CHAIR OF COMPUTATIONAL SOCIAL SCIENCE AND BIG DATA

Polarization in LGBTQ

Group 8

Winter semester 2021/22

Dominik Urban Michael Chan

1 Topic

In recent years the growing amount of data makes it difficult to obtain the relevant and desired information. Making that even harder: the data is mostly unstructured. This is especially apparent in today's social media. Anybody with internet access can create accounts (usually without any fee), write posts and comments as users on online platforms.With this in mind, our research group was wondering what specific topic out of numerous topics we should tackle.

In the last few years, the community of LGBTQ people is growing steadily. Alongside the increase in the size of the LGBTQ community and support for and within the community, the voice and online presence of LGBTQ members also grew. With this ongoing growth of LGBTQ hate crimes against the community are also increasing. This applies to the real world and the web. Online speech often contains foul language. Additionally, given the ethnic background of us being Polish, Hungarian and Chinese it made even more sense to choose this topic. Why would that be the case? Because Poland, Hungary and China – among other countries – are relatively conservative. Regarding these countries the topic of LGBTQ is still controversial and not necessarily supported[1, 2, 3].

Therefore we chose the polarization in the LGBTQ community on social media platforms as the topic for this project to not only show the growing support for LGBTQ but also the presence of hate and toxicity against LGBTQ.

2 Research Questions

For the chosen topic of polarization in LGBTQ in online social media, our research group first wants to find out what words are most frequently used by the LGBTQ community. This is the first step to take a closer look at what the community is usually conversing about. To provide more context for the used words we also want to figure out what topics are mainly discussed in LGBTQ online forums. As shortly mentioned in the introduction the LGBTQ community is facing hate crime and verbal abuse in real life – but is the community also facing hatred online? And how much hate is present?

To sum up our research questions:

What words are most commonly used by the LGBTQ community (**RQ1**)?

What topics are discussed in LGBTQ online forums (**RQ2**)?

Is the LGBTQ community encountering hatred and profane language online (**RQ3**)?

To answer these questions we first need to gather data from suitable social media platforms.

3 Data

We chose to collect and analyze data from popular online forums **Reddit** and **4chan**. More specifically we will be looking at the five biggest and most relevant LGBTQ subreddits[4] and the **4chan board /lgbt/**[5]. A subreddit is considered relevant if the posted content does not mainly consist of pictures.
The **r/lgbt**[6] subreddit is the biggest LGBTQ subreddit with 855.036 members. It's *"a safe space for GSRM (Gender, Sexual, and Romantic Minority) folk to discuss their lives, issues, interests, and passions. LGBT is still a popular term used to discuss gender and sexual minorities, but all GSRM are welcome beyond lesbian, gay, bisexual, and transgender people who consent to participate in a safe space."* This makes the subreddit a fitting candidate for our data.
The second biggest subreddit with 451.079 users is **r/bisexual**[7] which is a group for discussion and support for those who identify as bisexual, pansexual, omnisexual, queer, non-straight, etc.
Another subreddit is **r/actuallesbians**[8] which is a place for discussions for and by people who are bisexual women, lesbians, bi-curious people, cis, trans lesbians and anyone else interested. This subreddit has 384.159 members.
The next subreddit with 243.165 users is **r/askgaybros**[9]. This is where anyone can ask gay men for their opinions on various topics. Because this subreddit contains a lot of discussions, it naturally has lots of text data to be gathered and inspected.
The fifth subreddit – **r/trans**[10] – has 243.165 redditors. As the name implies, this subreddit is about everything regarding transgender topics.

Subreddit	Number of users	Number of comments	Creation Date
r/lgbt	855.036	128.902	14.03.2008
r/bisexual	451.079	64.539	14.04.2009
r/actuallesbians	384.159	50.882	13.09.2009
r/askgaybros	303.892	78.450	02.11.2012
r/trans	243.165	87.461	09.05.2011

A further source of data is the **/lgbt/ board**[5] on 4chan. 4chan is known for its harsh language and not restricted nature of banning users for offensive comments. It is delivering the LGBTQ community a safe ground since the 18th March 2013 to voice their opinion inappropriately without fearing a ban. The topics discussed differ from simple gay relationship advice to hating on LGBTQ subgroups. Another advantage and disadvantage for us are that users can comment anonymously but that restricts us from getting a user count for the board. But to examine if the board is active we used 4stats.io and with around 13 posts per minute, 22 threads an hour and 14000 posts a day on 31.01.2022 the /lgbt/ board on 4chan remains an active forum[11]. Another problem with 4chan was that the platform is not saving threads for a long period and is replacing them with newer ones so we had to collect threads and comments daily. 4chan also has no specific topic in a header for new threads and users have to set a picture to introduce a topic. Therefore we had to get a topic out of the whole conversation context.

4chan board	Number of comments	Creation Date
/lgbt/	88.758	18.03.2013

4 Methods

First, we collected data from the aforementioned subreddits and the 4chan board /lgbt/ using the Reddit and 4chan API. For Reddit we make use of the **PRAW** package (Python Reddit API Wrapper)[12], for 4chan respectively we utilize **4chan's read-only JSON API**[13]. After the data has been obtained, preprocessing has to be performed. The following steps explain the preprocessing pipeline. Using Python's Natural Language Toolkit (**NLTK**) we can get a list of stopwords (words we want to filter out, e.g. pronouns, prepositions, common verbs, etc.). More characters we want to remove include punctuation, quotes and other special characters (e.g. /, @, $, etc.). We also make every letter a lower case letter. After the first preprocessing attempt, it became apparent that we have to add some additional words to the list of stopwords. By refining the preprocessing pipeline the first major step is done.

Google Perspective APIIn order to answer RQ2 – what are the discussed topics in LGBTQ online forums –, we perform topic mining. Our goal is to find out what topics are discussed in our documents (e.g. our subreddits; each subreddit can be represented as a document) and what topic is primarily present in each document. Using **Latent Dirichlet Allocation** (**LDA**)[14] we obtain the desired results. We accomplish this by using the gensim package in Python.

We used the same methods for 4chan but did two different operations on our data. First, we used every thread as a document for the topic mining and then we used the whole board to get a topic for all threads together and not separately.

For the measurements of toxicity in online LGBTQ communities we established a dictionary of bad words. We gathered entries for this dictionary through already existing online dictionaries[15] and refined the list by removing and inserting relevant terms. Then we checked each word from the obtained data if it is part of the dictionary.

For another toxicity measure, we used the **Google's Perspective API**[16] because we got access to 1200 queries per minute and not only the default 60 queries per minute. Our dictionary search also left many new combinations of swear words out like *"borschtn*ggers"* and didn't check if the context of the word is in a negative sense or positive, e.g. if we used *"fuck"* as a bad word and someone wrote under a picture *"I would fuck you"* it would mean that someone looks good and not mean the same as the *"Fuck you"* we know as an insult. Therefore we used the more advanced method. With the Perspective API, we could determine if a posted comment under a 4chan thread or subreddit post tends to be toxic or not.

Using these methods we will answer the previously determined research questions.

5 Results

As a general overview of the most frequently used words, we created a word cloud [Fig1] using Python to answer RQ1. This word cloud depicts the 50 most frequently used words from all five subreddits. As someone would expect words like *"love"* or *"gay"* are within the top percents of most used words. Our team was first somewhat surprised that the word *"people"* appears regularly in the online discussion of the five subreddits. But after some consideration, *"people"* is not that much of an outlier. LGBTQ is all about humans, a connection between individuals and the interaction of people. Therefore it is perfectly reasonable to find *“people“* among the top most frequently used words.

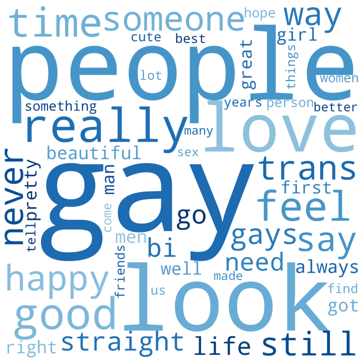

Figure 1: Reddit word cloud

Next up are the results from topic mining [Fig2] using LDA. We determined 5 major topics using the five chosen subreddits; each topic was determined with five words. Some interesting observations include the fact that often some words from topics overlap. Among these words are *“love“*, *“people“* and *“gay“*. Taking a peek at our previously depicted word cloud, those three words also occurred at the top of our most used words. In fact *“love“* and *“people“* can be found in every single topic. *“Gay“* appears in 4 out of 5 topics.
It is noteworthy that the five topics are pretty similar. This is because we considered each subreddit as a document which is why we expect subreddits, which are about LGBTQ, to have similar topics. We would probably receive more refined topics if we use each submission as a document for topic mining. Later on we used this strategy for 4Chan where we actually receive more distinguished and unique topics.

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
look	bi	love	gay	look
people	people	lesbian	people	love
love	bisexual	people	time	trans
gay	love	gay	really	beautiful
happy	gay	women	love	people

Figure 2: Reddit topics

Another mentionable finding is that the topics mostly match the general topic of each subreddit [Fig3]. For example, Topic 2 contains the words *“bi“* and *“bisexual“*. This topic has a 91% match with the subreddit r/bisexual. The third topic references the word *“lesbian“* which also matches the subreddit r/actuallesbians. The term *“trans“* also matches with topic 5 and the subreddit r/trans. This concludes RQ2 for Reddit.

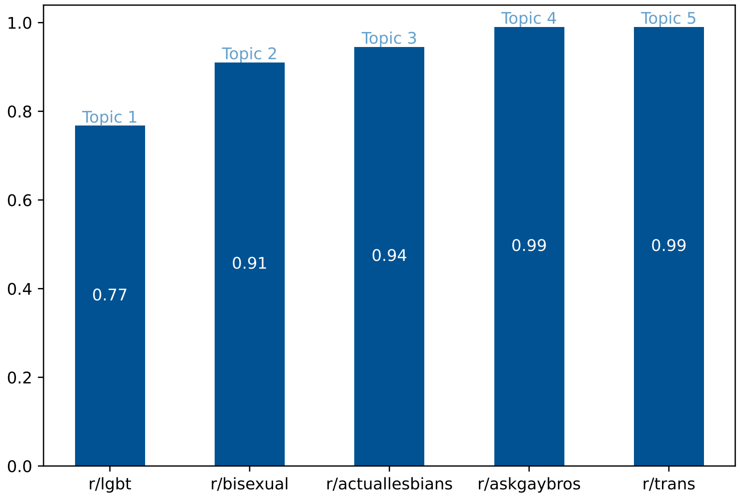

Figure 3: Coverage of main topic in each subreddit

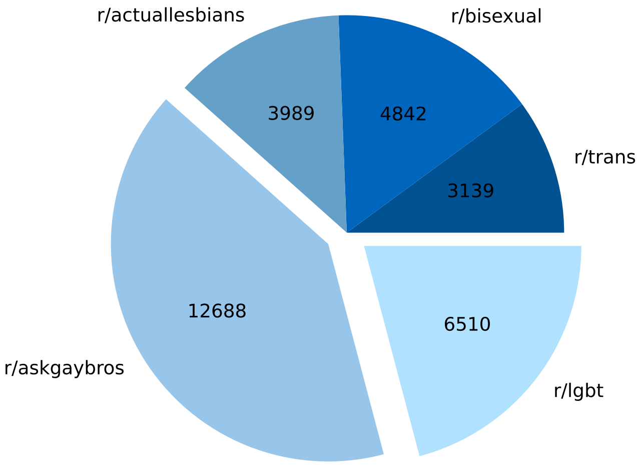

Figure 4: Number of bad words per subreddit

To answer RQ3 for Reddit we used the aforementioned dictionary of profane language and checked how many bad words occur. Out of 3517222 total words, 31168 of them were profane. This means about ~0.9% of all used words are considered offensive. This seems like a relatively small portion of online discussion in LGBTQ subreddits actually contains hate speech. Out of these 31168 words we analyzed how many words originate from which subreddit, which is depicted in the piechart above [Fig4]. The majority of words, 12688 (~41%), stem from the subreddit r/askgaybros. The second largest fraction with 6510 words, ~21%, is r/lgbt. The remaining 38% are made from r/actuallesbians (3989 words or ~13%), r/bisexual (4842 words or ~14%) and r/trans (3139 words or ~11%).

To answer RQ1 for 4chan we also made another word cloud [Fig5]. We can find similarities to the Reddit word cloud in words like *"people"*, *"good"*, *"really"*, *"look"* but we also can find 4chan typical words like *"anon"* that stands for the anonymous commenter on 4chan. We can also find that people write more about "hrt" on 4chan. The hormone replacement therapy seems to get addressed more on 4chan because people tend to write in a more "Not Safe For Work" manner with "NSFW" pictures of their body after an x amount of time in the *"hrt"*.

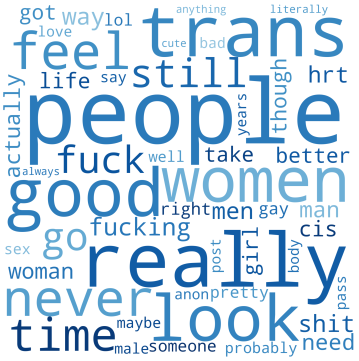

Figure 5: 4chan word cloud

For the RQ2 we also used the same methods as before for Reddit but we did two different analyses. For [Fig6] we used all threads as one document so we got 5 more diverse topics and for [Fig7] we used the whole board as one document and got 5 similar topics that's why we will consider only [Fig6] now. As we can see we have a wider range of different topics compared to Reddit but we also have 3 topics namely the first, second, and the fourth that can be counted to general discussions about the LGBTQ community because of the words *"people"* and the different sexual orientations (e.g. *"trans"*, *"gay"*). But if we look at the third and fifth topics we see - as mentioned before in the word cloud - that discussions about changing the gender are far more present. Not only in the word cloud but also for the topics we can find *"hrt"* again. We assume that the 4chan board gives people a good opportunity to discuss the positive but also the negative sides of the *"hrt"*. We have on one topic words like *"diaper"* and *"kms"* (Kill my self) that got used in discussions about the *"hrt"* and show that people can suffer from incontinence and underline it with *"kms"* to show that it's not only an opportunity to show someone's true self but can have some risk to it.

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
really	look	hrt	people	diapers
people	pass	dose	trans	boymoder
good	people	levels	gay	diaper
feel	really	take	woman	kms
time	good	estrogen	men	boymoders

Figure 6: 4chan topics for every thread on their own

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
people	people	people	people	people
really	really	really	good	really
trans	good	trans	really	good
shit	trans	good	look	trans
fucking	go	look	time	look

Figure 7: Topics for the whole 4chan board

Different from Reddit we didn't check the coverage of a topic to its subreddit but rather counted how often for the different threads the given topics get mentioned in [Fig8]. Topic 1, 2, and 4 got assigned nearly to all threads because as we mentioned these are the general topics that get discussed in the LGBTQ community. We can also see that nearly 30% of our threads discussed the topics about *"hrt"* which again shows us that 4chan gives a good ground for discussion about the *"hrt"*.

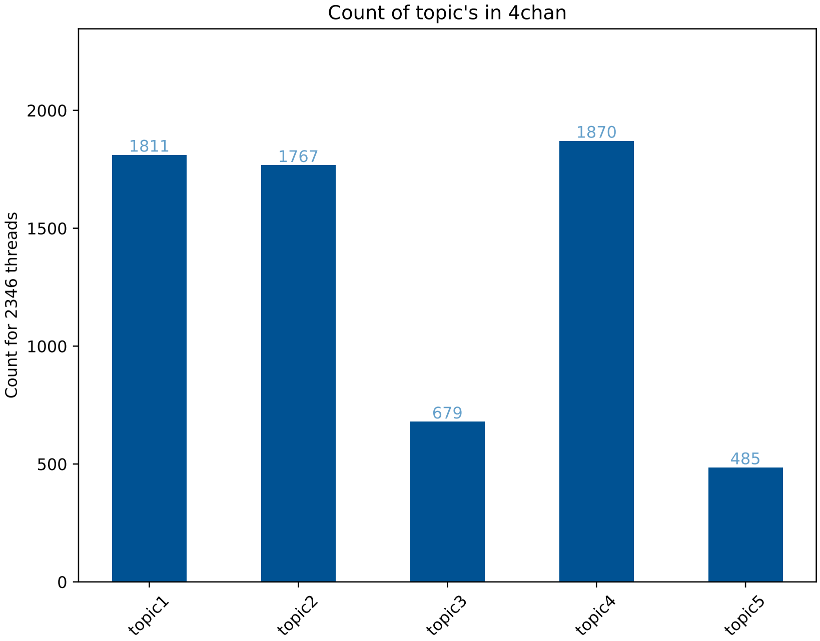

Figure 8: Topic count for 4chan

To answer RQ3 better than before we used the **Google's Perspective API**[16]. First, we looked at the scores for all threads from Reddit and 4chan together. In [Fig9] we see the scores assigned. 0.0 means a thread is neither neutral nor negative and rather very positive. A score of 1.0 would mean that a thread is the most toxic it could be and people will face either racism, profane language, or discrimination. But the LGBTQ threads are rather positive or neutral because most of the threads got assigned between 0.2 - 0.5. We also hand-checked our data and found out that the only comments with a score of 0.9 or more were 9 comments on 4chan and none on Reddit. That can tell us that the toxic comments get deleted on Reddit but not on 4chan. If we take a look at [Fig10] we see the average scores of the subreddits and the 4chan board. Surprisingly the 4chan board is not as toxic as expected before and only slightly tends to be more neutral than positive with the r/askgaybros subreddit but it is so slight that we can assume that the LGBTQ community for our analyzed subreddits and boards is a positive community without too much profane language.

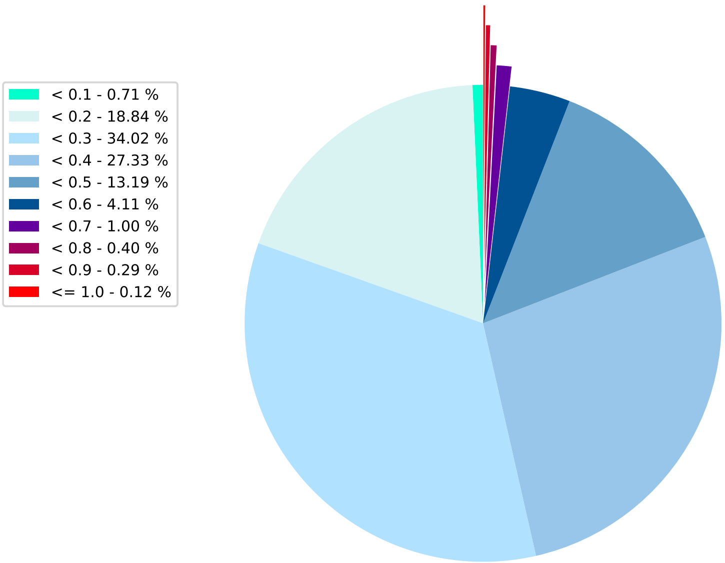

Figure 9: Toxicity score for all threads

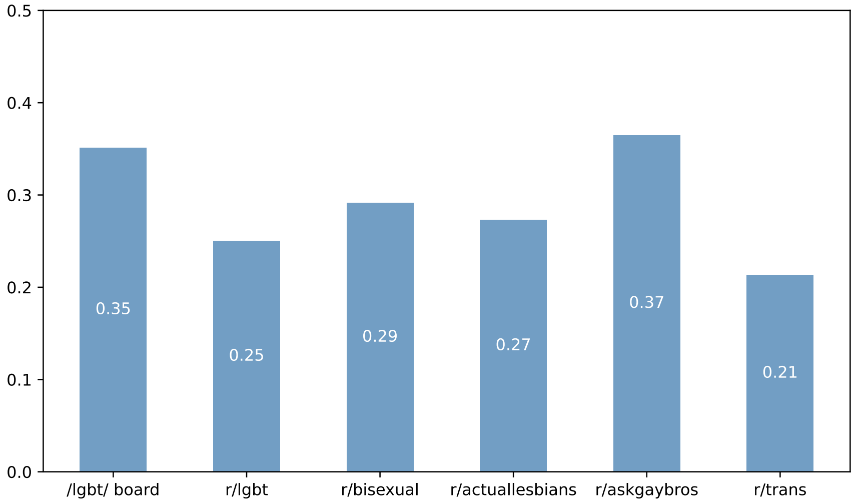

Figure 10: Perspective API average toxicity scores

6 Conclusion

Looking at our results there evidently are usages of profane language on the one hand. On the other hand, words are used with a positive effect like *"love"*, *"happy"* or *"beautiful"*. This means that polarization in LGBTQ in online social media is present to a certain degree.

Within the project, we also discovered several banned subreddits[17]. They were banned because of hateful speech and other inappropriate behavior towards the LGBTQ community. Deleted comments are also noteworthy. Further research regarding this kind of missing information could lead to even more insight.

The topic of polarization in the LGBTQ community in social media platforms is complex and much work can still be done in this research field. For example, we tried using Hatebase[18] but it was unsupported. Another API worth considering is The Weaponized Word[19]. It offers a dictionary of more than 5000 words and additionally a threat analysis tool.

Furthermore, it can be considered using other social media. Just a few examples are Facebook, Twitter and Instagram. Or using different or more subreddits is also a possibility.

It would also be interesting to look at certain parts of the world. Polarization in LGBTQ will certainly be different in Europe, North- and South America, Asia, Australia and Africa. Especially Asia and Africa would be interesting research areas because in some countries of those continents it is illegal to be gay.

Throughout the making of this project, we realized that there are a lot more possibilities in the research field of Polarization in LGBTQ. Using other data, different data analysis methods and various other APIs are just a few of them to continue research in this vast and complex research topic.

7 Contributions

All group members participated in the conception of the project idea "Polarization in LGBTQ".

Dominik Urban conceived the idea for using 4chan. He also performed the research, computations and analysis for 4chan.

Michael Chan conceived the idea for using Reddit. He performed the research, computations and analysis for Reddit.

Both Dominik and Michael discussed the results and contributed to the final poster and technical report.

References

[1] Pew Research Center. The global divide on homosexuality. https://www.pewresearch.org/global/2013/06/04/the-global-divide-on-homosexuality/.

[2] u/soratoumiga. Views of homosexuality by age group and coun-

try. https://www.reddit.com/r/dataisbeautiful/comments/phqxuc/oc_views_of_homosexuality_by_age_group_and_country/.

[3] Pew Research Center. Most central and eastern europeans oppose same-sex marriage, while most western europeans favor it. https://www.pewforum.org/2018/10/29/eastern-and-western-europeans-differ-on-importance-of-religion-views-of-minorities-and-key-social-issues/pf-10-29-18_east-west_-00-04/.

[4] Greg Seals. The best gay subreddits: Your guide to reddit's lgbt network. https://www.dailydot.com/irl/reddit-gay-lgbt-subreddits-gaymers-gaybros/.

[5] 4chan community support LLC. 4chan lgbt board. https://boards.4channel.org/lgbt/.

[6] Reddit Inc. r/lgbt. https://www.reddit.com/r/lgbt.

[7] Reddit Inc. r/bisexual. https://www.reddit.com/r/bisexual.

[8] Reddit Inc. r/actuallesbians. https://www.reddit.com/r/actuallesbians.

[9] Reddit Inc. r/askgaybros. https://www.reddit.com/r/askgaybros.

[10] Reddit Inc. r/trans. https://www.reddit.com/r/trans.

[11] 4chan statistics. https://4stats.io/.

[12] Bryce Boe. Praw: The python reddit api wrapper. https://praw.readthedocs.io/.

[13] Documentation for 4chan's read-only json api. https://github.com/4chan/4chan-API.

[14] Shashank Kapadia. Topic modeling in python: Latent dirichlet allocation (lda). https://medium.com/towards-data-science/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0.

[15] Code and lexicons for websci2019 paper: Exploring misogyny across the manosphere in reddit. https://github.com/miriamfs/WebSci2019/blob/master/Lexicon.txt.

[16] Google. Perspective api. https://developers.perspectiveapi.com/.

[17] Steven Asarch. Reddit bans transphobic superstraight subreddit for promoting hate. https://www.insider.com/superstraight-reddit-sub-subreddit-banned-super-straight-hate-2021-3.

[18] Hatebase Inc. Hatebase. https://hatebase.org/.

[19] The Weaponized Word. The weaponized word's threat analytics api. https://weaponizedword.org/.

All links were last accessed on 8th February 2022.