Social Media Analytics, 2023/24 academic year

Sentiment and community detection about winter transfer market window on r/ACMilan

Andrea Lucini Paioni, n. matr. 826578

Abstract

This project aim to analyze the community, the posts and the comments of the subreddit *r/ACMilan*: in particular, it will focus on the posts and comments of January 2024, that matches the winter transfer market window for Italian football. In this project, the phase of data acquisition and preparation will be followed by the developing of sentiment analysis methods (for both posts and comments), some topic modelling and a focus on community detection.

Contents

Intro

AC Milan is one of Italy's historic football clubs, known worldwide for his rich tradition of both national and international success, huge part of global history of football. The club had the most successful period between the 80’ and the first decade of the 21st century, cementing his spot among the most prestigious and known club in football; however, this period of great success was followed by a period of downfall in the second decade of 21st century, a period of time that only in the last couple of years seems to be near the end, with the return of European football thorough the seasons and the return to success with an Italian championship in 2022.

The team has passionate fan base, renowned as the “Rossoneri” (from the official colours of badge and the team kit), widely spread all over the world thanks to the success achieved in all of his history.

Reddit is a social media platform where users can share content in the form of text, video, images, links; founded in 2005, Reddit became one of the most popular social media: the composition of the platform, made of a huge number of communities focused on different interests (called subreddit), helped attract more and more users, with the chance to always find a topic of interest for everyone, and a related community.

With such premises, it’s easy to think that the AC Milan subreddit (*r/ACMilan*) is one of the most popular in football, and the most active in Italy compared to the other football related subreddits, with 62000 active users (or redditors, as the reddit users are commonly known). Born in 2010, the community grew year after year, establishing as one of the most active Italian subreddit about sport.

AC Milan started the season well, with a lot of changes in the composition of the team after the almost shocking sale of Sandro Tonali, part of the backbone of the team, but with a huge transfer market campaign in the summer, with over 120mln spent and 10 players arrived. Unfortunately, after a good start, a couple of losses alongside a series of injuries determined a bad period which led to an early exit from the Champions League, and few losses which led the team to a comfortable third-place, far from the other team fighting for a Champions League spot for next season, but also not close enough to Internazionale and Juventus to challenge them for the national championship.

With these premises, it's interesting to understand the sentiment of AC Milan supporters getting closer and closer to the winter transfer window: it's a chance to bring in some new players, to improve the team, to make up for the series of injuries that the team suffered in the previous months.

The goal of this project is to analyze the community belonging to *r/ACMilan* and to understand the sentiment of AC Milan supporters during the winter transfer window.

Project Goals

The project aims to:

- Collect data from r/ACMilan between 02/01/2024 and 27/01/2024 (the date when all the data were collected);
- Quickly analyze the data collected (after a cleaning phase), to analyze if some patterns show up;
- Apply methods of sentiment analysis on both posts and comments, to analyze the general sentiment of the community during the winter transfer window;
- Apply a method of topic modeling, to highlight the most common topic emerged with the posts;
- Apply community detection methods to analyze the composition of the subreddit.

Data collection

Data were collected with Python, using the library *request*, that allows to collect only a certain amount of data every minute (only 100 request per minute). For this reason, a number of consequent requests were made, to ensure the collection of all the data of interest.

Two datasets were created: the *Reddit_post* dataframe, and the *Reddit_comments* dataframe.

The *Reddit_post* had information about:

- '*title'*: the title of the posts;
- '*selftext'*: the text of the posts;
- '*upvote_ratio'*: the percentage of likes out of all votes;
- '*ups'*: the number of likes;
- '*author'*: the nicknames of the users who wrote the posts;
- '*date_and_time'*: the date and time when the posts were published;
- '*user_id'*: the id of the user who wrote the posts;
- '*kind_id'*: the id of the posts;
- '*num_comments'*: the number of comments of the posts;
- '*date'*: the date when the posts were published, from the variable '*date_and_time*';
- '*time':* the time when the posts were published, from the variable '*date_and_time*';

Only the posts between the 02/01/2024 (when the winter transfer window started) and the 27/01/2024 (the date of the last data collection) were collected. Also, some posts were filtered out (the ones about the matches, because they are among the most commented, but aren't technically related to transfer market).

The *Reddit_comments* had information about:

- '*name*: the id of the comments;
- '*ups'*: the number of likes, and dislikes (for negative values);
- '*author'*: the nicknames of the users who wrote the comments;
- '*author_fullname'*: the id of the user who wrote the comments;

- ‘*parent_id*: either the comments of the posts, or for the subcomments, the comments of which they were an answer;
- ‘*link_id’*: the id of the posts were the comments can be found;
- ‘*body’*: the text of the comments;
- ‘type_coms’: if the comments are comments to a post or to other comments.

It’s important to remember that the comments dataframe not only included comments to the posts, but also comments to other comments.

Data preparation

After acquiring and cleaning the data, to perform sentiment analysis methods, data (in particular text data from the posts) required further preparation steps. A number of steps involved a lowercase transformation, the elimination of punctuation, URLS, links and lemmatization. Finally, stepwords removal was performed, to ensure that only meaningful words were kept.

All the previous steps were performed both on the tokenized result obtained dividing the text in tokens, and on the full text. This is because some further analysis methods required the text divided in tokens, others required the full text (with all the preparation steps performed before).

Exploratory analysis

Through some data visualizations, some insights were collected:

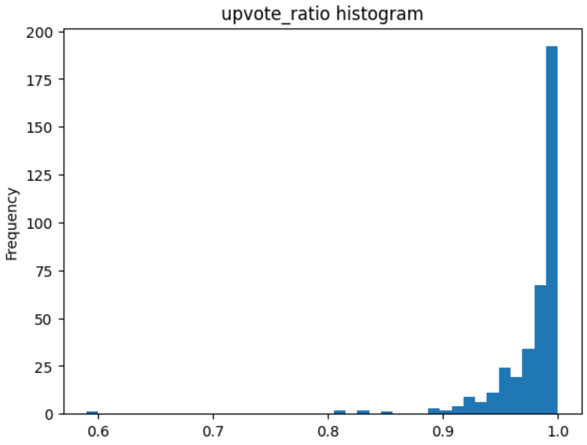

Figure 1: upvote_ratio *histogram*

The previous histogram showed the distribution of the ‘*upvote_ratio*’ variable from the *Reddit_post* dataframe: it’s easy to understand that the post are always upvoted more than downvoted, with just one post below the 80% of upvotes from the total of votes; there are a lot of posts with 100% (or near 100%) of upvotes.

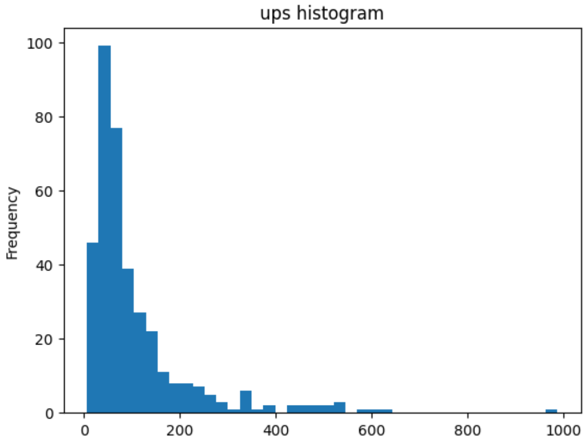

Figure 2: ups *histogram*

The previous histogram showed the distribution of the ‘*ups*’ variable from the *Reddit_post* dataframe: the vast majority of posts have a modest number of upvotes, usually under 150 upvots; but there are also some posts with a lot of upvotes, up to almost 1000.

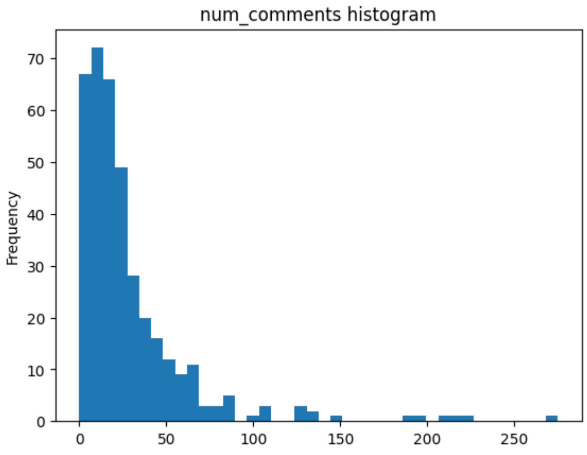

Figure 3: num_comments histogram

The previous histogram showed the distribution of the ‘*num_comments’* variable from the *Reddit_post* dataframe: the graph shows that the majority of posts only has less than 50 comments, but some posts involved more community, up to more than 250 comments for the most commented one.

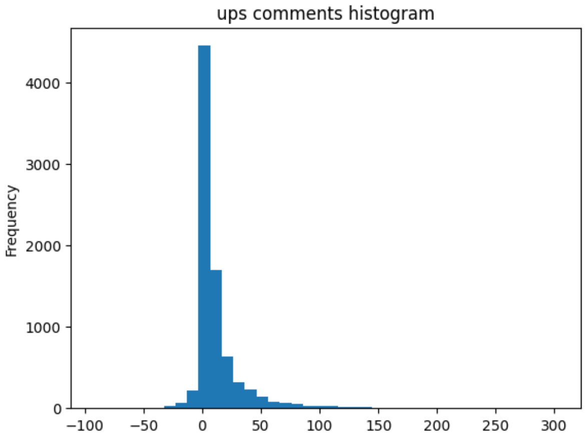

Figure 4: ups comments histogram

Finally, the last histogram shows the distribution of the '*ups*' variable from the *Reddit_comments* dataframe: the distribution is around 0, with an asymmetric distribution towards positive values, but also with negative values (so comments generally not appreciated by the community).

Topic Modeling

The goal of the topic modeling phase was to identify a series of topics included in the text from the posts of the subreddit, to highlight some of the possible interests of the community for January 2024.

A Wordcloud was developed starting from the text of the posts, containing the main topics of the month:

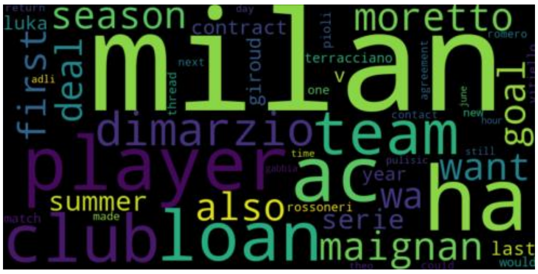

Figure 5: wordcloud

Some words are just common token in a subreddit related to Milan: the words 'milan', 'team', 'player', 'goal', 'club'… But it's not a common month, it's the month of winter transfer market window, and it can be seen with other words like 'loan', 'dimarzio' and 'moretto' (two famous Italian journalists, expert on market aspects), 'deal'…

Sentiment Analysis

Sentiment analysis was performed in order to obtain useful information about the general sentiment of the Rossoneri community on reddit.

Three sentiment analysis models were developed, both on posts and on comments:

- The AFINN Lexicon Base Approach (a list of English terms manually rated for valence with an integer between -5, or negative, and +5, or positive)
- The Opinion Lexicon (another lexicon-based approach)
- The Dictionary Based emotion detection (emoticon lexicon, called NRC, that defines some score for a series of emotions and considerations about text, such as 'positive', 'anticipation', 'disgust', 'joy', 'sadness', 'surprise', 'trust', 'anger', 'negative' and 'fear')

Sentiment on posts

We now focus on sentiment performed on posts text.

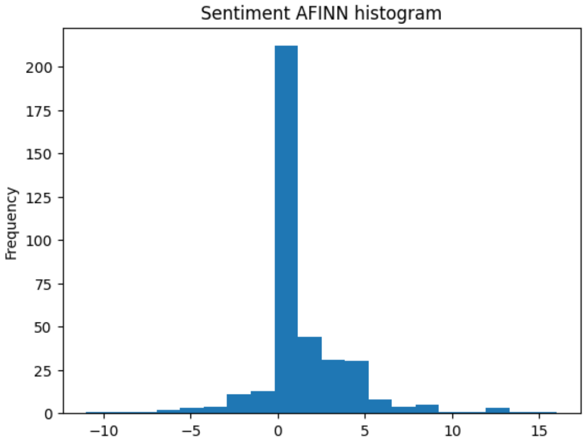

Figure 6: sentiment AFINN histogram on posts

The histogram above states the general sentiment obtained with the AFINN method: positive values indicate a positive sentiment. It's easy to see that values are typically positive, with just a few posts with negative sentiment (a trend following the distribution of the variable '*upvote_ratio*' seen in the exploratory analysis).

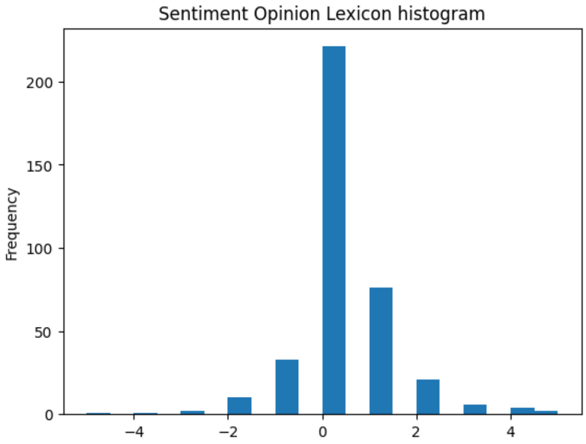

Figure 7: sentiment Opinion Lexicon histogram on posts

The histogram above shows the general sentiment obtained with the Opinion Lexicon method: as the previous method, values are usually positive, with just a few posts with negative sentiment (a trend following the distribution of the variable '*upvote_ratio*' seen in the exploratory analysis).

With the Dictionary Based emotion detection, the result is similar to everything obtained until now: calculating the mean values of the emotions reported, it was possible to see that positive posts are way more common than negative ones, and other common sentiment are trust and anticipation; some of the least common sentiment are disgust, anger, sadness and fear.

Sentiment on comments

We now focus on sentiment performed on comments text.

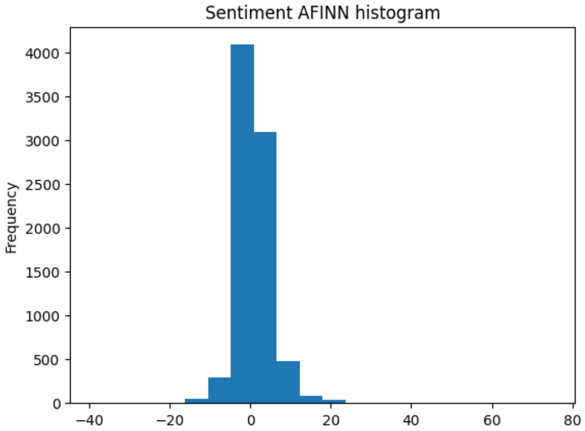

Figure 8: : sentiment AFINN histogram on comments

The histogram above states the general sentiment obtained with the AFINN method. Values are different than the ones seen with the posts, there's more balance between negative and positive sentiment.

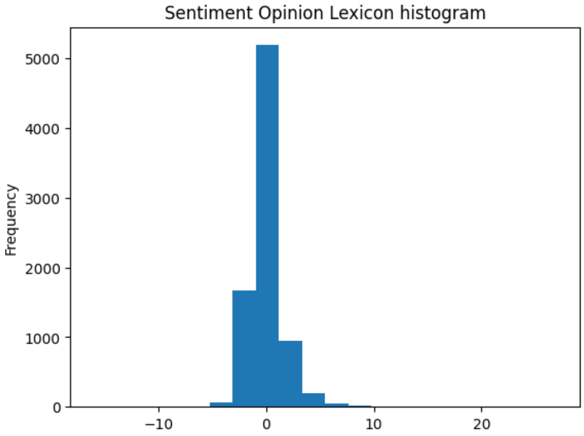

Figure 9: : sentiment Opinion Lexicon histogram on comments

The histogram above shows a situation similar with the Opinion Lexicon method applied on posts: values are usually positive, with just a few posts with negative sentiment.

With the Dictionary Based emotion detection, the result is similar to everything obtained with the posts: calculating the mean values of the emotions reported, it was possible to see that positive comments are way more common than negative ones (but with less difference now), and other common sentiment are trust and anticipation; some of the least common sentiment are disgust, surprise, anger and fear (with higher values than the ones seen before on the post analysis).

Evaluations on sentiment

The confusion matrix found while comparing the different sentiment analysis methods developed on the posts showed a good level of accuracy (71.6% for the 3x3 confusion matrix with positive, neutral and negative scores; 94.4% for the 2x2 matrix with negative and positive scores).

The confusion matrix found while comparing the different sentiment analysis methods developed on the comments showed a good level of accuracy (70.9% for the 3x3 confusion matrix with positive, neutral and negative scores; 89.8% for the 2x2 matrix with negative and positive scores). Still high scores, but lower than the posts one.

Community detection

Community detection methods are aimed at getting information about the community involved in the subreddit of interest. In particular, community detection helps understand the connection within the community active in *r/ACMilan*.

For this purpose, a graph was made, linking every post to every comment made by the community, so that each member who either posted or commented a post is linked with everyone who interacted with the post/comment.

Before building the graph, every post and comment made by an user with a deleted (or blocked) account was removed by the analysis, because there was the risk of manipulate the results.

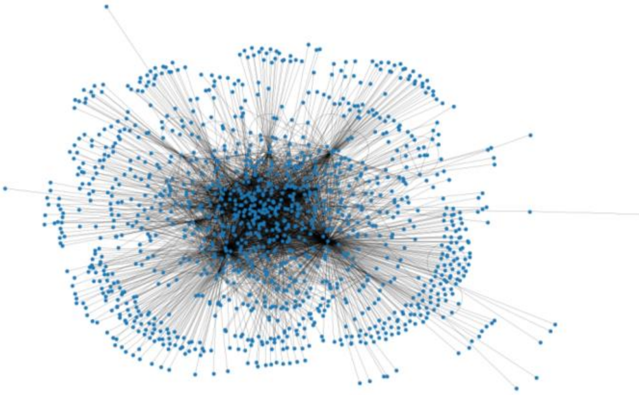

Figure 10: graph

The graph obtained is formed by 1293 nodes (which are the redditors) linked with 3624 edges (connections made with comments).

Some metrics were calculated, such as:

- Betweenness centrality: it helps find very influent nodes over the flow of the graph; redditors with some of the highest values are Claija79 (0.38), mercurialsaliva (0.34) and HeirOfRhoads (0.09).
- Degree centrality: it helps find nodes with the highest number of edges coming in or out; redditors with some of the highest values are Claija79 (0.43), mercurialsaliva (0.39) and HeirOfRhoads (0.18).
- Closeness centrality: it helps find out how close a node is to all other nodes in the graph; redditors with some of the highest values are Claija79 (0.61), mercurialsaliva (0.59) and HeirOfRhoads (0.51).

The three users above can be easily considered as some of the most influential members of the subreddit.

The degree of assortativity, having a value of -0.4, indicates relationship between nodes of different degree. The eccentricity is between 4 and 6, meaning that the maximus distance of a vertex from other vertex is always between these two values.

After the graph, a community detection method was implemented, to analyze if there are hidden structures underneath the data. The Louvain Community Detection algorithm was implemented, with the purpose of finding clusters in the graph: it doesn't require neither the number of communities searched nor their size. There are multiple steps, until there aren't more changes in the network, and max modularity is achieved.

The method found 16 communities, with different dimensions (from a minimum of 24 redditors for the smallest, to a maximum of 198 redditors). The corresponding modularity coefficient achieved a value of 0.34, far from the optimal value of 1 searched. As a result, the graph obtained before can be shaped like this, showing the communities found:

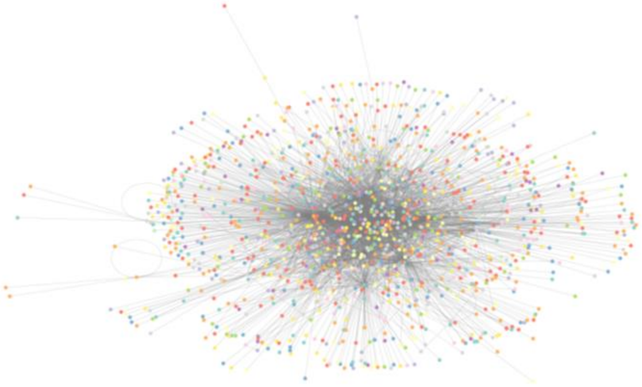

Figure 11: graph with community detection

The resulted graph shows the communities not well separated but almost mixed and overlapping.

As a consequence, looking at statistics about the sentiment analysis methods developed in the sentiment analysis phase, there aren't great differences between the communities built with the Louvain Community Detection algorithm.

Conclusions

This project aimed to analyze the community, the posts and the comments of the subreddit *r/ACMilan*, with a focus on the posts and comments of January 2024, matching the winter transfer market window for Italian football.

The phase of data acquisition and preparation was followed by the developing of sentiment analysis methods (for both posts and comments), some topic modelling and a focus on community detection.

The sentiment analysis, with the three methods developed on both posts and comments text, showed similar results: overall positive sentiment (with little negative sentiment), with little distinctions between posts and comments.

The topic modeling showed not only words related to AC Milan in general, but also terms related in particular to transfer market aspects, as it was expected.

The community detection phase helped found some of the most influent redditors of the community (Claija79, mercurialsaliva and HeirOfRhoads), and tried to build some communities inside the subreddit, with mixed results.

References

https://www.reddit.com/r/ACMilan/

Social Media Analytics notes

https://python-graph-gallery.com/