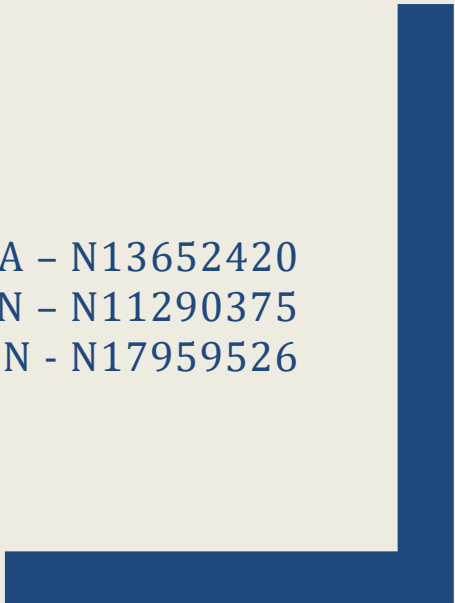




Dealing with Data- Project Report

# FOOTBALL TRANSFER RUMOR VALIDATION

ADITYA VASHISTHA – N13652420  
SAURABH SHARAN – N11290375  
SIDDHARTH JAIN - N17959526



## **ABSTRACT**

Rumors/ Fake news, defined by the New York Times as “a made-up story with an intention to deceive”, often for a secondary gain, is arguably one of the most serious challenges facing the news industry today. In a December Pew Research poll, 64% of adults said that “made-up news” has caused a “great deal of confusion” about the facts of current events.

We are living in a time where we have more information available to each of us than ever before in history. However, we are not all proficient at distinguishing between real information and fake news. Rumors/ Fake News is a Real problem.

For our python project, we are working on addressing variety of news related to football (soccer) player transfers and which news source amongst our identified ones are the most trustworthy and reliable when it comes to such speculations.

Fake news is getting more and more influential through social media and too many news sources which is misleading many times. In various football leagues, players transfer happens every year. So, fake player transfer rumors numbers are increasing.

## **MOTIVATION**

Dubiously sourced rumors about football transfers spread wildly on social media, and while experts say they don't usually affect where players end up, they can put pressure on clubs and move betting markets.

We are passionate football fans and closely watch premier league. In this digital world, we interact with all kinds of platforms to access information. It could be news websites like BBC, ESPNFC, Goal.com, etc., news aggregators and social media platforms like Facebook, twitter, Instagram and so on.

There are millions of football followers across the globe just like us who read information of transfers on daily basis. Football player transfers are highly speculated, so it is very likely that some news sources publish the right information more often than the others.

## **PROBLEM STATEMENT**

To find the most reliable sports website when it comes to player transfer during a football transfer window.

## **GOALS**

- Accurately classify fake/ real player transfer news
- Analyze which web sources are more reliable than the others
- Visualize the results

## **CLUBS**

- Arsenal
- Liverpool
- Manchester City
- Manchester United

These clubs belong to the English Premier League and we selected these clubs because they are the most widely followed clubs in the world.

## **NEWS SOURCES**

Our main source of news article was the NEWS API, which contained articles from around 30000 websites and blogs. Since scrapping articles from all these sources would've been a huge task, we went with a few of them and some of the more famous one's are listed below.

- ESPNFC
- BBCSPORT
- Daily mail
- Goal
- 101greatgoals
- Bleacher Report

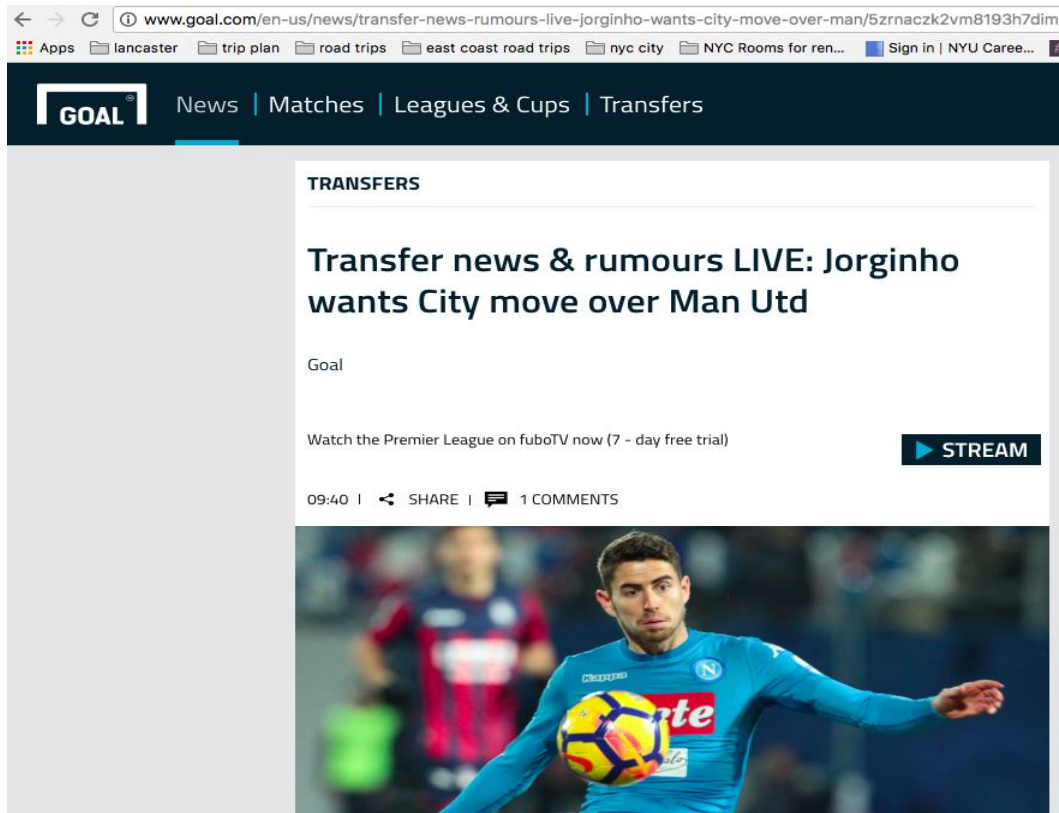
## **Football Transfer Windows**

In the football world, every year, there are two transfer windows in which a football club can transfer players from other clubs into their playing staff.

The transfer window frame is runs every year between January 01 to January 31 and May 17 to August 9 in the English Premier League of United Kingdom.

We are targeting the January 01 to January 31 of 2018 as our designated transfer window for the project.

## Sample transfer news article



## ASSUMPTIONS

1. Articles that have news only about transfers
2. Articles that were published between 1<sup>st</sup> Dec 2017 and 31<sup>st</sup> Jan 2018 only
3. Scraped only the proper nouns in the articles and not specifically “to” and “from” in the article
4. If a player’s first or last name appears in an article description and it matches our current squad rosters after the transfer window ended, then we consider the whole article to be valid

## **DATA SOURCES AND TOOLKIT**

### **SportRadar**

Using SportRadar API, we get the current squad of Europe's top 96 teams which is then inserted into the player database. This will be used to do comparison with the news from the various sources.

<https://api.sportradar.us/soccer-xt3/eu/en/tournaments/>

### **News API**

The API that we are using to get data is called the News API. News API is a simple HTTP REST API for searching and retrieving live articles from all over the web.

<https://newsapi.org/v2/everything>

**Search for articles with any combination of following criteria,**

- **Keyword or phrase** -find all articles containing the query term (q)
- **Date published**- find all articles published between 2017-12-01 to 2018-01-31
- **Source name.** find all articles by ESPNFC, Goal,101greatgoals, etc.
- **Source domain name.** find all articles published on espnfc.com.
- **Language.** find all articles written in English.

### **Stanford Named Entity Recognition tagger (NER)**

It is an alternative to NLTK's named Entity recognition (NER). In the named entity recognition tagger, we used three class models for recognizing locations, persons, and organizations. We are using Stanford NER to recognize and extract only the player names from the news articles.

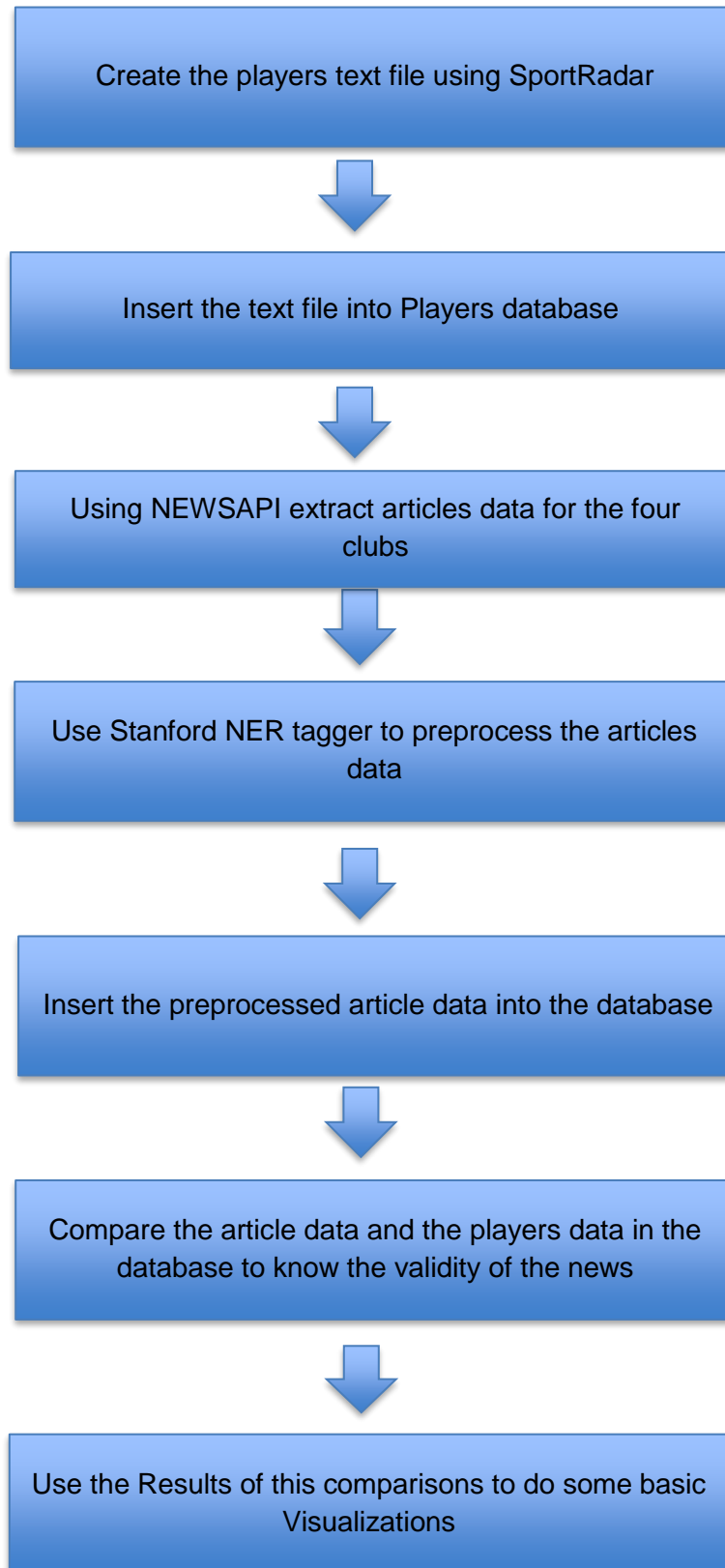
### **Python IDE**

- PyCharm

### **Libraries used**

- Pymysql
- Json
- Stanford NERTagger
- Bokeh

## DATA FLOW



## DETAILS

### Scraping from the Web

For every club, we are fetching data from 40 pages and on every page, we are taking the first 10 articles. So, we are fetching approximately 400 articles with the keywords “transfer AND football AND <club name>”, inserted dynamically, in a two-month time period.

```
for cname in club_list:
    try:
        for i in range(1, 40):
            parameters = dict(q='transfer AND football AND ' + cname, from_parameter="2017-12-01", to="2018-01-31",
                               language='en', pageSize=10, page=i,
                               apiKey="1101c8a20f244a82bebbaf45561ba8a4")
            r = requests.get(url, params=parameters)
            footballdata_dict = json.loads(r.text)
```

### Preprocessing on the articles

Once we get the response (r) from the API, for every club, we are using the Stanford NER to extract the “PERSON” from the description and “Source” from the embedded URL from the footballdata\_dict object, as shown in the code snippet above, and writing it to the club’s specific json. An example element of the json file is shown below,

```
{
  "PERSON": [
    "Luiz",
    "Giroud",
    "David",
    "Luiz",
    "Olivier",
    "Arsene",
    "Wenger"
  ],
  "Source": "goal"
},
```

## Creating JSON for every club

For every club, we are creating a json which will have the dictionary with keys “PERSON” and “SOURCE”. We have 4 Json files namely –

- Footballdata\_Arsenal.json
- Footballdata\_Liverpool.json
- Footballdata\_ManchesterCity.json
- Footballdata\_ManchesterUnited.json

## Storing it in databases

Database – MySQL

Tables –

### 1. “Players” table and attributes:

**ID (CHAR)** – Unique identifier

**Player\_Mapping (VARCHAR)** – Sports radar ID of each player

**Player\_name (CHAR)** – Name of the player

**Club (CHAR)** – Club of the player before the transfer window period

**Club\_Changed (YES/NO)** – Whether after the transfer window the player club got changed or not

ID	Player_Mapping	Player_name	Club	Club_Changed

### 2. “Articles\_<ClubName>” table and attributes:

**AID (CHAR)** – Articles Unique identifier

**Persons (CHAR)** – All the Player names extracted from the article

**Source (CHAR)** – News sources of the article

**Validity (1/ 0)** – After comparison, whether the article is valid or not

AID	Persons	Source	Validity



## Comparison

Articles table has articles only about the players for whom there were some news published during the transfer window. Therefore, for every article in the articles table, we compare each of the words in the person names with the players in the players table. This is accomplished via the 'LIKE' operator in SQL. Upon comparing, if any of the matched tuples from the "players" table has the "club\_changed" attribute as "YES" then we set the "validity" for the article as 1. Setting it to 1 means that the player transferred to a new club in 2018 and hence the article rumor was right. In the other case if all the "club\_changed" attribute values are "NO" then we set it validity to 0, which means that it was a false rumor.

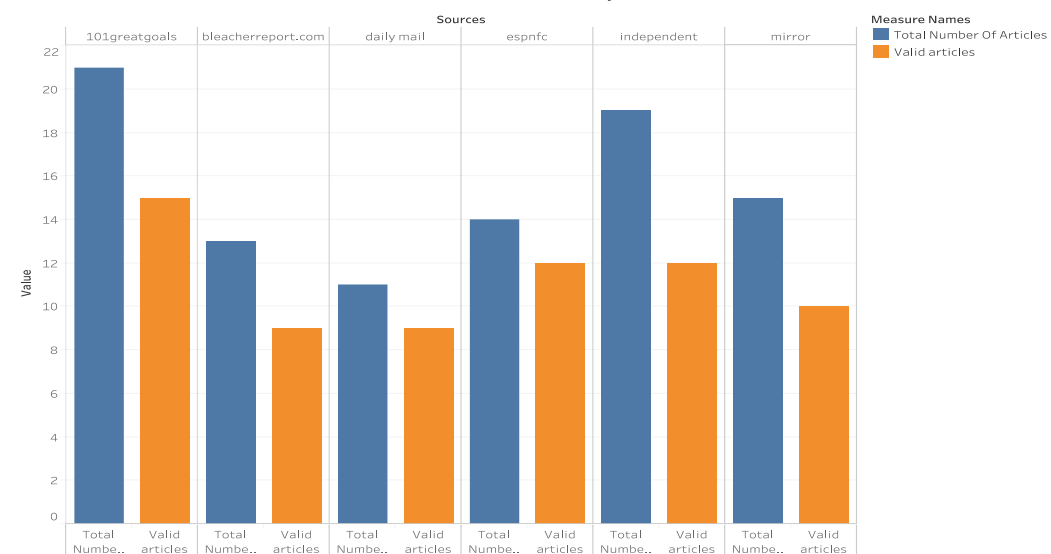
## RESULTS AND VISUALIZATIONS

### 1. Arsenal – Total number of articles vs Valid Articles per source

From the plot below, we can say the following –

- 101greatgoals, ESPNFC and Independent are the most reliable news sources for Arsenal transfer rumors
- Bleacherreport.com and Daily Mail are the most unreliable sources

Arsenal- Total number of articles vs Valid articles for every source



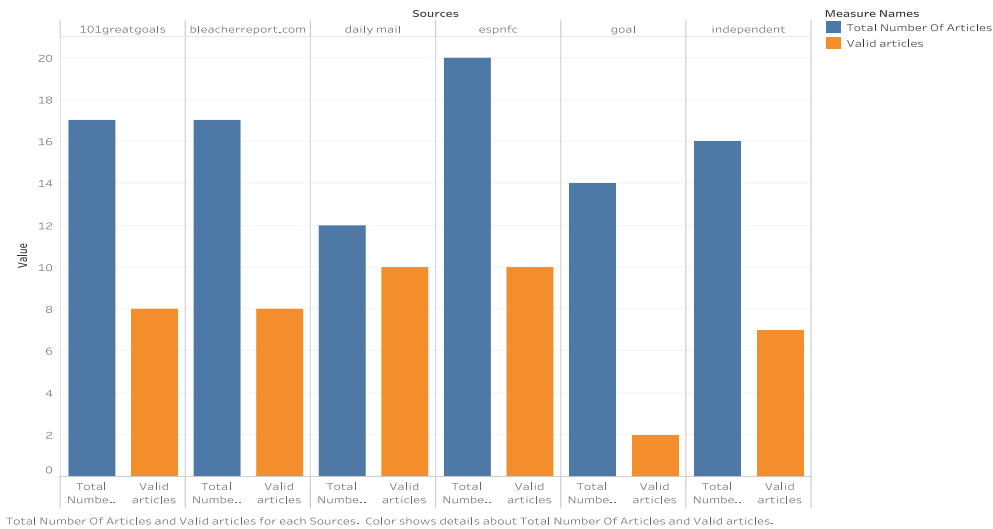
Total Number Of Articles and Valid articles for each Sources. Color shows details about Total Number Of Articles and Valid articles.

## 2. Liverpool - Total number of articles vs Valid Articles per source

From the plot below, we can say the following –

- ESPNFC and Daily Mail are the most reliable news sources for Liverpool transfer rumors
- Goal and Independent are the most unreliable sources

Liverpool - Total no of articles vs Valid articles for every source

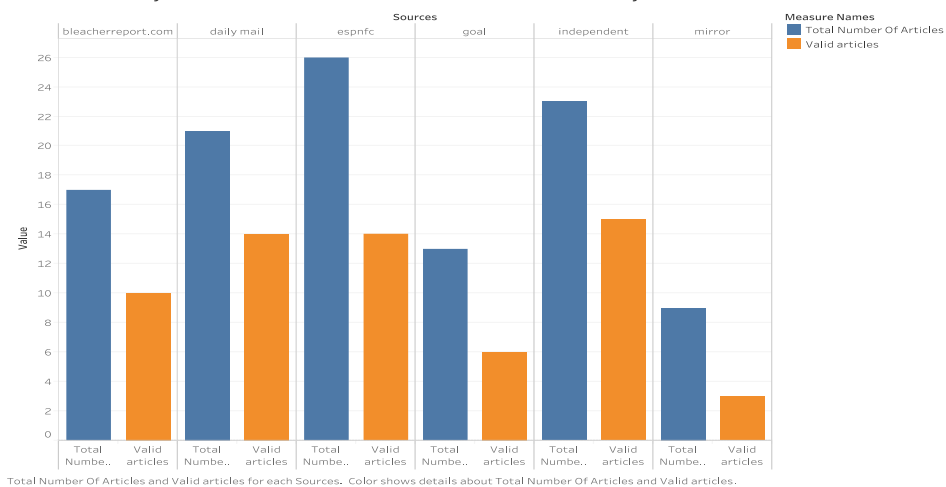


## 3. Manchester City - Total number of articles vs Valid Articles per source

From the plot below, we can say the following –

- Independent, ESPNFC and Daily Mail are the most reliable news sources for Manchester City transfer rumors
- Goal and Independent are the most unreliable sources

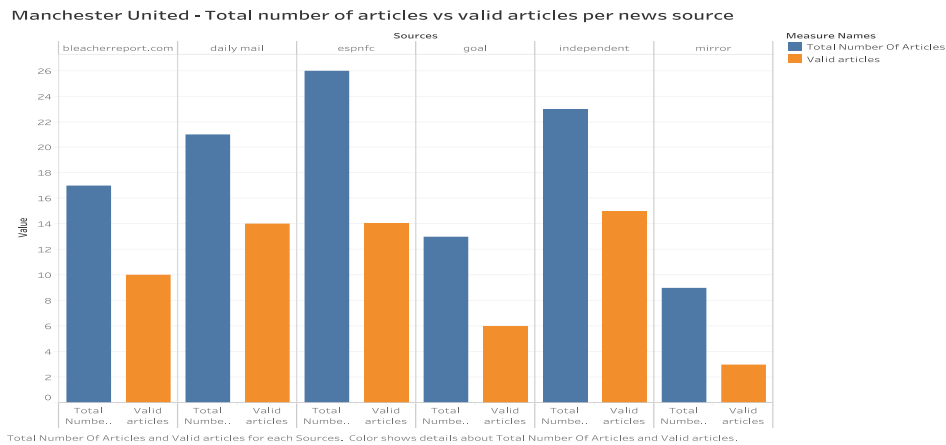
Manchester City - Total number of articles vs Valid articles for every news source



#### 4. Manchester United - Total number of articles vs Valid Articles per source

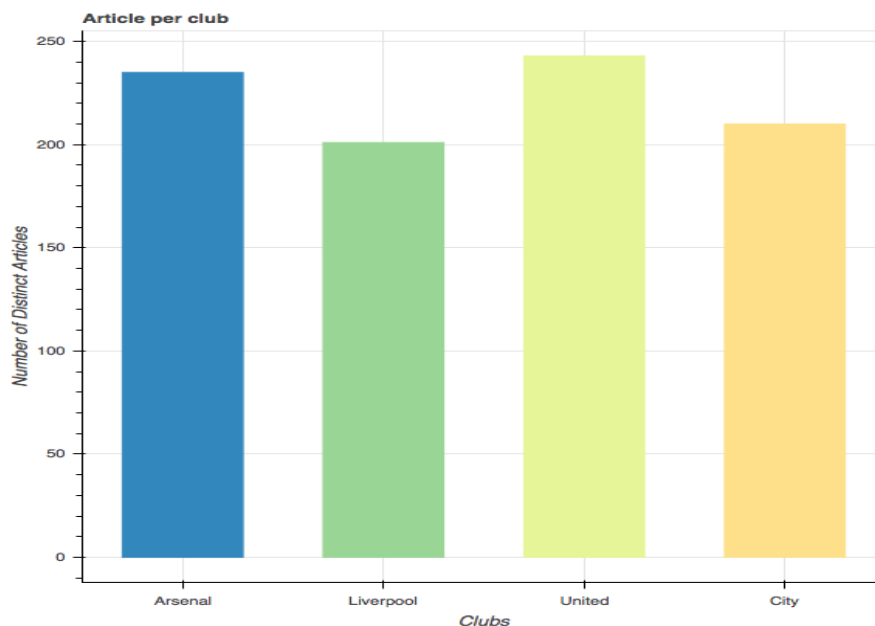
From the plot below, we can say the following –

- Independent, ESPNFC and Daily Mail are the most reliable news sources for Manchester United transfer rumors
- Mirror and Goal.com are the most unreliable sources



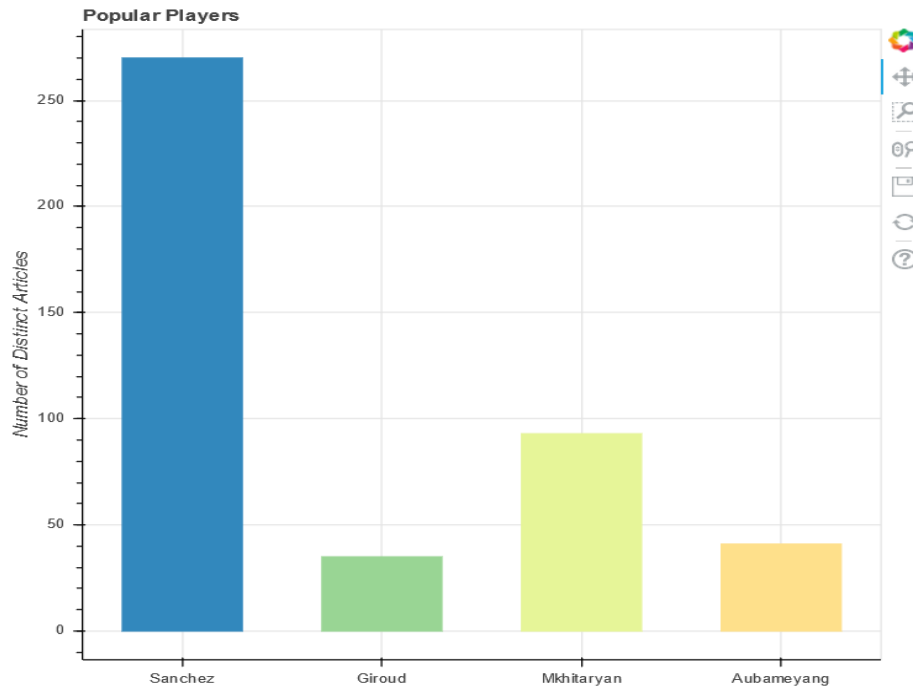
#### 5. Most popular clubs

From the results, we conclude that Manchester United and Arsenal were the most popular clubs as they had the most number of articles. This was due to the fact that there was a swap deal of 2 star players namely, Alexis Sanchez and Henrikh Mkhitaryan, between Arsenal and Manchester United.



## 6. Most popular players

Sanchez and Mkhitarian were the most popular players. Sanchez was highly talked about because he belonged to Arsenal and he was speculated to move to Manchester City, but he eventually moved to Manchester United. Mkhitarian belonged to United before the transfer season and was involved in a swap deal with Sanchez.



## FUTURE SCOPE

1. We can make it better by specifically scraping "to" and "from" of a player from the article instead of just the proper nouns.
2. It can be extended to the other transfer windows and other leagues.
3. We can take the averages of the number of articles published per player, set a threshold and then predict the transfer of future players
4. Refine the database and use it for data mining to predict clubs that are likely to spend the most money in a season by collecting more historical data and looking for patterns in the data.

## **REFERENCES**

- <https://api.sportradar.us/soccer-xt3/eu/en/tournaments/>
- <https://newsapi.org/v2/everything?>
- <https://nlp.stanford.edu/software/CRF-NER.shtml>
- <http://www.google.com>
- <http://www.stackoverflow.com>