r/marvelstudios

COMMUNITY AND CONTENT ANALYSIS AFTER THE RELEASE OF AVENGERS: ENDGAME

Roberto Ferrari 852220
Davide Prati 845926
Marco Sallustio 906149

CONTENTS

- Main Goals
- Data Collection
  - Data Preprocessing
- Data Exploration
- Social Network Analysis
  - Graphs
  - Metrics
  - Community detection
- Social Content Analysis
  - Sentiment Analysis
  - Topic Modeling

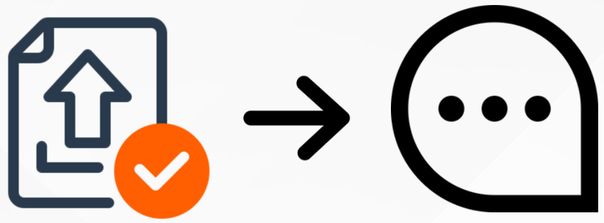
MAIN GOALS

The main objective of this project consists in the **analysis of the r/marvelstudios subreddit.** In particular we decided to analyze the interactions generated for one of the most popular Marvel films: Avengers Endgame.

We can identify some of the main objective of our analysis:

1. **Identify the relationships** generated within the subreddit and within our discussion of interest;
2. **Identify the most influential users** through Social Network Analysis;
3. **Understand the feedback** that the users of the subreddit have had through the sentiment generated by each one;
4. **Visualize the main topics** discussed in the community and highlight groups of words that tend to relate to a specific topic rather than another.
5. **Identify any communities** present and their characteristics.

DATA COLLECTION



Keywords: "avengers" OR "endgame"
Period of time: April 26, 2019 - April 26, 2020

Features extracted:

- **author_submission**: author of the post
- **title_post**: title of the post
- **score_post**: score number of votes received by the post, indicates the "popularity" of the post
- **url_post**: URL of the post
- **created_utc_post**: date and hour of the creation of the post
- **author_comment**: author of the comment
- **body**: text of the comment
- **created_utc_comment**: date and hour of the creation of the comment
- **score_comment**: score number of votes received by the comment, indicates the "popularity" of the comment

DATA PREPROCESSING

Before Pre-Processing operations:

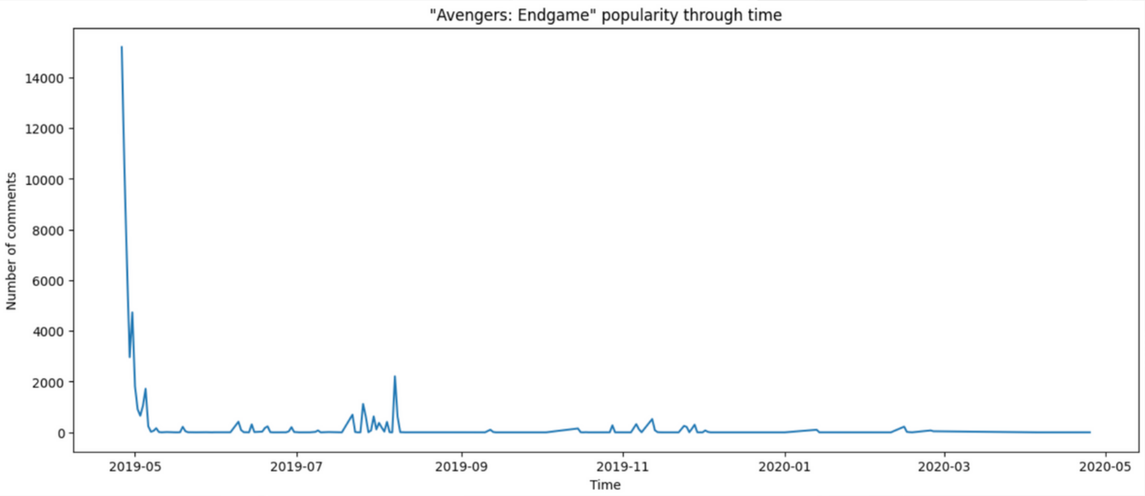
- **Delete comments in a language other than English** → pycld2 library
- **Manage contractions in English**: “I’ve” → “I have”

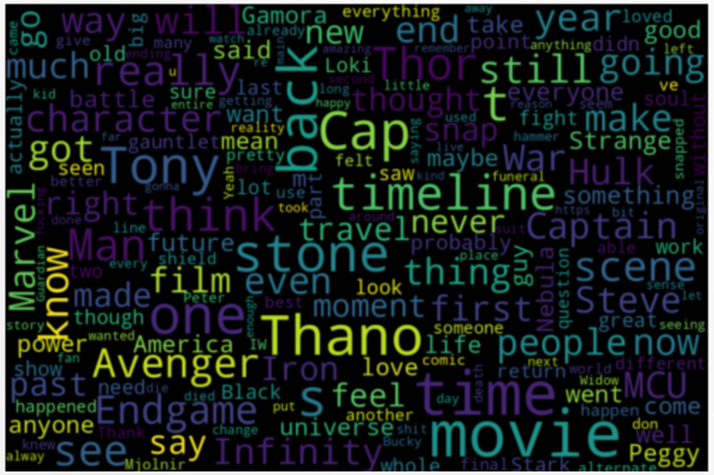
At this point, we performed the following **Pre-Processing operations**:

- Lemmatization;
- Removal of links;
- Removal of e-mail addresses;
- Removal of numbers;
- Removal of extra-white spaces;
- Removal of punctuation;
- Removal of emoji;
- Transforming text to lowercase.

DATA EXPLORATION

The exploratory operations **focus** on the **visualization** of significant **features** of the dataset. The first to be observed is **popularity** of the topic: the highest peak in the subreddit is registered with 15202 comments, corresponding to the first days after the official release of the film, only to drop significantly a week later.





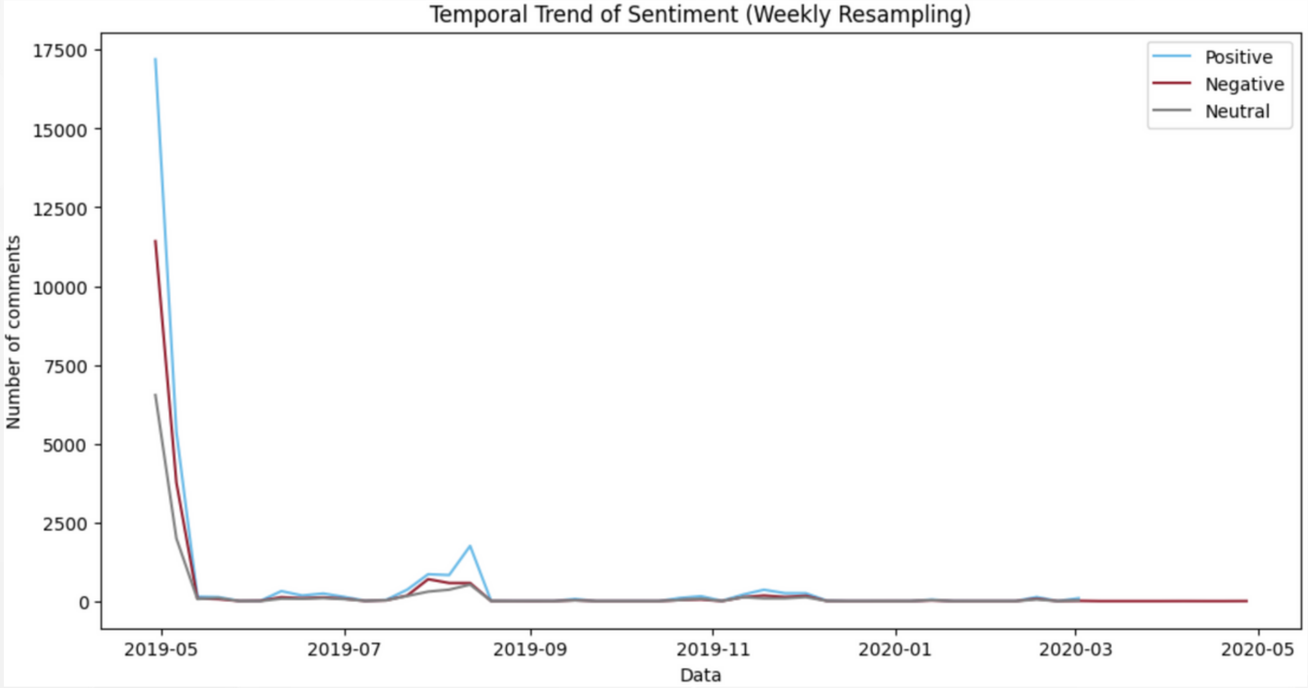
The second feature is the frequency of the **words adopted** and to achieve this, it was performed a Word Cloud analysis: the most picked ones are words related either to the characters of the series, particular elements of this final chapter or more common words such as *movie* and *Endgame*.

SENTIMENT ANALYSIS

- We used the model VADER that is sensitive to both polarity (positive/negative) and intensity (strength) of emotion.
- It relies on a dictionary that maps lexical features to emotion intensities, known as sentiment scores. The sentiment score of a text can be obtained by summing up the intensity of each word in the text, which ranges between -1 and 1.
- To have even more detailed values we have also created three different functions for each polarity (positive, negative, neutral) which return the sentiment score rather than a compound score.
- For example, for the first row of our dataset we have:
  - **Compound Value**: 0.9745
  - **Negative Value**: 0.328
  - **Positive Value**: 0.065
  - **Neutral Value**: 0.607

SENTIMENT ANALYSIS

To perform some analysis we decided to create a new column in our dataset that directly contained the sentiment value associated with the comment (Positive, Negative or Neutral). To do this we used the CompoundValue obtained previously and we established a specific threshold in order to convert the numerical value into the associated sentiment.

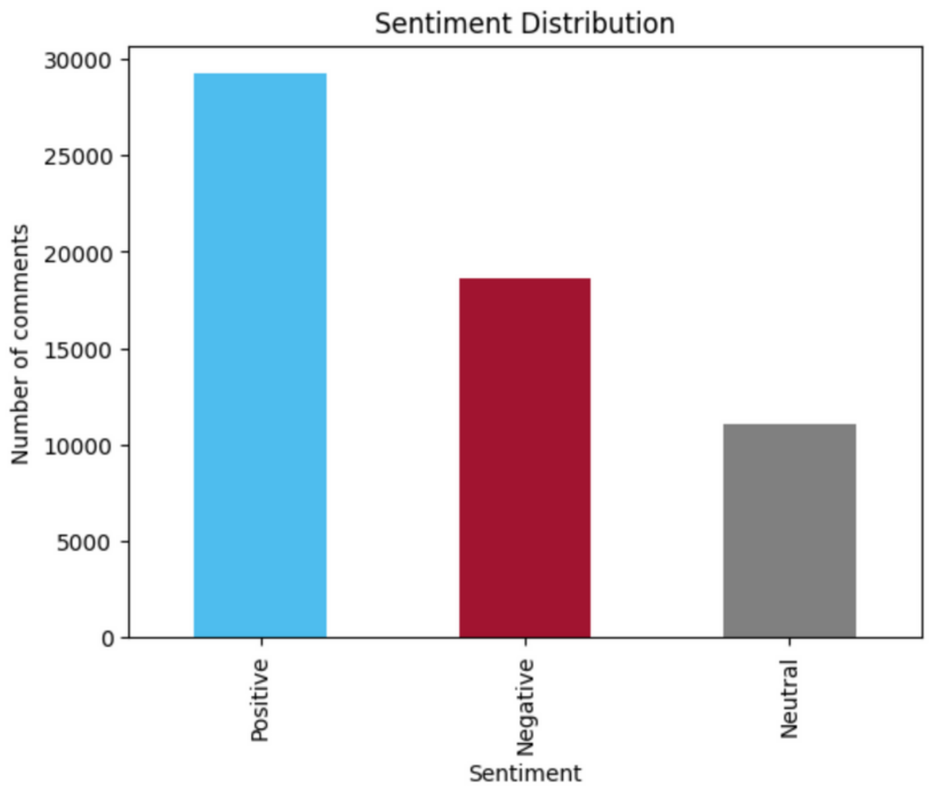


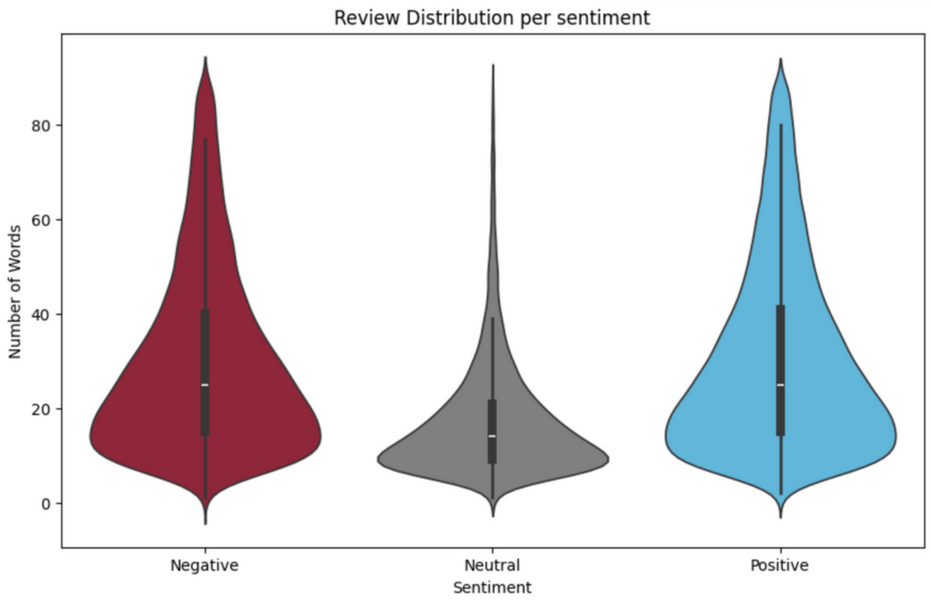
Positive WordCloud

Negative WordCloud

SENTIMENT ANALYSIS

We also observed the distribution and the length of the comments in relation to each sentiment label.





One interesting aspect is related to the length of the comments, i.e. **number of words**, observed **for each sentiment label**. To highlight this, we created three violin plots, that represent the three sentiment labels.

SENTIMENT ANALYSIS

Conclusions:

- Thanks to sentiment analysis we also managed to satisfy the objective in which we asked ourselves to understand user feedback regarding the analyzed topic of interest.
- We have identified the most common words for negative and positive sentiment
- And we were also able to analyze the weekly trend of comments to understand the type of sentiment generated in particular periods.

Future developments:

- To get even more information on the feedback from the users, an emotion analysis could also be performed.
- In this way we would be able to understand a person's emotional state or the emotional expression associated with a user's comment.

TOPIC MODELING

Topic modeling was performed to group comments based on **common themes or topics**, analyzing the frequency and distribution of words in one case compared to others.
To do so, we employ an **LDA** (Latent Dirichlet Allocation) model.

Metrics used for evaluation:

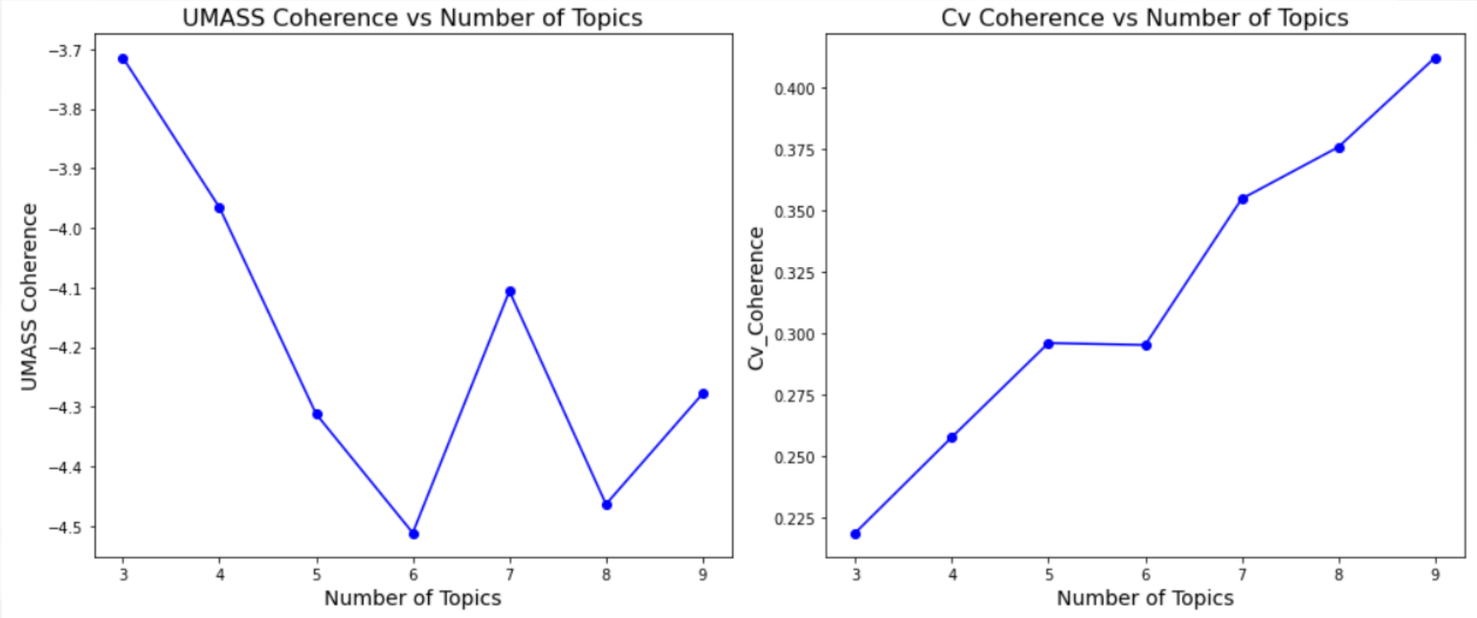
- U_mass coherence
- C_v coherence
- human judgment

TOPIC MODELING - TEXT REPRESENTATION

To obtain a text suitable for fitting the model and fully leverage its features, we implemented the following steps:

- **tokenization**
- **stop words removal**, with reference to the list of English stop words, from which we removed 'not' and added default Reddit words.
- **creation of bigrams and trigrams**: this involved identifying a total of 1047 bigrams and 59 trigrams, which were added to the corpus.
- **removal of most and less frequent words**: removal of words occurring in only one or two documents, and those present in more than 15% of them.
- represent the text with the **Bag of Word approach**.

TOPIC MODELING - LDA



Running the model with the number of clusters varying from k=3 to k=9, we ultimately opt for the model with seven topics, which takes the following parameter values:

- -4.106 U_mass coherence
- 0.355 for C_v coherence

TOPIC MODELING - LDA

Finally, we generated WordClouds for each of our seven topics. With our interpretation, we attempted to provide titles for the topics:

1. **general opinions** on the movie, expectations
2. focus on particular **memorable scenes** and **favorite moments**
3. **actors and characters' development**, differences between this film and the previous ones
4. **critical reflections** on the plot, actions, and thoughts of the characters
5. astonished and sensationalistic **comments on impactful moments** and unexpected plot twists
6. **specific characters and their superpowers**, with reference to a cut scene, likely featuring Dr. Strange as its protagonist
7. focus on the **most iconic chararcters**: Thor, Captain America, Iron Man

TOPIC MODELING

Conclusions:

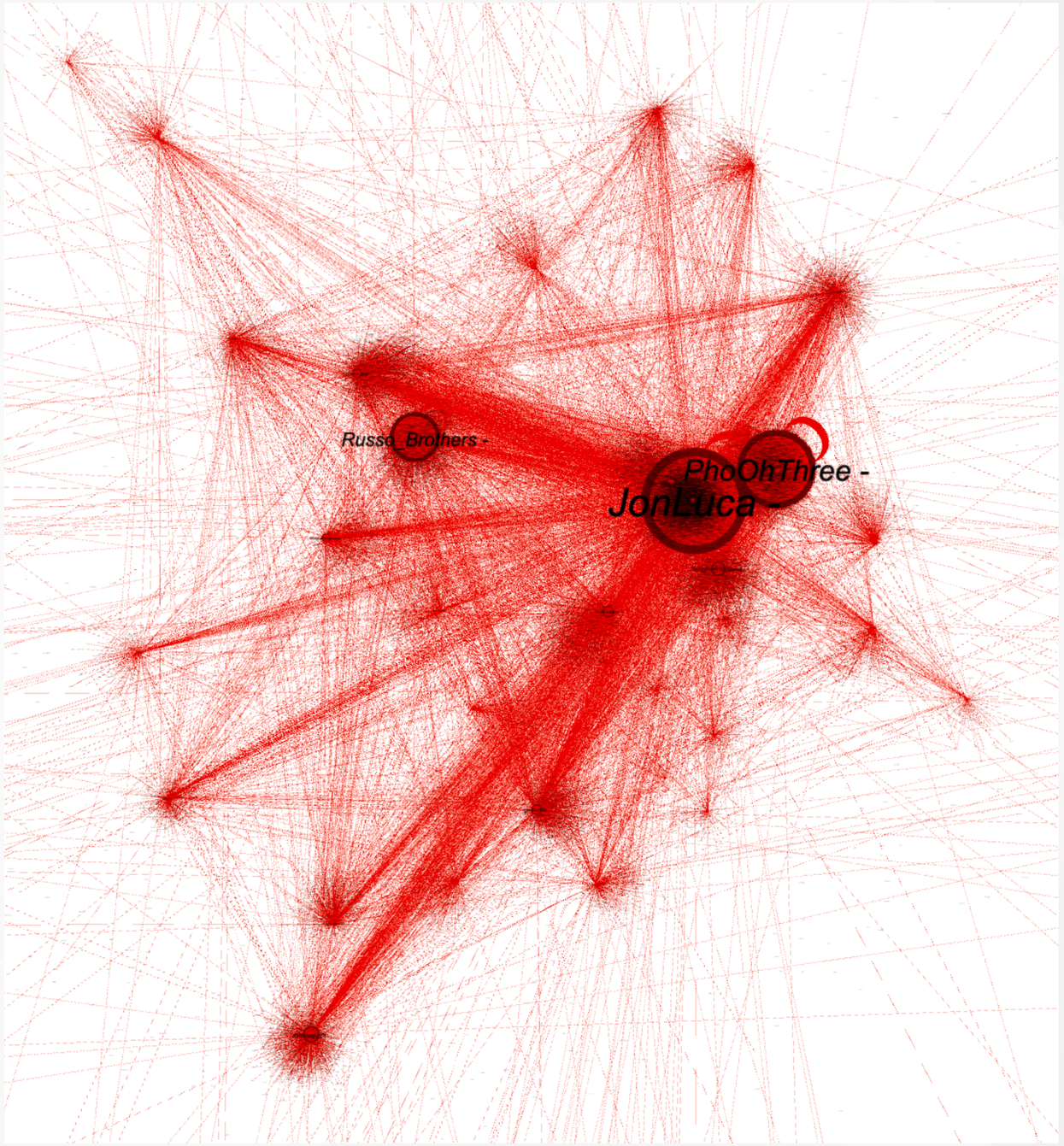
- the topics primarily revolve around opinions on the film's quality and plot, along with considerations about the story and characters
- despite the film being among the highest-grossing in cinema history, none of the detected topics explicitly references the box office success or the production aspect.

Future developments:

- attempt evaluating the word clouds for other values of k, favoring a different balance between C_v and U_mass coherence than the one initially considered
- explore the development of other topic modeling models, such as LSA or BERT, using different text representations like TF-IDF, and observe the levels of C_v and U_mass coherence they achieve.

GRAPH

- **Source node** = author of the comment
- **Target node** = author of the commented content
- **Weight**= Comment score
- **# Nodes** = 25.388
- **# Edges** = 32.747



GRAPH METRICS

We calculated some general graph metrics that give us important information about the nature of our network

Metric	Value
Density	0
Avg. Degree	1,29
Avg. Weighted Degree	88,25
Avg. Path Length	1,764

GRAPH METRICS

In order to understand who the main influencers of the network are, were computed two centrality metrics: **closeness centrality** and **betweenness centrality.**

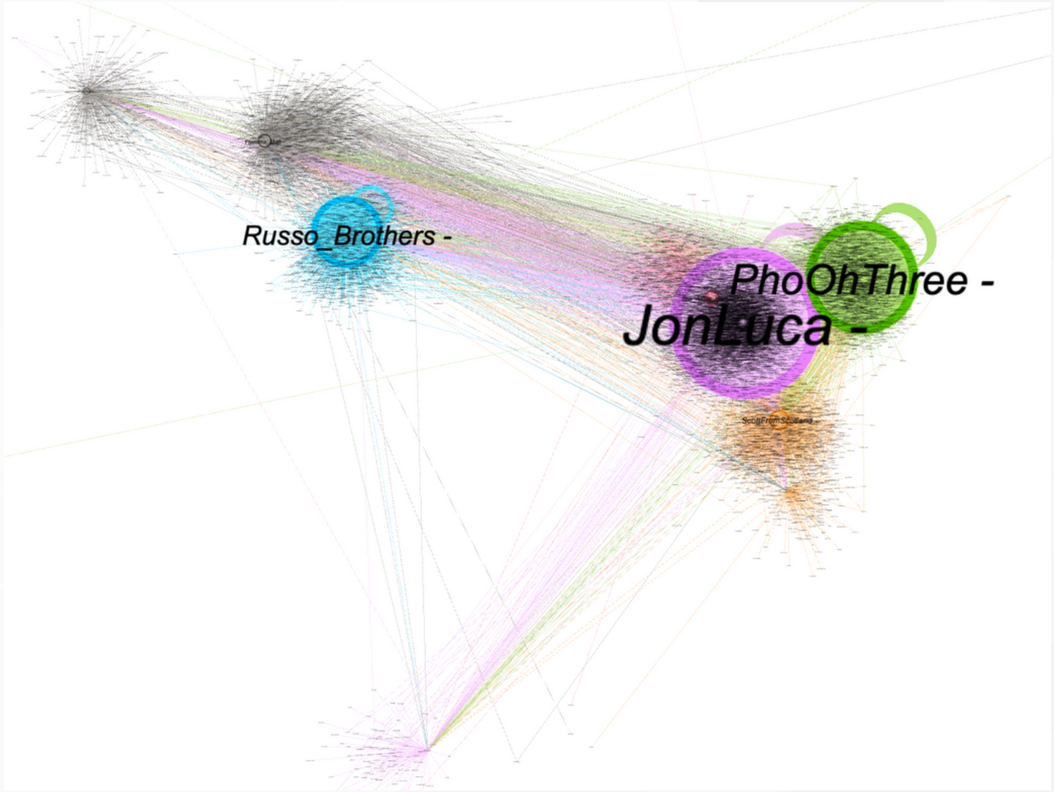
User	Betweenness Centrality
The_Asian_Hamster	14530.1
KostisPat257	13641.5
Flamma_Man	9227.5
Sisiwakanamaru	4520.5
PhoOhThree	3600.6
kahlkorver	1983
MangoJam18	1294
LordHyperBreath	670.9
dannys717	570
ENusatron	392.08

User	Closeness Centrality
The_Asian_Hamster	0.58
KostisPat257	0.85
Flamma_Man	1
Sisiwakanamaru	0.42
PhoOhThree	1
kahlkorver	0.5
MangoJam18	0.61
LordHyperBreath	1
dannys717	1
ENusatron	0.66

COMMUNITY DETECTION

With the aim of analyzing the communities of interest, was calculated the modularity value of our network.

0	(32,39%)
8	(11,74%)
353	(6,66%)
4	(6,44%)
1	(5,87%)
3	(4,99%)



Results: 485 communities with a modularity value of 0,744

COMMUNITY DETECTION

To better analyze the communities found, we report the 6 top communities and the **associated sentiment values** to understand whether sentiment influenced the division of the communities

Community	Populousness	Positive	Negative	Neutral
0	8131	49,61%	33,20%	17,10%
8	2948	46,80%	32,50%	20,65%
353	1673	61,92%	19,78%	18,29%
4	1817	48,67%	32,40%	19,35%
1	1474	47,15%	33,24%	19,60%
3	1253	47,72%	33,28%	18,99%

- We have very similar sentiment values. This indicates that people within each community interact similarly, with a balance between positive, negative and neutral comments.
- The only different values are for community 353 which has a higher positive sentiment value (almost 62%) and more unbalanced than the other communities.
- Observing the previous graph we know that this community includes the user 'Russo_Brothers' whose node is positioned in a more distant position than the others. This may be because this community may cover topics or content that is significantly different from other communities.

COMMUNITY DETECTION

Conclusions:

- Thanks to community detection we met one of our goals regarding the need to find how our network is divided into communities.
- We obtained that in the largest communities found, all the nodes with a higher weighted degree value corresponded.
- This phenomenon leads us to support the thesis that:
  - central nodes with a high degree of weighting could have a greater probability of belonging to larger communities
  - communities could grow around central nodes with a higher degree of weighting, since these nodes attract more connections over time and indeed some nodes with a higher degree of weighting could play a key role in maintaining the integrity and cohesion of communities, leading to the formation of larger communities around them.

THANK YOU

Università degli Studi Milano Bicocca