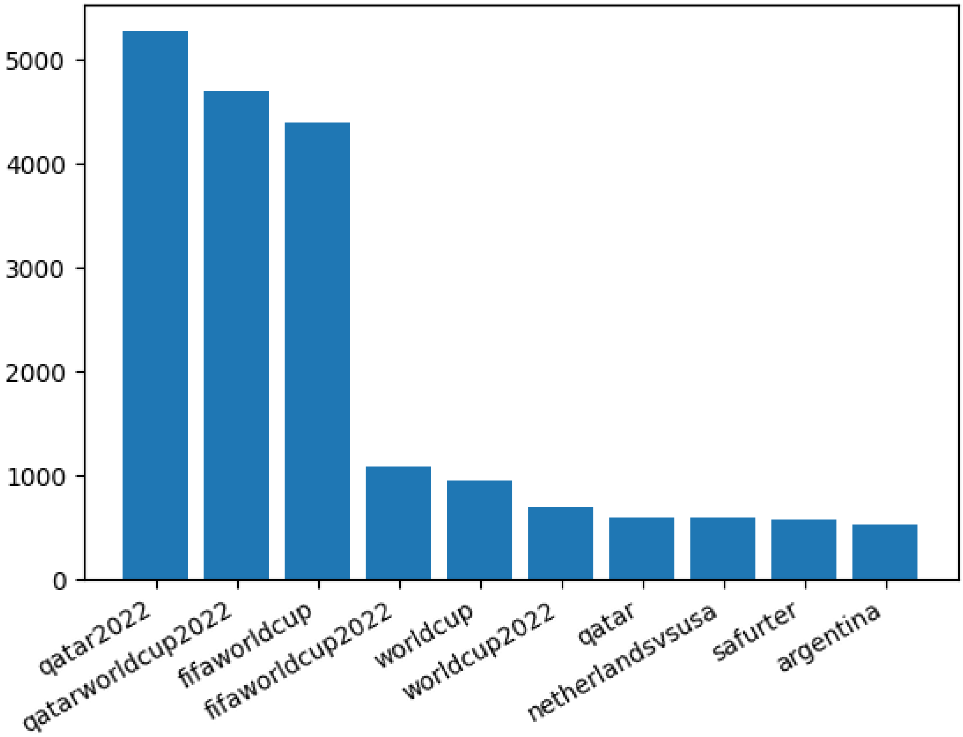
> *"this was amazing to watch so much respect for the audience and culture"*
> *"first game today best of luck you got this"*
> *"...so far they have really put together an amazing event"*



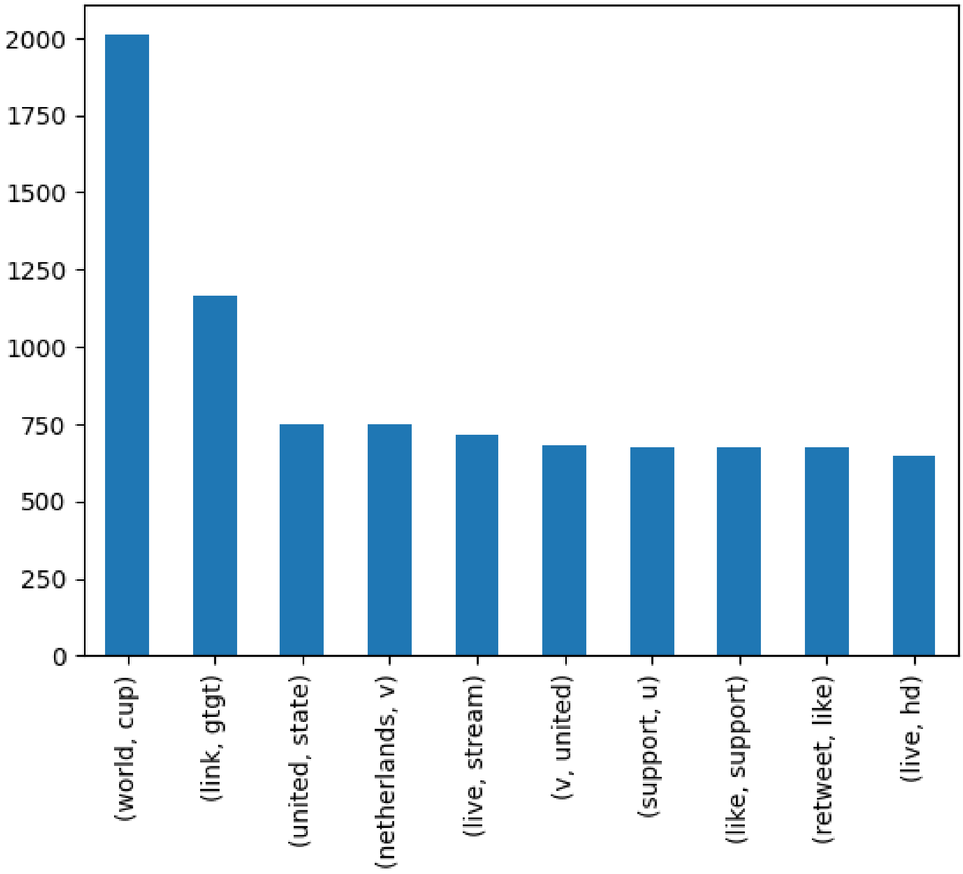
Wordcloud of texts from positively labelled tweets

Among the most frequent words we still find those related to streaming services, however, we find phrases that we can assume come from users wishing to express their satisfaction with their national team or the match they have enjoyed.

> *"Let's go Argentina. Today is a very important day. Today Argentina win…"*



The most frequent hashtags from positively labelled tweets



The most frequent bigrams from positively labelled tweets

Plotting the most frequent hashtags and bigrams directly points to the frequent repetition of similar content in the nature of advertising. Therefore, the next step was to group positive tweets on the basis of text content.

The most frequently duplicated messages:

> *“live hd stream…free live stream link gtgt link gtgt retweet like to support us”*
> *“support your favorite team and win total millions of rupiah prizes become a that should be together join now”*
> *“make the best of the world cup increase your winning chances with predictions by prediktr try for days for free”*

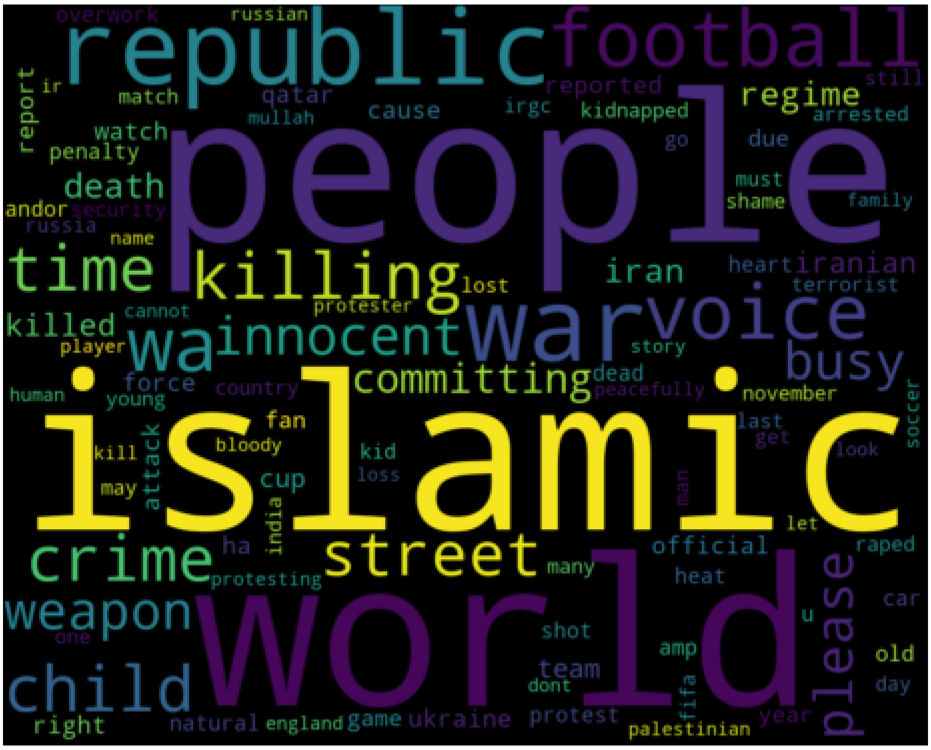
When reading the non-repeated content, we still encounter publicity content, however, in the midst of this, the messages of the real football fans emerge:

> *“No controversy Messi is simply the greatest of all time what a way to end a glorious and illustrious career”*
> *“Not only better the one France organized in but the best ever organized in the planet congratulations Qatar”*
> *“Love it without a doubt this is the most exciting and game show ever”*

Negatively labelled tweets

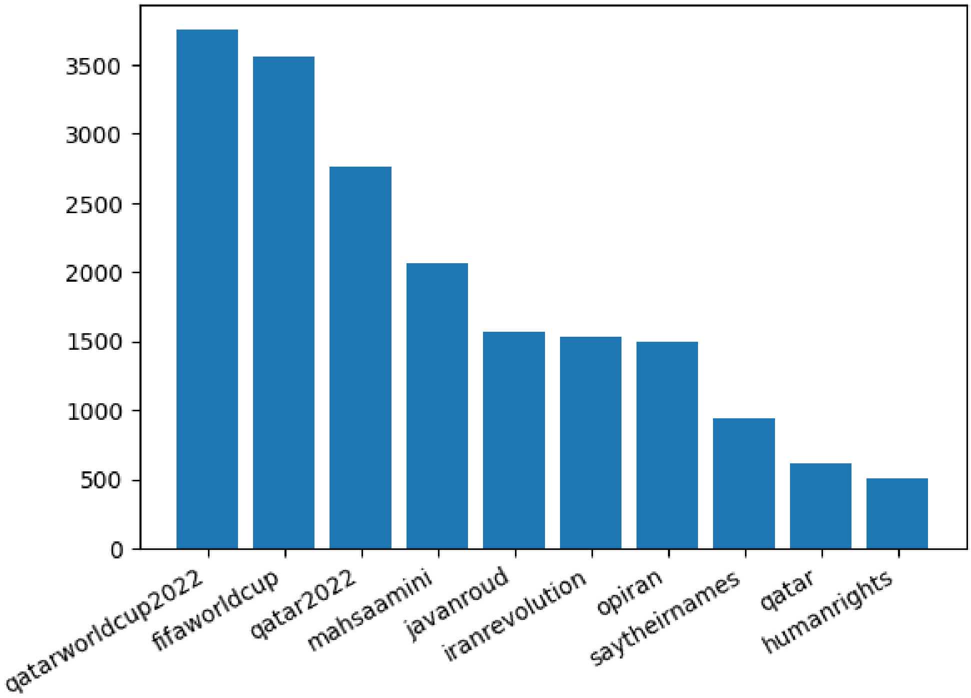
Examples of tweets

> *“Why a sensible man would allow such stupidness on a football field…”*
> *“Europe said to be inclusive but did not respect the local culture of Qatar…”*
> *“This is not right there is happening right under where all this murder and killing is happening in”*



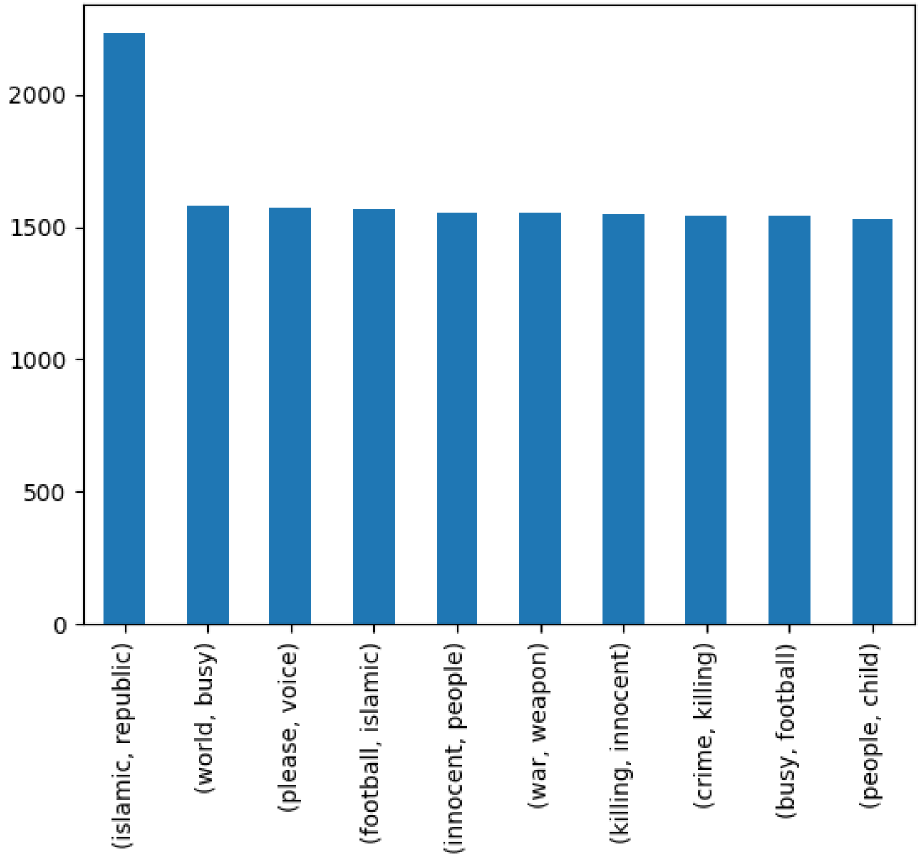
Wordcloud of texts from negatively labelled tweets

As opposed to positively labelled data, here we can observe a significant occurrence of words related to uncomfortable topics. At this stage, it can be already assumed that the WorldCup event is not the main source of the negative feelings.



The most frequently used hashtags from negatively labelled tweets

The results of the plot are directly indicative of the latest developments in Iran. Although these phrases were the only ones on the most frequent list, our attention was also drawn to words such as War, Weapon and Russia, which we linked to the ongoing conflict in Ukraine.



The most frequent bigrams from negatively labelled tweets

Equivalence of the mentioned bigrams indicates their usage within one massively repeated message. By grouping these negative tweets based on their content, we were able to investigate this particular message.

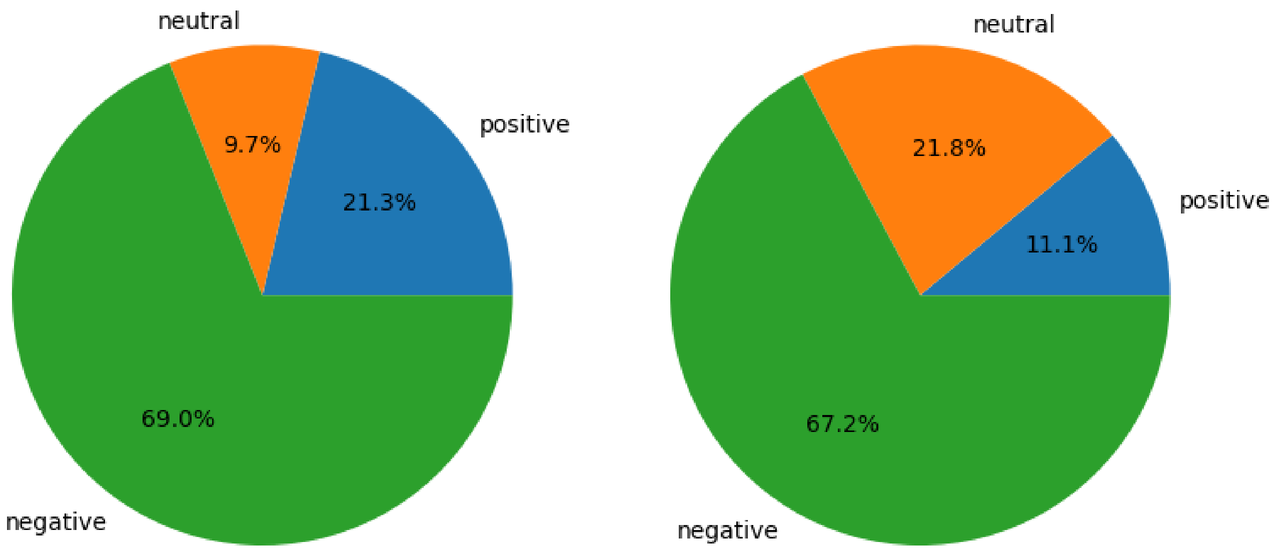
> *“At this time when the world is busy with football the islamic republic is committing crimes and killing innocent people and children with a war weapons in the street please be our voice”*

Few more examples:

> *“Islamists’ republic has nothing to do with Iran’s rich history…”*
> *“Iran's football team does not represent the people of Iran…”*

> *“We have stopped them… Zelensky speaks about Ukraine’s progress in war with Russia”*
> *“War, Peace and Politics…”*

As the analysis of the negative tweets yielded interesting results, we decided to additionally analyse those related to events in Iran and Ukraine. The search was conducted by using hashtags respectively “mahsaamini”, “javanroud”, “iranrevolution”, “saytheirnames”, “opiran” and “ukraine”, “russia”, “stop_war”, “ukrainerussiawar”, “war”.



Sentiment plotts for Iran (left) and Ukraine (right) related tweets

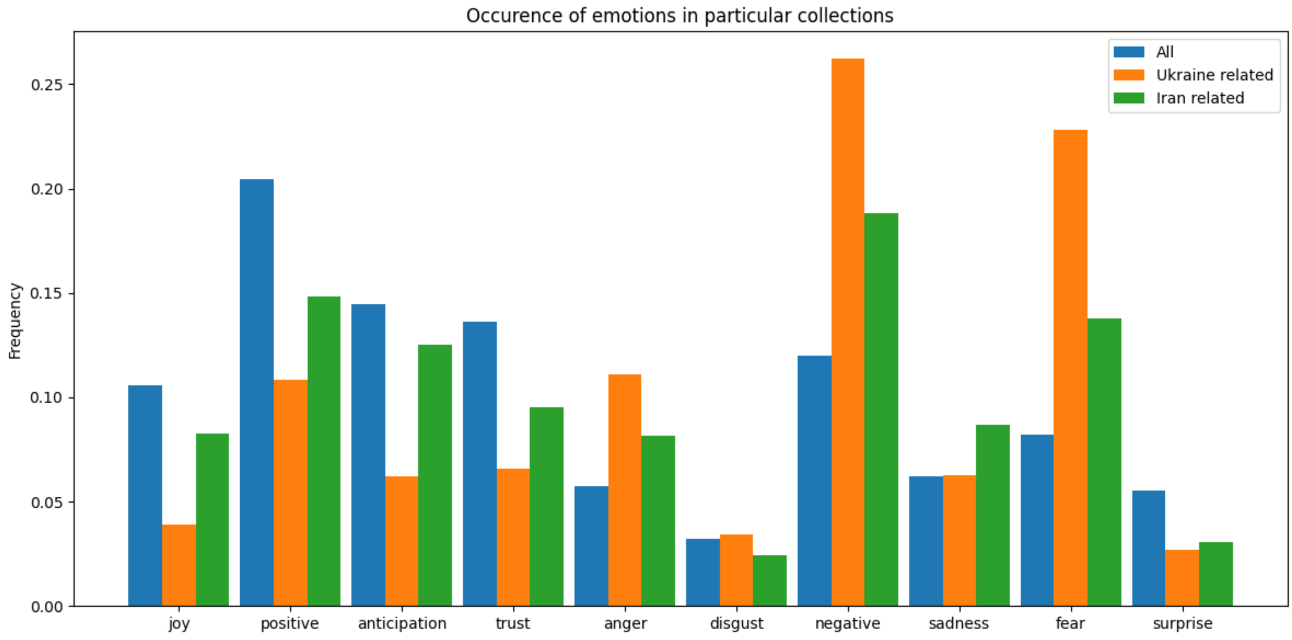
Both sets related to events score similarly in negative sentiment scores. Having examined tweets marked as positive, we conclude that those were mostly mislabelled by the tool or were blindly used by unaware users in order to share their messages further.

However, there are a few exceptions in the form of thanks for kind words or satisfaction with the progress

> *“GREAT VICTORY! Ukrainian Army Liberates Kreminna…”*
> *“Great news to Ukraine…”*
> *“Thank you England for being our voice… #MahsaAmini”*

Emotion analysis

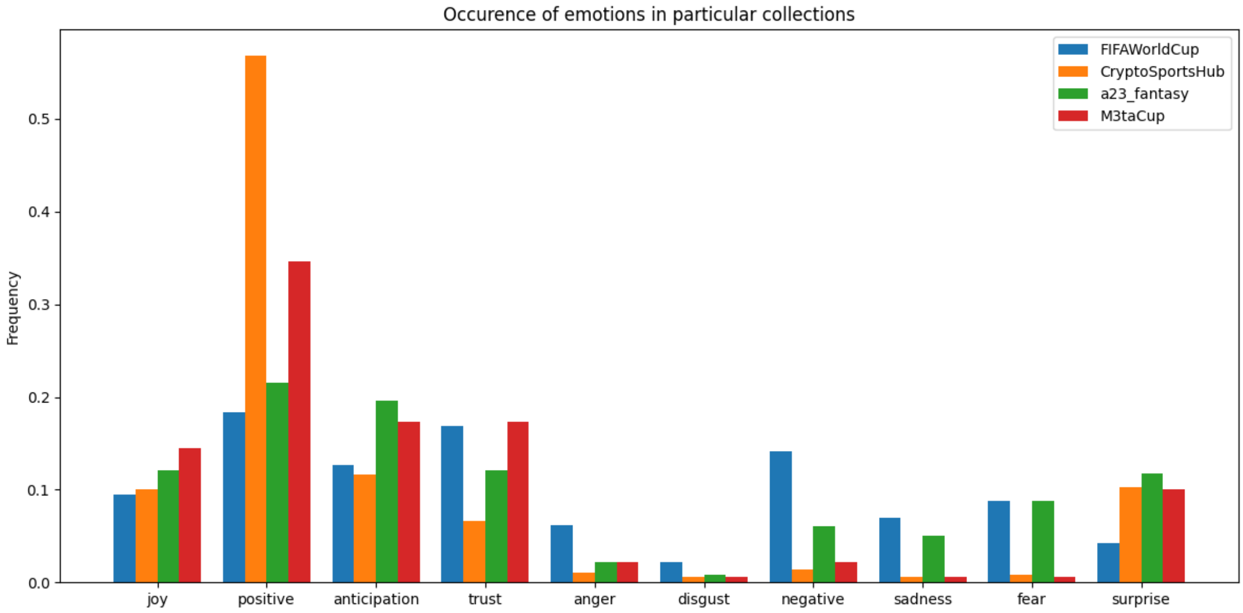
With the help of the NRC Emotion Lexicon (NRCLex) we have been able to conduct emotion analysis on our data. As examining tweets related to events in Iran and Ukraine returned unusual results, we have decided to include them in the following analysis.



Both Ukraine and Iran related tweets obtain dominant values for the negative, fear and anger emotions, which is no surprise considering tragic events to which they refer. High rate of the positive and trust emotions for the whole tweet set may be the result of a significant number of messages with an advertisement or competition structure. The relatively high anticipation emotion value for the tweets related to Iran might come from requests to be heard and support from other countries.

Emotions in different communities

As we have encountered four communities during network analysis, we have decided to apply the same NRC tool on each of them in order to better understand their structures.



The high bar for both second and fourth communities relates to their main characteristic, which is participation in competitions.

Community analysis

The biggest community is well described by its name because there are all the users posting about Mundial... but not only. We saw on the previous graphs that it also contains Iran-related tweets. We observed that the most centralised nodes are common for the whole graph and the red community, which shows how magnificent part of the network it is.

Despite the topic differentiation, we can see that it is the one with the less positive tweets and the most negative ones. All the others are way more biassed and focused on cryptocurrencies and giveaways, and these are always full of fake positivity. The leader is the Crypto Sports Hub community which also contains the famous YourChaifully (it does not have the highest degree because of graph reduction).

Example of CryptoSportsHub tweet:

> *“WORLD CUP CARNIVAL: DAY 22*
> *🥉 The battle for third place!*
> *PREDICT, EARN POINTS AND WIN*
> *To earn point, predict the correct score to be part of a 100 USDT prize pool💰”*

Investigation of node closeness centrality for each community all, by simply checking the profiles on Twitter, exhibited the same topics of interest. For example, a profile called Styroa Eleanor, who had the second highest centrality in the blue community (the first one is an account that does not exist anymore so it could have been a BOT) has a bio: “Crypto Enthusiast”.

The a23_fantasy and M3TACup were hosts of competitions, first encouraging people to predict the score of the game and the second to do joggling and record it. People, who tend to take part in that kind of challenge, usually don't end up with one, therefore the groups of mentions grow. The cryptocurrencies topic and related competitions are attracting them successfully.

Conclusion

Our main objective was to explore the general feeling about the sporting event, however, this task proved to be much more challenging. At every step we had to take into account the significant amount of advertising content in our data. After analysing the communities, we came to the conclusion that three of the four are deeply related to cryptocurrencies and contests. In-depth observation of positive tweets showed that the vast majority of users are satisfied with the individual matches or the performance of their team. However, there are exceptions in the form of entries expressing admiration for the culture there and the enormity of the event itself. The analysis of negative content proved much more fruitful as, in addition to tweets expressing dissatisfaction with the main event and the controversy surrounding it, we find those referring to the tragic events in Iran and Ukraine. Many of these messages have been repeated many times, however, this confirms our belief that such an action was intended to give the best possible publicity to a particular cause.

To conclude, the vast majority of the audience is critical of the event, with speakers pointing to numerous deaths of workers during preparations and human right violations. The number of people expressing satisfaction with the event is negligible. The popularity of the event has been exploited in various ways, mainly for publicity purposes, but also to publicise certain tragic events such as the crisis in Iran and Ukraine.