02805 Social graphs and interactions: Project Assignment B

Ioannis Vlasakoudis s232755, Max Heiberg Bestle s194574, Jonatan Rasmussen s183649

December 11, 2024

In this paper, we analyze the 20 Newsgroups dataset to understand how online users interacted across different topics. We construct a network where nodes represent users who posted across the newsgroups, with edges connecting users based on their interactions. We then apply community detection algorithms to uncover natural groupings of users.

Social networks | Text analysis | Sentiment Analysis | TF-IDF | Data Science

In the early days of the internet, online communities formed through newsgroups where people shared opinions, discussed various topics, and engaged in conversations through email-like posts.

These newsgroups represented one of the first large-scale digital social networks, making them an interesting subject for both network and text analysis. In this paper, we analyze the 20 Newsgroups dataset to understand how these early online communities were structured and how users interacted across different topics.(1).

Research questions

Our analysis combines network science with natural language processing to investigate three main questions: How did users form communities across different newsgroups, and are they aligning with the predefined newsgroup categories? What were the main topics of discussion within and across these communities? And how did the sentiment of discussions vary across different topics and communities?

Methodology

To answer these questions, we constructed a network where nodes represent users who posted across the newsgroups, with edges connecting users who interacted with each other. We then applied the Louvain method to uncover the structural communities of the users. These communities are analyzed using both topic modeling techniques (Latent Dirichlet Allocation, more on this later) and sentiment analysis to understand the nature of discussions within each community. For sentiment analysis, we use the LabMT wordlist(2) for assigning a 'happiness rank' to common English words, which allows us compare the positivity/negativity across forums. *Link to our notebook: https://github.com/maxx1559/socialgraphs-final/tree/main*

Data overview and source

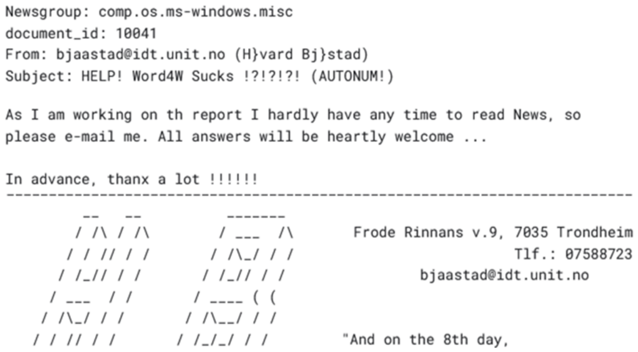
As previously mentioned, our data has been sourced from the 20 Newsgroups dataset. The version of the dataset that we are using has been downloaded from Kaggle.com, see the references section.(3) In terms of dataset content, each newsgroup is discussing a specific subject, such as religion, politics, or sport. An example of what posts look like can be seen in figure 1.

Each newsgroup submission has the following attributes: document id, the newsgroup, the author, and subject title. However, the dataset is a big text file containing 1000+ posts, so we decided to do some pre-processing and turn it into JSON files. When splitting the raw data into posts, we ended up with a total of 18,773 posts. The posts are distributed approximately evenly across the newsgroups, with most containing around 1,000 samples, see figure 2. The average length of each post is 193 words, which is quite long by modern social media standards. This means that we have almost 4 million words in our data set!

Data partitioning. The 20 Newsgroups website suggests grouping the newsgroups into six broader subject areas: Computers, Buying & Selling, Recreational, Politics, Science, and Religion, see figure 3. Generally speaking, it is also easier to display 6 categories in a plot compared to 20 categories, which is why we are using this partition for parts of our work.

Significance Statement

This project work provides insights into how early internet users formed social bonds and communities around shared interests, and how these communities sometimes transcended the formal boundaries of individual newsgroups. Understanding these patterns can help us better comprehend the evolution of online social networks and community formation in digital spaces.



```
Newsgroup: comp.os.ms-windows.misc
document_id: 10041
From: bjaastad@idt.unit.no (H}vard Bj}stad)
Subject: HELP! Word4W Sucks !?!?!?! (AUTONUM!)

As I am working on th report I hardly have any time to read News, so
please e-mail me. All answers will be heartly welcome ...

In advance, thanx a lot !!!!!!
-------------------------------------------------------------------------
          --   --          -------
         / /\ / /\        / ___  /\       Frode Rinnans v.9, 7035 Trondheim
        / / // / /       / /\_/ / /                         Tlf.: 07588723
       / /_// / /       / /_// / /                    bjaastad@idt.unit.no
      / ___  / /       / ____ ( (
     / /\_/ / /       / /\__/ / /
    / / // / /       / /_/_/ / /        "And on the 8th day,
```

Fig. 1. An example of a newsgroup post. This is what our data looks like!

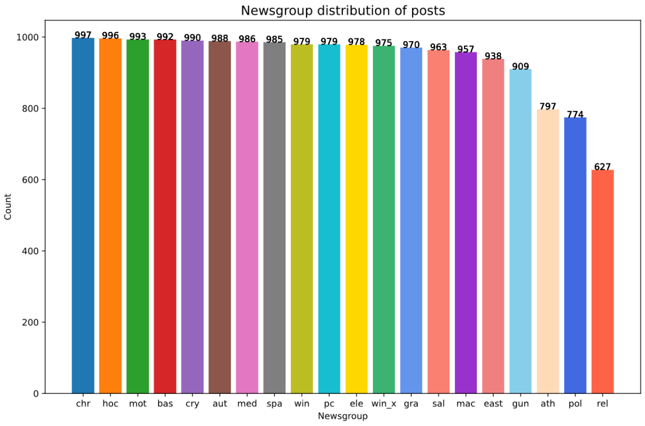


Fig. 2. Distribution of posts across newsgroups. It appears that the original dataset sampled 1000 posts from each newsgroup. However the religion.misc newsgroup stands out as an outlier, having only 627 samples. Still, this should not cause any issues for our data analysis work, as it's a very healthy sample size.

Results and key findings

This section is a presentation and visualizes of our key findings. Discussion regarding our results can be found in the discussion section on the next page of the report. For our code and additional data, check out the notebook.(4)

Individual Networks. In order to construct the large network between users in all newsgroups, we first had to analyze and prepare the individual newsgroup networks. As mentioned in Methodology, we created the networks, with nodes representing the newsgroup's users and they were connected to each other if there was an interaction between them (for example, a user mentioned the name of another user/node in their post). For each newsgroup, only the largest connected component was selected for analysis and the result was 20 networks with an average of 246 nodes and 574 edges, with the largest being the network of *Computer Graphics* with 636 nodes and 1222 edges (see Figure 4).

Unified Network and Communities. After the analysis of the individual networks, it is time to create the large unified network. For that, we had to begin by identifying the common users across these 20 networks. There are 347 users that have posted in different newsgroups and most of them between atheism and religion (32). After the combination, a unified network of 4306 nodes and 11098 edges emerged (Figure 5).

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

Fig. 3. The 20 newsgroups partitioned into 6 subjects.

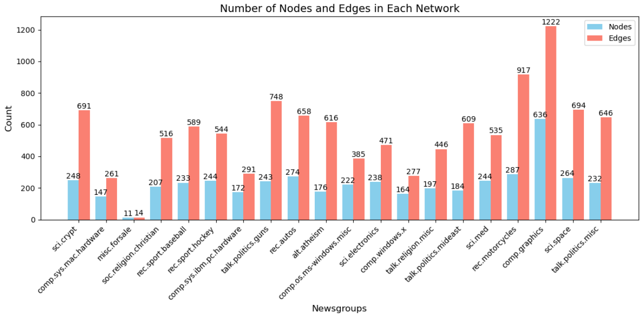


Fig. 4. Number of Nodes and Edges of Individual Networks

The next step was to apply the Louvain method and detect the communities, which could offer us valuable insights. From this method, 21 communities were derived. Afterwards, we proceeded with the analysis of these communities and calculated the distribution of users, as well as the total number of common users in each community (Figure 6). Looking at the distribution of degrees within the large network, we found that while the network did seem to follow a power-law distribution, it did not follow a scale-free regime, for either in-degrees, out-degrees or the total degrees.

TF-IDF and Most Frequent Words. Analyzing the word frequency is an additional way to deeply understand the communities and how the users in them think. Thus, we found the 10 most used words for each community (after some text processing) and for a better, comprehensive analysis, we also calculated the TF-IDF scores of the words, in order to identify which words are the most important for each community. For example, the top words for community 3 are *'gun', 'people', 'one', 'government', 'right', 'get', 'think', 'weapon', 'law', 'fire'*, and they can reveal a lot about the context of the posts.

Sentiment Analysis. To measure the language used across different topics, we have used sentiment analysis to quantify the positivity/negativity of each topic. During our analysis, we found that the *forsale* newsgroup had an average sentiment score of 5.566 (±0.179), which was the highest of the sentiments, whereas the political newsgroup had the lowest average sentiment score, which was at 5.312 (±0.162). The sentiment all the subjects can be seen in figure 7. We see that the difference is statistically significant, which means that the sentiment actually differs across topics. Perhaps it is not too surprising that for-sale posts contain positive words and language to promote the items that are being sold, whereas political discussion is more likely to spark conflicts and disagreeable language.

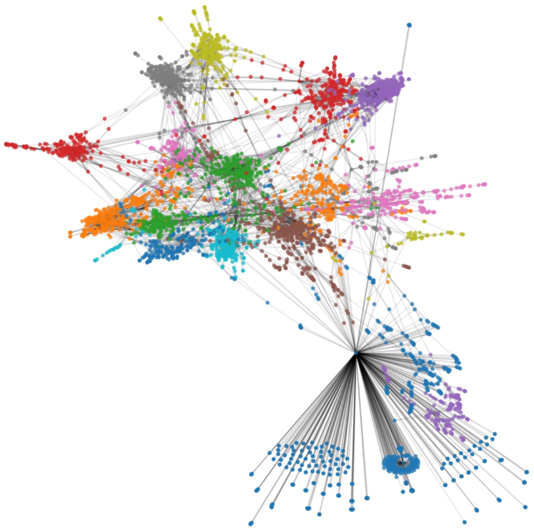


Fig. 5. The Unified Network

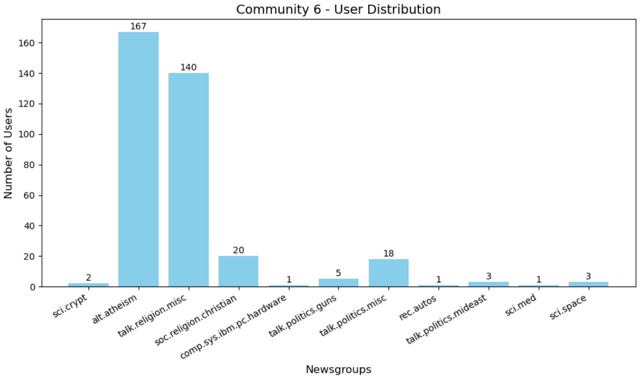


Fig. 6. User Distribution of Community 6

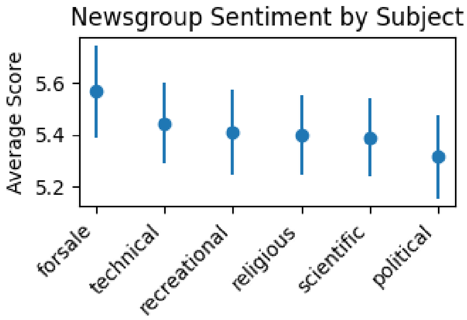


Fig. 7. Newsgroups sentiment by subject. We observe that the *forsale* newsgroup has the highest sentiments, whereas the partitioned political newsgroups have the lowest sentiment. The other four partitioned newsgruops have nearly identical sentiment, which is also somewhat surprising as they differ widely in subject.

Community Sentiments. Now that we have the sentiments for each newsgroup, we can use them as a comparison metric for the sentiment analysis of the communities. The highest sentiment is that of community 17 with 5.683 (±0.854), which has users who have only posted in the newsgroup regarding Windows operating system. Community 3 has the lowest sentiment score of 5.350 (±1.147) and includes users with interests in guns, politics, religion, space, and other topics.

LDA and latent topics. Latent Dirichlet Allocation (LDA) is an algorithm that can discover "hidden" topics in any text by counting which words frequently appear together. The output of the algorithm is a list of the most frequent words for each of the self-discovered n topics. For example, we can set n=6 and execute the LDA-algorithm on our entire dataset to discover 6 latent topics which may or may not be the same word distributions as our partitioned newsgroups. Noteworthy findings include Topic 5, where the top words are *"armenian", "turkish", "israel", "israeli", "arab", "jew", "turk", "said"* and *"war"*. Without using any of the newsgroup labeling, the computer found these terms to be especially tightly linked. Now compare that to Topic 0, which is *(would, one, get, like, dont, time, good, know, im, could)* or Topic 4 (people, one, would, god, say, think, dont, know, u, right) which from a human standpoint seems much more random. In that context, it is remarkable that Topic 5 is a word clustering that both humans and the LDA algorithm agrees belong together.

Discussion

Here we present the key insights gained from the analysis. Starting with the individual newsgroup networks, the first thing to be noticed is that almost all networks are balanced in terms of node numbers, except for newsgroup *for sale*, which is a very small network and this is understandable, as we can assume that people in this category mainly communicated with specific users who were selling something. Thus, a big network was unlikely to be created here. Furthermore, looking at the top nodes/users by degree in each newsgroup network, we notice some interesting patterns. In car and computer related topics, the out-degree of the top users is, in most cases, significantly higher than their in-degree. This indicates that they initiate interactions more frequently than they are contacted/messaged by others. This could be happening because these users have the knowledge to give advice or solve technical problems that other users have, related to these certain topics. In general, the in/out-degree distributions tent to be either relatively balanced or the out-degree being higher, which means that the top users of many newsgroups mainly reach out to people, but it is, in the majority, a back and forth conversation.

Looking within each of the individual networks, we looked at the communities that might arise in each newsgroup. Here, most of the networks have a fairly high modularity, the lowest of them still being over 0.4, showing that communities within them do seem to appear. This could be because the communities are created by small back and forth interactions between users, meaning they often reference each other in their posts. Some users could also represent one single often referred to post, meaning their e-mail is often represented in other users' posts.

Moving to the large unified network, it is a very sparse network with a *Density* of 0.0011, which is a common characteristic among real-life social networks. Looking at the degree distribution, we can see that the degree exponents are all along the lines of $\gamma > 3$ for either in- or out-degrees, as well as total degrees. This means the network is more or less random, or at least hard to distinguish from a random network of the same size(5). This mirrors the network for emails, where it still follows a power law distribution, but is more or less random. Another characteristic that indicates the existence of a real social network, is the high level of local connectivity between the users (*Average Clustering Coefficient*: 0.4731). That is how local communities are formed.

And talking about communities, we can draw multiple conclusions from analyzing the communities derived by the Louvain method. First and foremost, we can see that the *Modularity* is very high (0.85), which indicates that the network has well-defined communities with strong structures. This was an expected result, if we think the way this network was build (combination of smaller networks through common nodes). Looking at the community user distributions, we can observe that users in them predominantly talk/post about a single topic, except for some specific communities. In these specific communities, the users talk/post about relatively similar topics. For example, in community 14, the users are talking about computer hardware, software and electronics or in community 6, they are talking about atheism, Christianity and politics. Thus, we can conclude that the communities contain users with the same interests like technology, social matters, sports etc..

Regarding the text analysis of each community, it can help us understand what the main context of the posts is, without even looking at the user distribution. Especially for simple word frequencies, understanding the topic is very easy by looking at the top 10 words of a community. On the other hand, looking at the words with the highest TF-IDF scores, does not provide as much insight into the topics and interests of a community as we initially thought. For most of the communities, we get terms that could be characterized as key words for the topic of a community, but there are also some words like names or other nonsense words that cannot help us distinguish one community from another. For example, in communities . Thus, we can use this only as a complementary analysis

Last but not least, we have the sentiment analysis. With this analysis, we can understand what is the overall "mood" of each community. To justify the results, we are going to use also the sentiment analysis of the newsgroups. It can be noticed that when the interests of a community are politics (especially Middle-Eastern) or guns, there is a negativity in the posts, while for Christianity or technical topics there is a positivity. There are also some examples, like community 19, which ranked third in overall sentiment and the context of the posts was about religion. Although religion ranks in the middle to low range in terms of sentiment, this community is ranked quite high. This indicates that some communities are also divided in terms of sentiment. By this, we mean that certain communities contain users who interacted with each other and the context of their conversation was positive, even though the general topic of their conversation may have a low sentiment score.

Conclusion

In terms of our three research questions mentioned at the start of the paper, we have reached the following conclusions:

How did users form communities across different newsgroups, and are they aligning with the predefined newsgroup categories? While users often stayed within their primary interest areas, we found significant cross-posting between related topics. The 21 communities detected by the Louvain method showed that users tended to group around broader interest areas rather than strictly following newsgroup boundaries.

What were the main topics of discussion within and across these communities? Our text analysis revealed both highly focused single-topic communities and broader multi-topic discussions. The LDA analysis uncovered some surprising topic clusters, showing deeper patterns in how discussions naturally organized themselves.

How did sentiment vary across different topics and communities? We found statistically significant variations in sentiment across topics, with political discussions being notably more negative than buy-and-sell discussions. Community sentiments generally reflected their primary topics, though some communities showed unique emotional characteristics independent of their subject matter.

Overall, we found that users don't necessarily tend to stay within the same topics. Many users post in different topics, and because of this, communities do seem to arise based on shared interests. Additionally, our sentiment analysis found some of the newsgroups to be more unhappy than others, with the political communities being the most unhappy, while the buy and sell and technology related communities seemingly using the most positive language.

Contributions

Max: Data pre-processing and preliminary network analysis
Ioannis: Network analysis and community analysis
Jonatan: Newsgroups text, sentiment analysis and LDA

Jupyter notebook

Link to our notebook:
https://github.com/maxx1559/socialgraphs-final/tree/main

References

1. Ken Lang. The 20 newsgroups data set. http://qwone.com/～jason/20Newsgroups/, 2007. Accessed: 2024-12-10.
2. Peter Sheridan Dodds, Kameron Decker Harris, Isabel M. Kloumann, Catherine A. Bliss, and Christopher M. Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. https://doi.org/10.1371/journal.pone.0026752, 2011. Published: December 7, 2011, accessed 2024-12-10.
3. Chris Crawford. 20 newsgroups: A collection of 18,000 newsgroup documents from 20 different newsgroups. https://www.kaggle.com/datasets/chriscc/20-newsgroups, 2017. Accessed: 2024-12-10, updated 7 years ago.
4. Ioannis Vlasakoudis, Max Heiberg Bestle, and Jonatan Rasmussen. Python notebook with code for this project. https://github.com/maxx1559/socialgraphs-final/tree/main, 2024. Accessed: 2024-12-10.
5. Albert-Lazlo Barabasi. Network science. https://networksciencebook.com/chapter/1, 2016. Accessed: 2024-12-10.