Text Mining & Search project

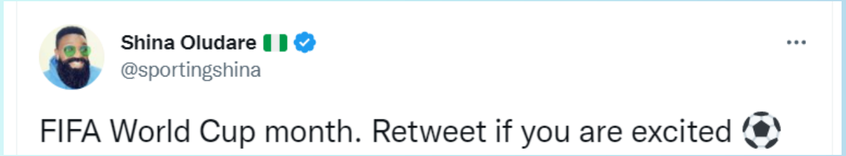
TEXT MINING OF TWITTER DATA FOR QATAR WORLD CUP

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

OUTLINE

- Introduction
- Data acquisition
- Preprocessing
- Text clustering
- Text summarization
- Conclusions

INTRODUCTION



- Highly **polarized** event with different opinions.
  - Amazing tournament.
  - But many controversial episodes.

- Entirely new and recent dataset.

- Analyzing tweets by means of **text clustering** and **text summarization**.
  - Grouping tweets into a discrete number of clusters with similar properties.
  - Generating summaries regarding the main hashtags.

DATA ACQUISITION

- Data obtained from Twitter by **scraping** (snscrape.modules.twitter library).

- Specific **query** for tweets.
  - Gathering tweets with most popular hashtags regarding the tournament.
  - Only English language.
  - Minimum of 5 likes.
  - Up to December 19, 2022.

- Tweets entered into a **dataframe**, with the following features:
  - Text, username, likes, retweets and timestamp.

- Final dataset composed by **151012** tweets.

PREPROCESSING - 1

- 2 main steps to make the dataset suitable for both **natural language** and models requiring **further pre-processing steps** (e.g. lemmatization).

- **Hashtags**, **mentions** and **URLs** removal.
- Extra-space removal and case folding.
- Repeated character removal.
- Removed those tweets that provide the **streaming** of the matches.
- Kept only those tweets with more than **10 likes**.
- Removed **symbols** not used in natural language (i.e. we kept punctuations).
- **English** language control.
- Removed tweets with less than **10 words**.
- Duplicated tweets removal.

Preprocessed dataset:

89261 tweets

PREPROCESSING - 2

- Further preprocessing required for text representation methods such as **TF-IDF** and **Doc2Vec** embedding.

- **Demojizing**: emoji have been transformed into the respective text.
- Numbers removal.
- Contractions fixed (e.g. from "you’re" to "you are").
- Punctuation removal (we kept the ’_’ symbol because it’s the delimiter for the emoji).
- **Tokenization** (nltk TweetTokenizer).
- **Lemmatization**.
  - More meaningful then stemming.
  - Tweets by themselves are difficult to treat due to their high level of insignificance.
- Stop-words removal.

TEXT CLUSTERING

Three different kinds of embedding technique:

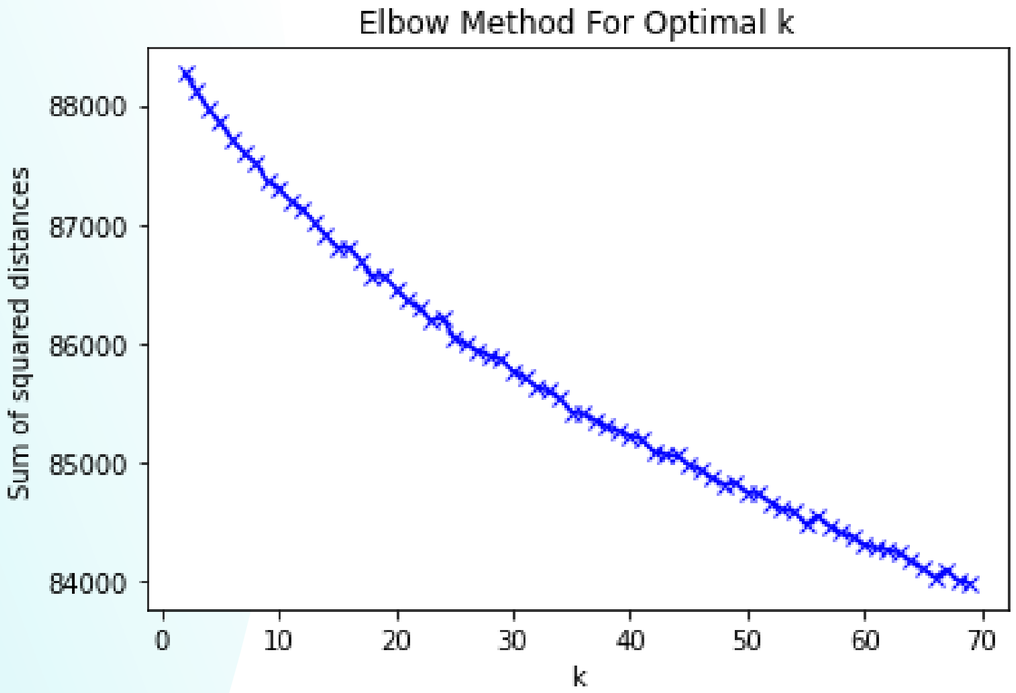
- **TF-IDF**.
- **Doc2Vec**.
- **BERTweet**.

Two different kinds of clustering approach:

- **K-means**.
- **Hierarchical**.

TF-IDF

K-MEANS

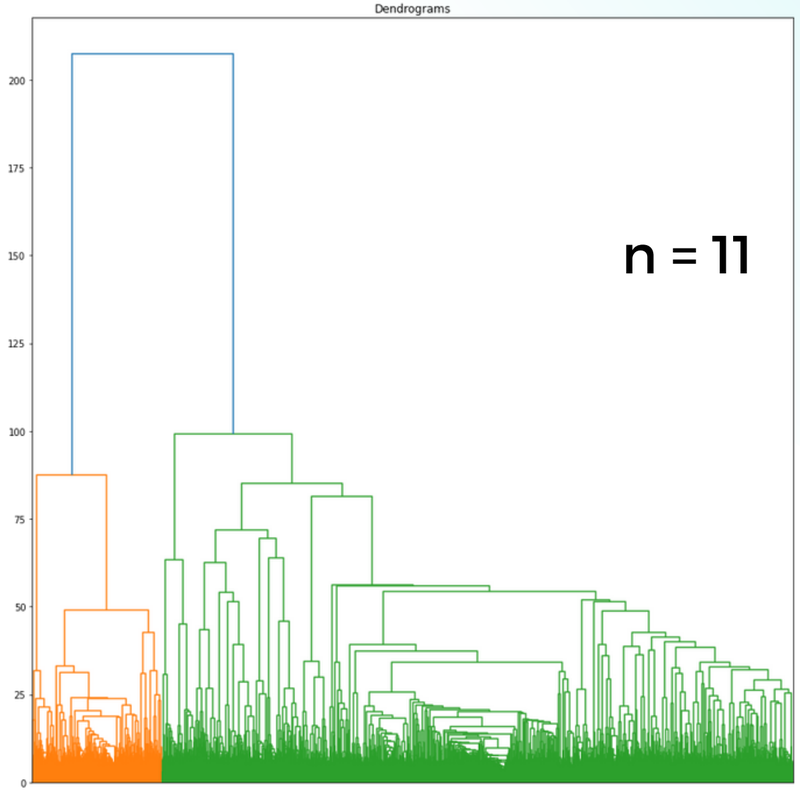


K = 40

Silhouette coefficients:

- Euclidean distance = 0.0093
- Cosine distance = 0.0176

HIERARCHICAL CLUSTERING

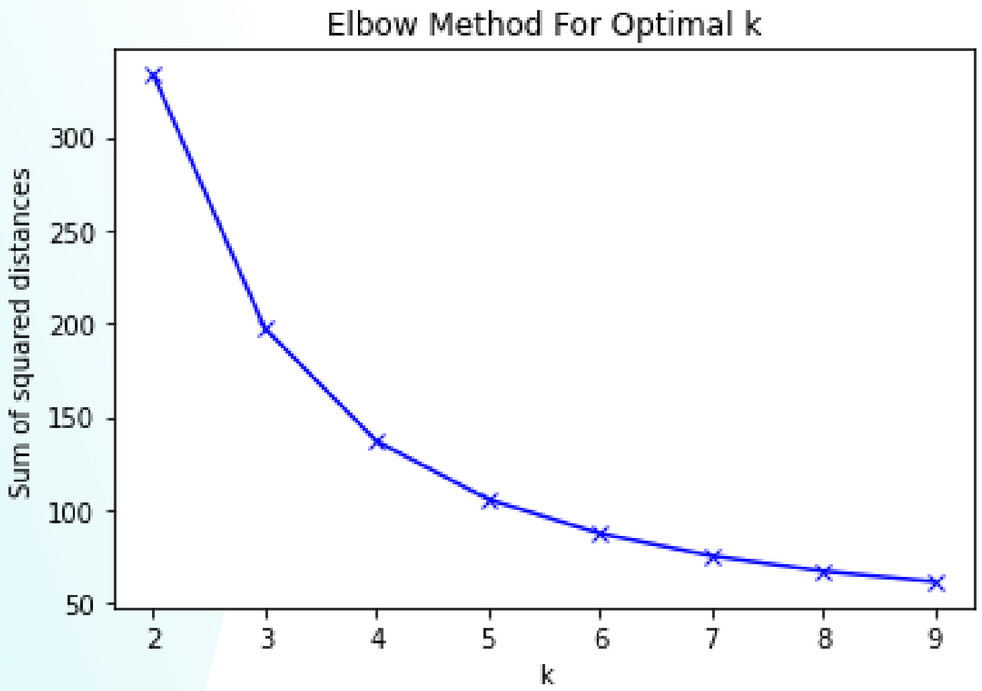


Silhouette coefficients:

- Euclidean distance = 0.0039
- Cosine distance = 0.0070

DOC2VEC

K-MEANS

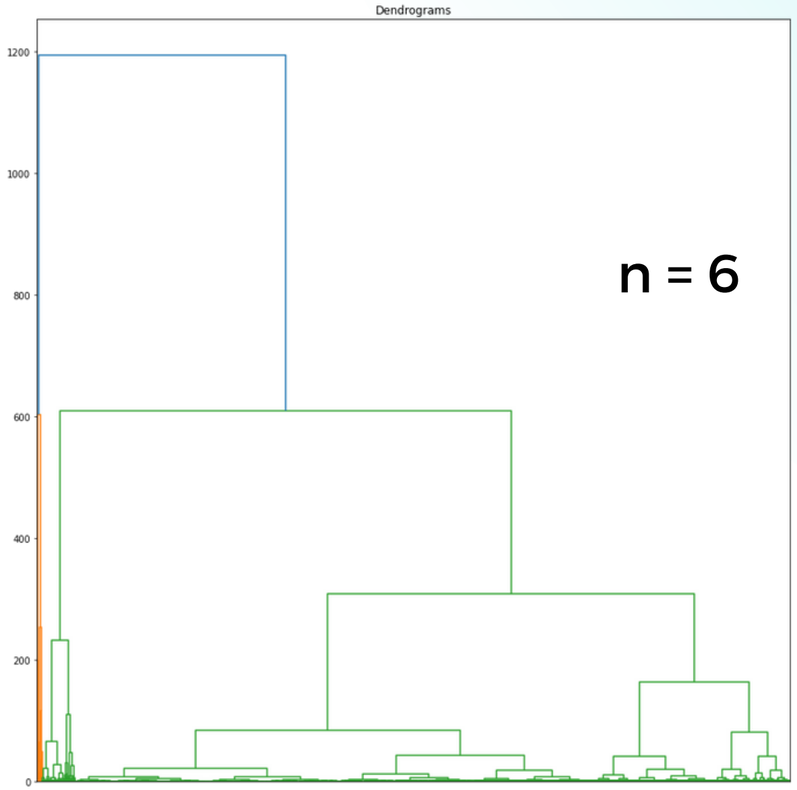


K = 4

Silhouette coefficients:

- Euclidean distance = 0.4097
- Cosine distance = -0.0841

HIERARCHICAL CLUSTERING

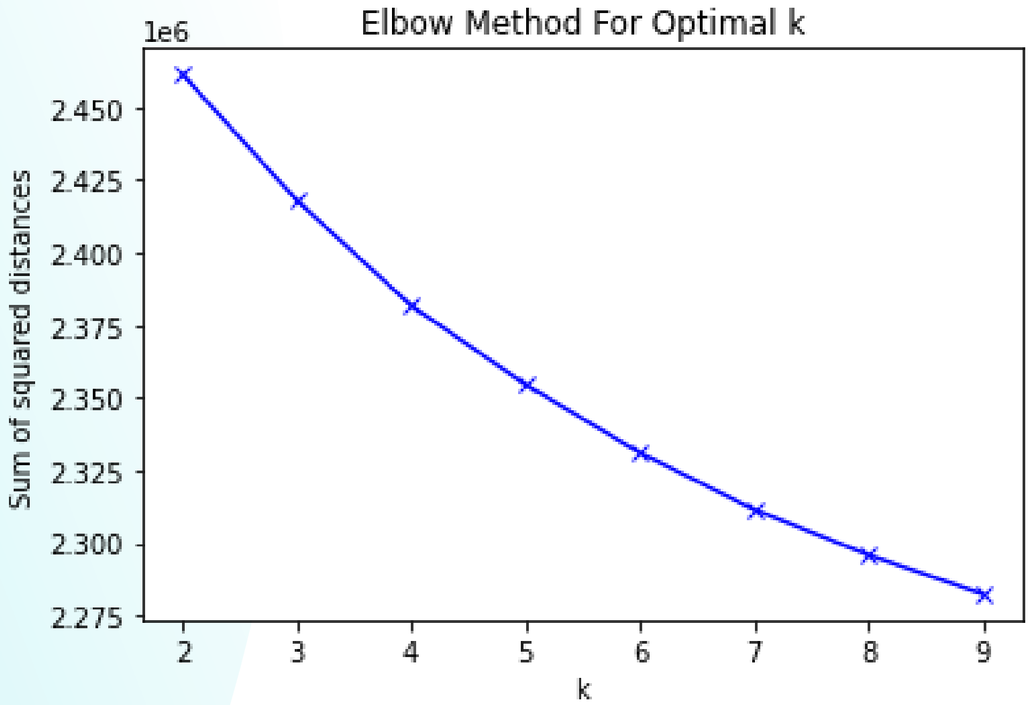


Silhouette coefficients:

- Euclidean distance = 0.2019
- Cosine distance = 0.3115

BERTWEET

K-MEANS

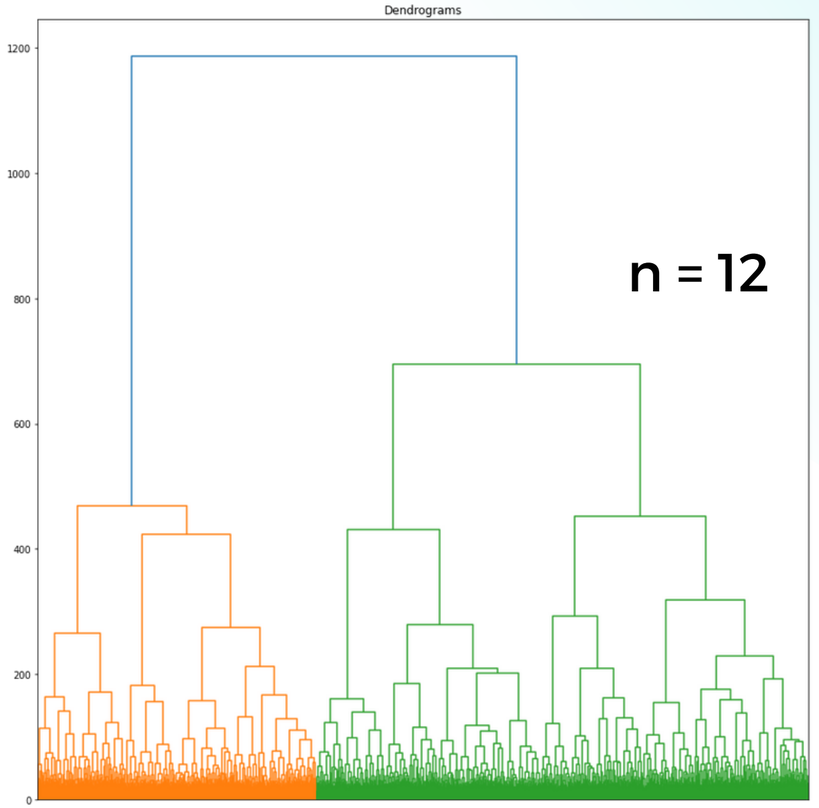


K = 5

Silhouette coefficients:

- Euclidean distance = 0.0229
- Cosine distance = 0.0374

HIERARCHICAL CLUSTERING



Silhouette coefficients:

- Euclidean distance = -0.0117
- Cosine distance = -0.0206

TEXT CLUSTERING

SUMMARY EVALUTATION

We also empirically evaluated the clusters by extracting a summary for each cluster.

- **Abstractive** summarization.
- **BART** pre-trained model.
- Several interesting summaries concerning:
  - Negative opinions.
  - Positive opinions.
  - General statistics and episodes.

- Practical way to assess **significance** of clusters (together with silhouette coefficients).

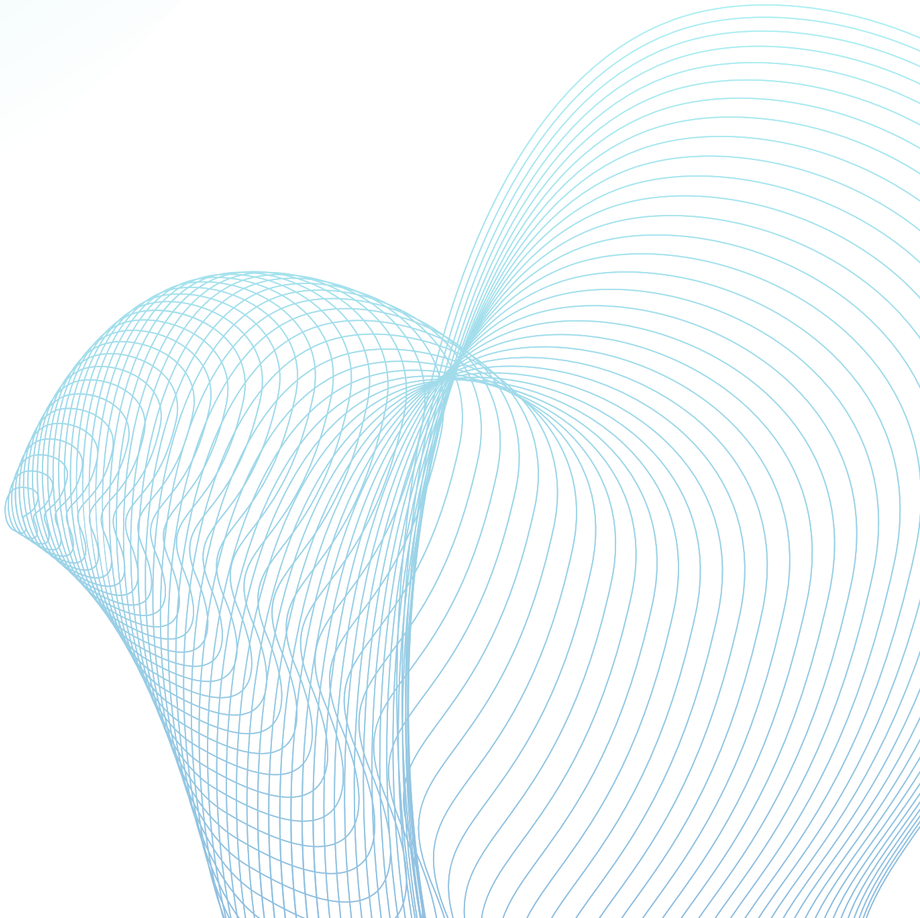
TEXT SUMMARIZATION

Grouped the tweets based on the following 8 hashtags:

- **#Qatar2022**
- **#SayTheirNames**
- **#BoycottQatar**
- **#Messi**
- **#Argentina**
- **#France**
- **#Croatia**
- **#Morocco**

Two different approaches :

- **Abstractive** Summarization.
- **Extractive** Summarization.



TEXT SUMMARIZATION

ABSTRACTIVE APPROACH

We used this technique with two pre-trained models:

- **T5** (Text-to-Text Transfer Transformer).
- **BART** (Bidirectional Encoder Representations from Transformers).

Problems and limitations:

- Informal language.
- Noise.
- Lack of training data.
- Lack of domain-specific knowledge.

TEXT SUMMARIZATION

ABSTRACTIVE EVALUATION

- Human assessments.

T5

#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.
#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 .

Highlighting Blue: repetitions and those parts not relevant to the specific hashtag.
Highlighting Red: parts that do not make sense or wrong.

BART

#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 [...]
#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 [...]

TEXT SUMMARIZATION

EXTRACTIVE APPROACH

We used two algorithms based on the idea of **graph-based** centrality:

- **LexRank**.
- **TextRank** with three text representations:
  - **TF-IDF**.
  - **Doc2Vec**.
  - **BERTweet**.

TEXT SUMMARIZATION

EXTRACTIVE EVALUATION

- Human assessments.
- Score of **sentences** for LexRank.
- Score of **tweets** for TextRank.

		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 1: Top-4 scores for extractive methods

CONCLUSIONS

- Novel dataset composed by **tweets**.
- **Specific preprocessing** with respect to different **models** and **embeddings**.
- **Clustering** tweets according to two different algorithms (k-means + hierarchical) and three different representations (TF-IDF, Doc2Vec, BERTweet).
- **Summarization**: abstractive (T5 + BART) and extractive (LexRank + TextRank with three different representations).
- The application of these two tasks showed the **extreme bias** of the event.

THANKS FOR ATTENTION