

# EasySport Understanding

Artificial Intelligence final project

Lorenzo Pirola - mat. 816418 Matteo Romanato - mat. 816852 Youssef Karrati - mat. 817435

# **Outline**

**Introduction** 

- 2 ) Dataset
- (3) NER & NEL

4 Word Embeddings

- 5 Results
- 6 Web Application
- **7** Conclusions

# 1 Introduction

# What is the problem?

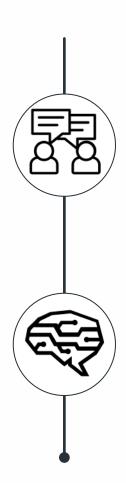
People found it difficult to follow new sports due to a lack of knowledge



# Solution

An application where people can improve their knowledge about sports exploiting common knowledge retrieved from the web





# Idea

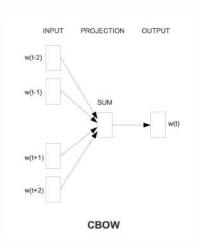
### **Distributional Representations**

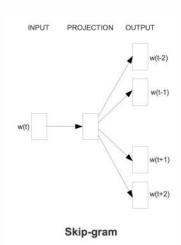
Represent words by n-dimensional dense vectors derived from the usage of words in some text corpus

# **Background Work**

# Efficient Estimation of Word Representations in vector Space

"We propose two novel model architectures for computing continuous vector representations of words from very large data sets"





# **Background Work**

### Compass-aligned Distributional Embeddings for Studying Semantic Differences across Corpora

"We present a general framework to support cross-corpora language studies with word embeddings, where embeddings generated from different corpora can be compared to find correspondences and differences in meaning across the corpora"

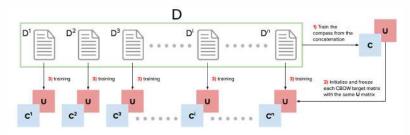
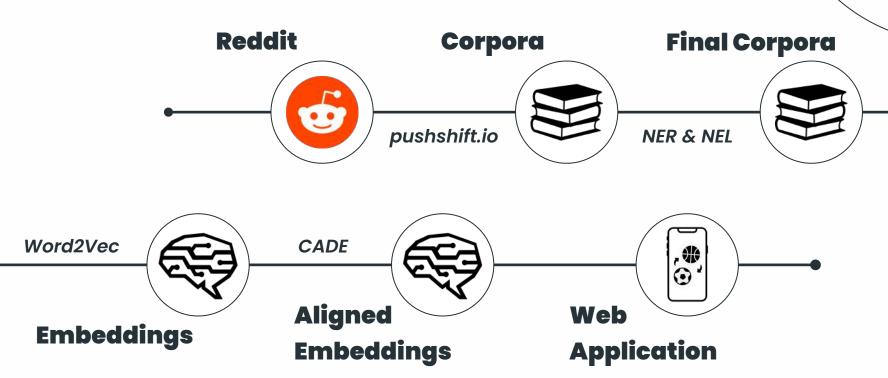


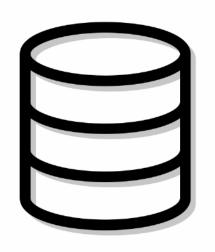
Figure 5: The CADE model.

# **Methodological Approach**



# 2 Dataset

# **Dataset**



### **Subreddits**



r/soccer



r/nba

### Seasons

- 2015/16 (31/07/2015 31/07/2016)
- 2017/18 (31/07/2017 31/07/2018)
- 2019/20 (31/07/2019 31/11/2020)

# **Preprocessing**

- Punctuation and Stop Words removal
- Emoji:

• Numbers:

The Golden State Warriors (7-6)

defeat the Los Angeles Lakers (11-4)

The Golden State Warriors defeat the Los Angeles Lakers

• Acronyms:

Curry in tonight's win: 8/25 **FG**, 5/20 Curry tonight win **field goals** points from 3, 25 points

NER & NEL using DBpedia Spotlight with confidence equal to 0.6

For each entity were retrieved and saved (if present):

- Its **foaf:name** in English
- Its dbo:abstract only first sentence
- Its **rdf:type** (most specific)





8859

Total entities for **soccer** 



5736

Total entities for **basketball** 

# Issues

### Ambiguity and polysemy:

**Antonio Conte** signs year deal Chelsea

germany vs italy friendlies

antonio\_del\_pollaiolo conte signs year deal chelsea\_f.c.

gianluigi\_buffon play game
italy\_national\_football\_team tonight
all-time record azzurri

• Missed entities:

Diego Costa **red card** incident

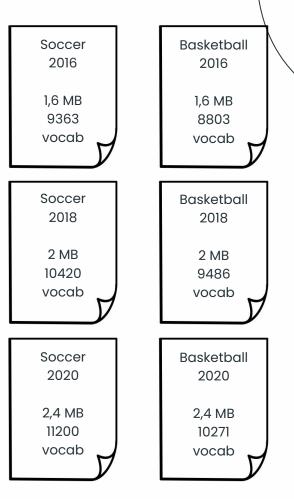
diego\_costa **red card** incident

# 4 Word Embeddings

# **Word Embeddings**

For each sport and season (plus sports in general):

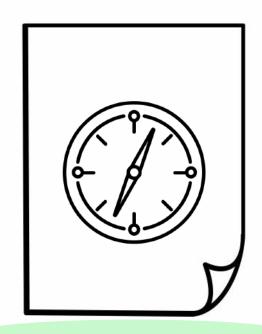
Word Embeddings using Word2Vec



# **Word Embeddings**

For each sport and season (plus sports in general):

- Word Embeddings using Word2Vec
- Embeddings alignment with CADE



# **Evaluation**

Without a ground-truth database, the obtained embeddings were evaluated using some analogies and our knowledge Evaluation was based on different concepts:

Teams

Moves

Players

Awards

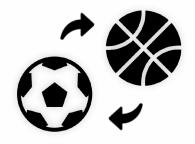
Coaches

# **Evaluation**



### **Intra-Corpora Analogies**

- Similar entities within the same corpus
- To expand user's knowledge about a known sport



### **Cross-Corpora Analogies**

- Similar entities between two different corpus
- To understand a new sport



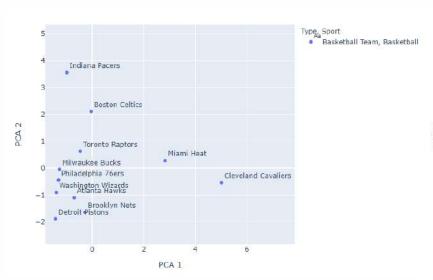
### **Short-term Meaning Shift**

 To study the evolution of an entity over different seasons

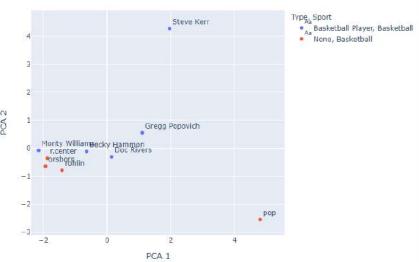
# 5 Results

# Intra-Corpora Analogies

### **Most similar to Boston Celtics**

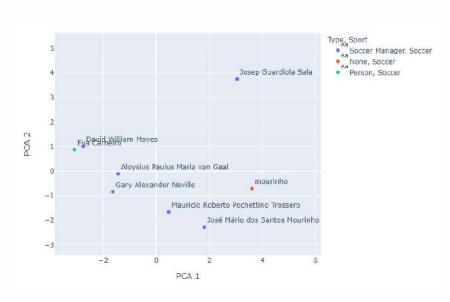


# Most similar to Steve Kerr - GSW + Spurs

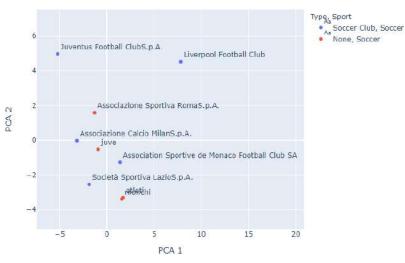


# Intra-Corpora Analogies

### Most similar to José Mourinho



### Most similar to F.C. Liverpool - EPL + Serie A

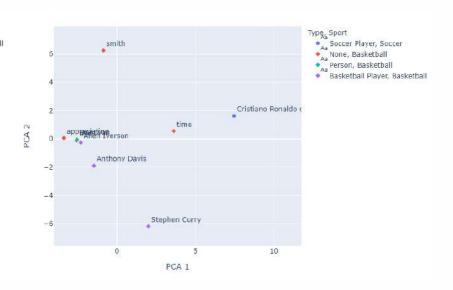


# Cross-Corpora Analogies

### **Stephen Curry in Soccer**

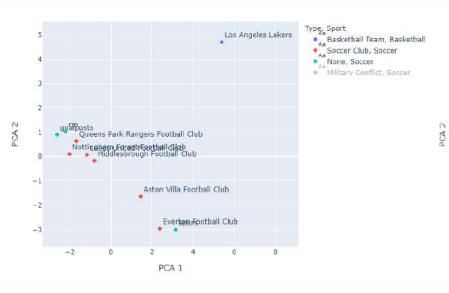
# Type, Sport As Basketball Player, Basketball As Album, Soccer As Soccer Player, Soccer As None, Soccer None, Soccer As David Oxlade-Chamberlain Dated Andre Sturridge O 5 10 PCA 1

### Cristiano Ronaldo in Basketball

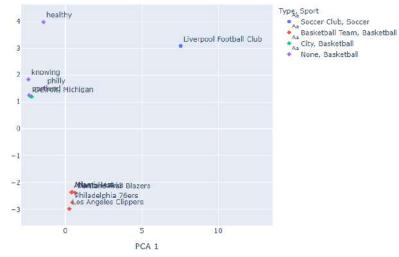


# Cross-Corpora Analogies

### Los Angeles Lakers in Soccer



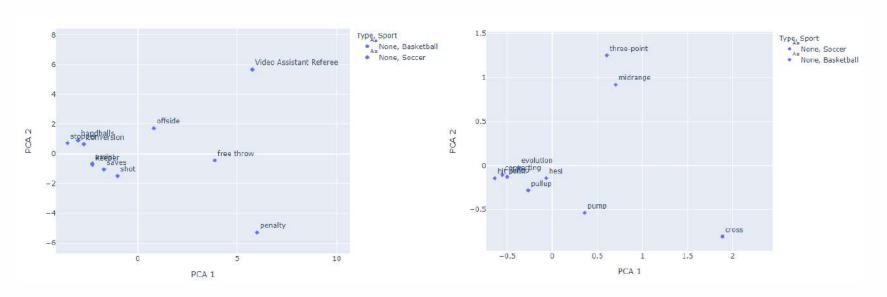
### F.C Liverpool in Basketball



# Cross-Corpora Analogies

### Free throw in Soccer

### **Cross in Basketball**



# **Short-term Meaning Shift**

### **Cleveland Cavaliers in 2016**

### **Cleveland Cavaliers in 2018**

### **Cleveland Cavaliers in 2020**

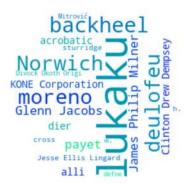
NBA Championship Detroit Pistons Philadelphia 76ers Atlanta Hawks Los Angeles Clippers Cleveland Cavaliers Houston Rockets
Golden State Warriors Miami Heat Chicago Bulls lossless Dallas Mavericks NBA Finals Brooklyn Nets cavaliers



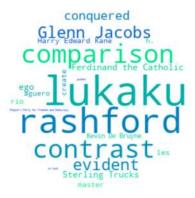
Denver Nuggets snapping Washington Wizards Cleveland Cavaliers New Orleans Pelicans Detroit Pistons Phoenix Suns Chicago Bulls
Los Angeles Clippers Anthony Tolliver Toronto Raptors blowing

# **Short-term Meaning Shift**

### Lukaku in 2016



### Lukaku in 2018



### Lukaku in 2020

Maurizio Sarri
Lukaku

Cagliari Calcios.p.A.

Detrayed

cellino
Comune di Milano
italiano, lingua Martinus Brands

Martinus Martinus Brands

mixed Sarri

mixed Sarri

# **Short-term Meaning Shift**

### Orlando Magic in 2016

Charlotte Hornets

Brooklyn Nets

Dallas Mavericks

Sacramento Kings
Philadelphia 76ers torusto believes

Washington Wizards
Utah Jazz

Orlando Magic
Phoenix Suns Laugus lossless

Atlanta Hawks
Detroit Pistons
Boston Celtics

Los Angeles Clippers

Memphis Grizzlies

### Orlando Magic in 2018

Dallas Merinana Portlands Mavericks Mean Minnesota Lynx Chicago Bulls Atlanta Hawks

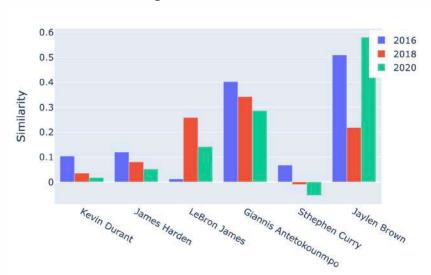
Orland Orlans Pelicans Orland Charlotte Hornets Ocharlotte Hornets Miami Heat

### Orlando Magic in 2020

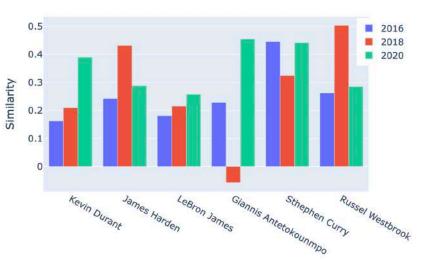
```
training
bubble dec.
july june grampus
approvedprotocols prior
Orlando Magic
coronavirus outbreakf
schedule
restart friday
wednesday
```

# Temporal Analogies

### **Similarity with Boston Celtics**

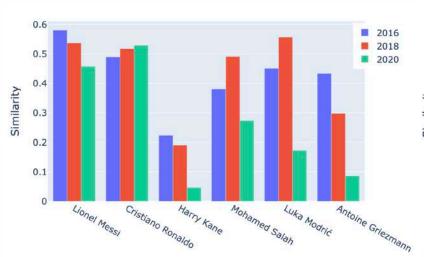


### Similarity with MVP

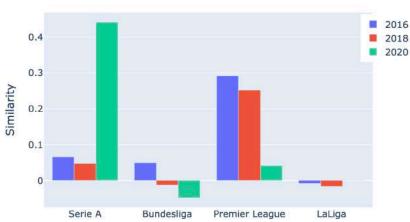


# Temporal Analogies

### Similarity with Ballon d'Or



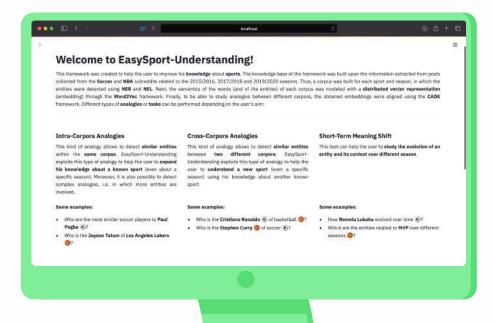
### Similarity with Romelu Lukaku



# 6 Web Application

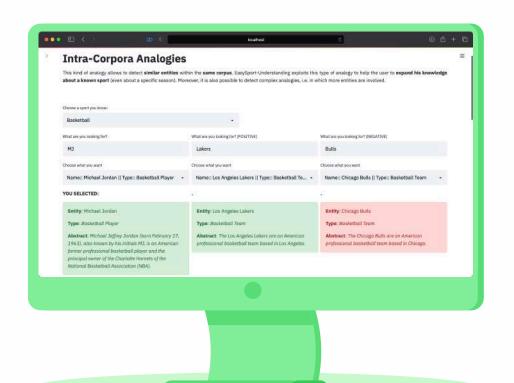
# **Application**

- Python
- Streamlit framework



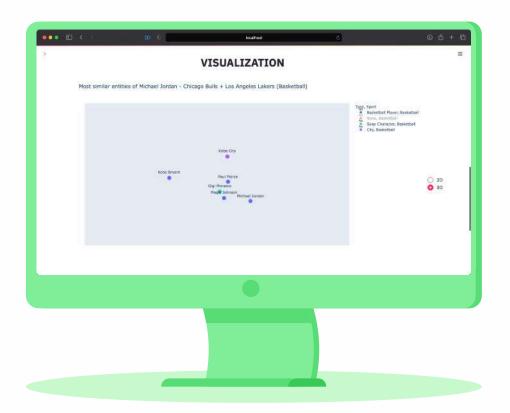
# Intra-Corpora

- Choose a sport
- Choose an **entity**:
  - Optional positive entity
  - Optional negative entity
- Entities' info are shown



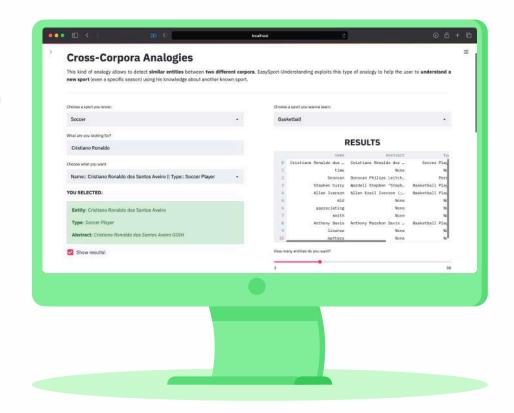
# Intra-Corpora

- Results are shown in a table
- 2D Visualization
- **3D** Visualization



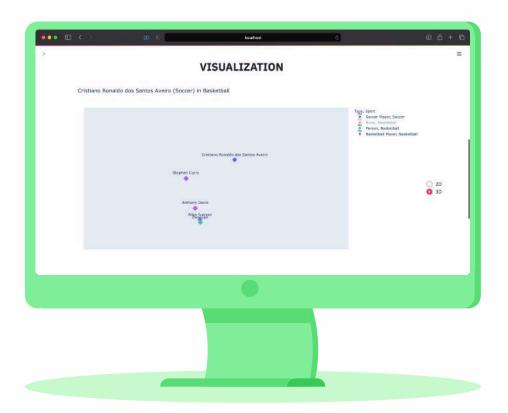
# Cross-Corpora

- Choose a known sport
- Choose an unknown sport
- Choose an entity
- Entity **info** are shown



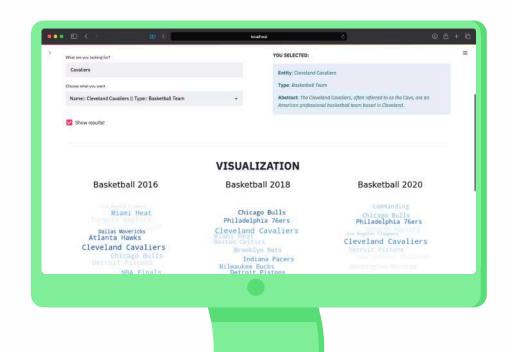
# Cross-Corpora

- Results are shown in a table
- 2D Visualization
- 3D Visualization



# **Meaning-shift**

- Choose a sport
- Choose an entity
- Entity' **info** are shown
- Word Cloud visualization



# 7 Conclusions

# Conclusions

### **Considerations**

- Good results but not optimal, ambiguity is the main problem
- Low quality sources
- Missing entities

### **Future studies**

- More sports
- Test other approaches
- Better NER & NEL (different techniques or confidence)



# References

- Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).
- Bianchi, Federico, et al. "Compass-aligned Distributional Embeddings for Studying Semantic Differences across Corpora." arXiv preprint arXiv:2004.06519 (2020)

# Thank you!

# Let's try a demo!

