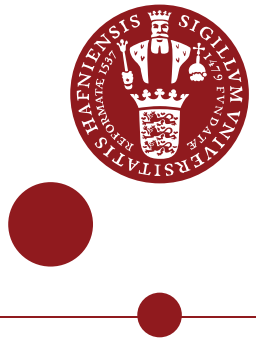
UNIVERSITY OF COPENHAGEN
DEPARTMENT OF ECONOMICS



A PAPER IN SOCIAL DATA SCIENCE

Frederik Ølund Larsen, exam nr. 245

Anne-Mette Landgren, exam nr. 230

Freja Christine Thim Hansen, exam nr. 42

WHAT'S UP WITH 4CHAN?

A text-analysis of 4chan's *politically incorrect* board /pol/

Contributions:

Exam nr. 245: 1, 2.3, 4.1.1, 4.2, 4.3.2, 5.2

Exam nr. 230: 2.1, 2.4, 4.1.2, 4.3, 4.4, 5.3

Exam nr. 42: 2.2, 3, 4.1.3, 4.3.1, 5.1, 6, 7

Seminar: Introduction to Social Data Science, 2200-S20

Submitted on: 28 August 2020

Keystrokes: 34,439

What's up with 4chan?

A text-analysis of 4chan's *politically incorrect* board /pol/

Frederik Ølund, Anne-Mette Landgren og Freja Thim

University of Copenhagen

28 August 2020

Online forums like 4chan become increasingly more prominent in shaping the opinion of especially young people. This paper investigates the topics discussed on 4chan's anonymous board /pol/, and how the debate evolves over time. To address this question, we collect over 150,000 posts from /pol/ and perform a text analysis using topic modelling and a Neural Network Language Model. The topics we identify can be labelled as anti-Semitism, racism, race, women, America, politics, history and war, Youtube and gaming as well as small hateful comments and a big *other* category. Our Neural Network Language Model successfully groups words by semantic similarities. The model reveals a vast use of discriminating language on /pol/.

1 Introduction

Over the past few months, there have been several events here among Covid-19, Black life matters and the upcoming US election that have led to huge discussions both in the public debate, social media and other internet forums. The online discussions opens up the debate to more people and more diverse voices. However, it also opens a discussion on who's reality is the truth. Donald Trump has been accused of spreading fake news but has also himself accused the established media of the same. Twitter are now taking new initiatives and are fact-checking posts to avoid fake news and misinformation to spread among users (Culliford & Paul (2020)). In our modern society, we increasingly rely on the internet for gathering and sharing information. Not all website takes responsibility for the content shared by users. This has enabled people with extreme political views and toxic behaviour to occur online at an unprecedented scale. A well-known forum for sharing extreme opinions is the board /pol/, short for politically incorrect, on 4chan. Forums like 4chan have over the past decades evolved from being a mean of communication to having an increasing impact on several aspects of today's society.

4chan is an anonymous imageboard website that hosts various boards dedicated to topics like anime, video games, sports, politics etc. The board /pol/ was created in end October 2011 and is, like the rest of 4chan, a popular place to spend time for teenagers and young adults. During the last decade, 4chan has been linked to the white supremacist massacre in Christchurch, New Zealand (Reuters (2019)), an anti-women terrorist attack in Canada (Kassam & Cecco (2018)) and cyberbullying of young individuals (Dewey (2014)). Furthermore Qanon, a conspiracy theory stemming from 4chan, is currently gaining followers and could possibly impact the upcoming presidential election (Roose (2020)).

Even though 4chan has 22 million monthly users, the discussions on the board are unknown to most people. This paper seeks to investigate the topics discussed on the anonymous board /pol/, and how the debate evolves over time. We address this issue in the following order. Section 2 describes how we gathered and cleaned the data. Section 3 talks about ethical considerations. Section 4 describes the theory behind topic modelling and word embeddings, which are used for the analysis in section 5. Section 6 discusses the drawbacks of our models and data. Section 7 concludes.

2 Data

2.1 Data gathering

To investigate posts from 4chan's board /pol/, we collect data using an API key from archive.4plebs.org. 4pleps is an unofficial archive of certain boards from 4chan, including the board /pol/. The API contains metadata on all original posts (OP's) and additional comments from /pol/ dating back to November 2013. We chose to collect both OP's and comments in our dataset to ensure that the topics and the wording reflects the users on /pol/. In the rest of this paper, the term *post* will refer to OP's and comments collectively.

Each day thousands of posts are created on /pol/. Due to the time constraint on this project and limited computer power, we sample data in the following way: (1) Data for OP's was gathered on the 25$^{\text{th}}$ of August 2020. We collect all OP's on a given page from 30-12-2013 to 23-08-2020, skipping 100 pages between each collection of OP's. We chose to skip 100 pages, as this ensures data from almost single every day throughout the period. The posting frequency will be reflected in our dataset as high activity periods automatically takes up more pages and we skip the same amount of pages each time. (2) Data for comments was gathered on the 23$^{\text{rd}}$ of August 2020. Each OP can have up to several hundred comments, together OP's and comments are called threads. The index-API only contains the latest five comments in all threads at a given page. This enables us to have a more diverse dataset. We skip 200 pages between each collection for comments in the period 29-12-2013 to 22-08-2020. The amount of skipped pages is larger than for OP's as there are considerably more comments per page compared to OP's.

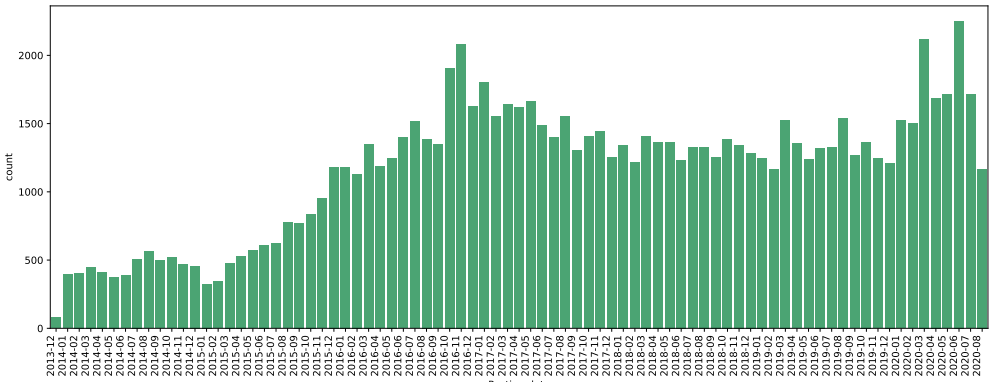
The two datasets are merged together for further data analysis. Figure 1 depicts the number of comments and OP's in our dataset. If we compare the two plots with the actual frequency of posts published on 4pleps (see figure in appendix A), we find that the overall trend is present in our dataset. However, the comments might undersample the presidential election in November 2016 if the treads from this month are very long compared to the average thread.

2.2 Data log

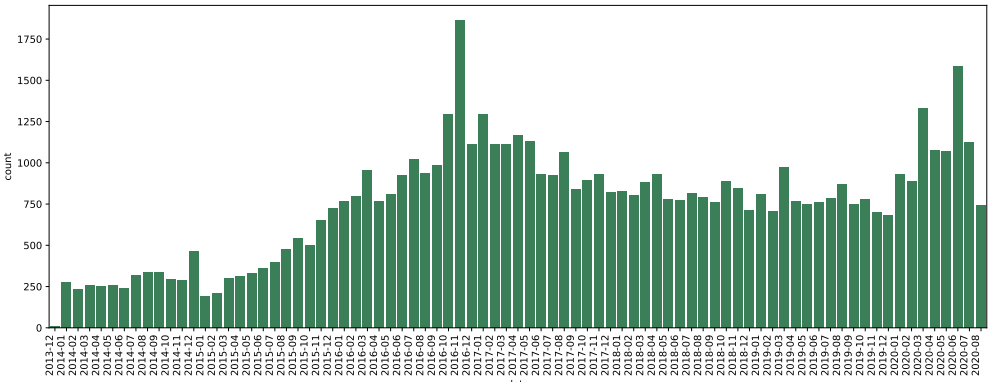
A Data-log were used to document the data collection. The data log makes the gathering of data more transparent and trustworthy as it keeps track of the connecting calls form the IP-address to the 4plebs API. Furthermore, it indicates what happened if an error occurs or if the program crashes. Examining our datalog, we find no data with systematic missing items (see attached logfile). Appendix B depicts our data collection process. The collection of comments took around seven hours and the collection of OP's took around eight hours. The call time looks relatively

Figure 1: Monthly activity on 4chan in our samples

(a) Number of comments



(b) Number of OP's



Note: Sample size of comment-data: 94,335 obs. Sample size of OP-data: 60,866 obs.

Source: Own data samples collected from 4plebs.org

stable. The size of the responses are equally plotted, and is approximately 100 KB throughout the entire period.

2.3 Data cleansing

In order to progress with the data, some initial data cleaning is needed. We remove all empty posts from the datasets, which reduces comments from 96,256 to 94,335 observations and OP's from 63,000 to 60.866 observations. Our analysis will be based on 155,201 merged posts from /pol/ covering the time-period November 2013 to August 2020. This final dataset weights OP's proportionally more than comments because of the data collection. This is done to reflect that OPs are read by more users and that they might be a better indicator of what topics are up for debate.

2.4 Preprocessing of data

As we are working with unstructured data preprocessing is required to obtain an applicable dataset for further analysis. We start by removing all links from the posts as we cannot categorise based on the links. Then we remove punctuation and transform all letters into lowercase. Furthermore, as many of the comments in our sample contain references to post id we also remove digits. We then remove stopwords from the dataset. We create our own list of stop words based on the NLKT stopwords. We add commonly used words on /pol/ such as slang for words already included in NLKT and other commonly used words that do not add any meaning to the post. At last, we lemmatise the dataset. The preprocessing is intended to ensure better results in the topic and coherent analysis.

3 Ethical considerations

The data is collected in accordance with several ethical considerations. Firstly, the data we analyse is composed of posts from 4chan where people write anonymously. Secondly, we chose to collect the data from the 4chan archive 4plebs rather than from 4chan directly, because we wanted to avoid illegal and unethical content. 4plebs removes child pornography, copyrighted material not permitted under fair use doctrine, and any information that identifies individuals according to the NIST definition (Wikipedia (n.d.)) (4plebs (n.d.)). One concern, which still remains is whether or not it is justifiable to include contents from the posts in this paper. Well aware of this issue, we mostly present frequently used pieces of posts. The included quotas are only incorporated to demonstrate the content of the non-classifiable clusters, and especially rude posts have been omitted, in order not to to promote these views. One could argue, that this paper still participates in spreading some of the hateful views from /pol/, by presenting counts and plots for the most frequently used discriminating words from 4chan. However, we believe it is important to shed light on what is debated on popular forums like this one, in order to understand how opinions for many teenagers are formed today.

4 Applied Methods

4.1 Tf-idf and k-means

4.1.1 Feature vectors

To categorize the topics discussed on /pol/ we apply *k-means* clustering to tf-idf vertorized data. Tf-idf is an abbreviation of the term frequency-inverse document frequency, which is a way to

transform text data into numerical feature vectors. A feature vector tells how often a particular word occurs in a post. The tf-idf method downweights these counts for frequently occurring words across posts, in order to assign uninformative, commonly used words less importance (Ian M. Smith. et al. (2001) p. 262). In this paper we compute the tf-idf vectorizer from the scikit-learn library, calculated as follows

$$\text{tf-idf}(t,d) = tf(t,d) \times (idf(t,d) + 1) \tag{1}$$

where $tf(t,d)$ is the term frequency defined as the number of times a word t occurs in post d. The inverse document frequency, $idf(t,d)$, is given as

$$idf(t,d) = \log \frac{1 + n_d}{1 + df(d,t)} \tag{2}$$

where n_d is total number of posts and $df(d,t)$ is number of posts d where word t is present. This amounts to the logarithm of the inverse share of posts this word occurs in, i.e. frequently used words across posts will have a larger denominator and a smaller $idf(t,d)$. To calculate the words, the vectorizer tokenizes the words in the posts. In this paper we chose to work with both 1-grams and 2-grams. This implies that we count the frequency etc. of each word and each combination of two words occurring next to each other in posts (Ibid., p. 2). 2-grams contain valuable information in our data, because terms like *alt* and *right* are somehow uninformative separately whereas *alt-right* unambiguously refers to *Alternative Right* a right-winged, social movement who supports Donald Trump (Ingraham (2017)). Before applying the tf-idf vectorizer, we preprocess the data as described in section 2.4. Furthermore, we choose to exclude words that occur in more than 80 pct. of the posts assuming these words are uninformative. In practise this restriction matters little, as we have already removed the most uninformative words by using stopwords. Likewise, we also exclude words with a document frequency lower than 0.1 pct. Finally, we only build a vocabulary based on the 3000 most frequently used words. These parameters are specified in the tf-idf vectorizer using max_df, min_df and $max\_features$.

4.1.2 Topic modelling

In order to do topic modelling, we need to apply some kind of unsupervised machine learning on our feature vectors. Unsupervised machine learning discovers hidden structures in data previously unknown to the ones who apply it (Ibid., p. 347). In this paper we work with the *k-means* clustering algorithm from the scikit-learn library. This is an easy and efficient way of clustering. Specifically, it identifies groups of posts more related to each other than to other posts (Ibid., p. 348). We apply the *k-means++* algorithm which improves the placement of the initial centroids and thereby improves the clusters compared to the simple *k-means* algorithm (Ibid., p. 353). To

get the *k-means* clustering algorithm working, we need to decide a number of clusters. In our case we choose to have 18 clusters. The *k-means* algorithm works in the following way (Ibid., p. 349):

- Place the initial centroids far away from each other by using the *k-mean++*.
- Assign each post to the nearest centroid $\mu^{(j)}$ for j = 1,2,...
- Recalculate the centroids as the center of the cluster formed by the posts
- Repeat the two latter steps until the algorithm reaches the maximum number of iterations, in our case we specified it as 300, or until the cluster assignments are stable.
- Run the *k-mean++* clustering algorithm multiple times with different initial centroids and choose the results that reaches the smallest SSE. We choose 25 different centroid seeds by specifying the n_init parameter.

For each repetition the *k-means* algorithm minimizes the Sum of Squared Errors (SSE) (Ibid., p. 350).

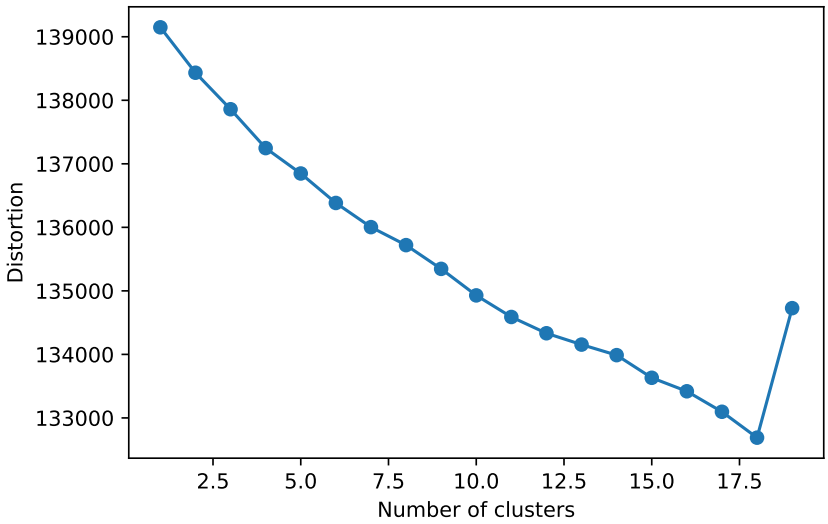
$$SSE = \sum_{i=1}^{n}\sum_{j=1}^{k} w^{(i,j)}||x^{(i)} - \mu^{(j)}||_2^2 \tag{3}$$

The latter part of the expressions, $||x^{(i)}-\mu^{(j)}||_2^2$, is the Euclidean distance in m-dimensional space, where $x^{(i)}$ is the feature vector of each i post and $\mu^{(j)}$ is the centroid of each j cluster. $w^{(i,j)}$ is an indicator function equal to one if the post $x^{(i)}$ is in the cluster j. The *k-means* algorithm minimizes the Sum of Squared Errors (SSE) by taking the first order conditions (FOC) with respect to $w^{(i,j)}$ and and $\mu^{(j)}$. The FOC with respect to $w^{(i,j)}$ assigns the post to the nearest centroid. The FOC with respect to $\mu^{(j)}$ follows by computing the new centroids of the clusters based on new assignment. These, are the steps described in the list above.

4.1.3 The elbow method

One drawback with *k-means* algorithm is that it requires us to take an uninformed decision about the number of clusters. To handle handle this issue, we use the elbow method. This means plotting the number of clusters against the distortion of cluster, calculated using the within-cluster SSE (Ibid., p.357). The optimal number of clusters is at the point where the distortion stops decreasing rapidly. At this number of clusters, the posts will be closest to the centroids. The elbow method is presented in figure 2.

Figure 2: The elbow method



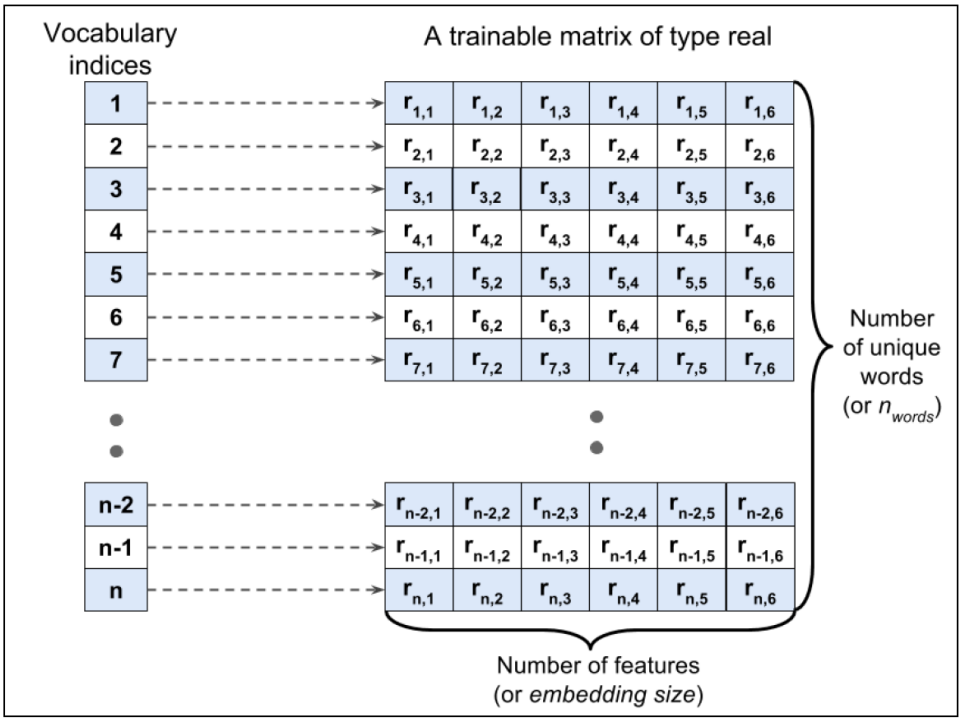
Source: own calculations on 4chan data collected from 4plebs

The distortion is in general quite high for our data, cf. figure 2. Surprisingly, the distortion increases from 18 to 19 clusters, which might indicate that either the elbow method or the k-means clustering is a bad fit for our data, a discussion that we return to in section 6. In figure 2 we see two small elbows at 4 and 14 clusters and then one at 18 clusters. After assessing the clusters' content with respectively 4, 14 and 18 number of clusters, we decide on 18 clusters as these topics are rather clear and the largest cluster takes up 64 pct. of the total posts. Even with 50 clusters the distortion is only slightly smaller than with 18 clusters but the cluster sizes are worse. Thus, we stick with the 18 clusters presented in table 2.

4.2 Word embedding

Words can be converted into input features in several different ways. Word embedding creates feature vectors for each word with a selected size as opposed to the more naive one-hot encoding which has n values corresponding to the size of the vocabulary. Furthermore embedding allows training of the model. Figure 4 depicts an embedding matrix with an embedding size of 6 and a vocabulary of n words. One-hot encoding would create n inputs in each row, therefore, word embeddings reduce the risk of the curse of dimensionality.

Figure 3: Caption



Source: Ian M. Smith. et al. (2001)

4.3 Neural Network Language Model (NNLM)

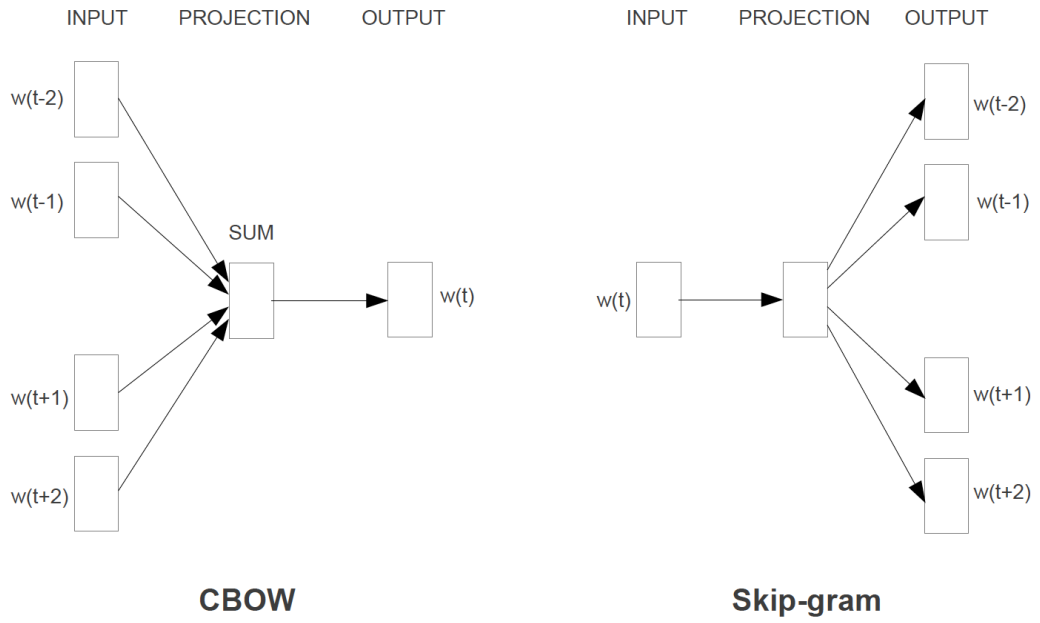
In 2013 google employees released the paper *Efficient Estimation of Word Representations in Vector Space* (Mikolov et al. (2013)). In the paper the authors introduce two new NNLM architectures. The Continuous Bag-of-Words Model (CBOW) and the Skip-gram model. These models were released under gensim as word2vec.

4.3.1 CBOW

The training criterion in CBOW is to correctly classify the flanked word, using context words. Since the order of the surrounding words does not affect the projection, the model can be thought of as a bag-of-words model like the tf-idf. The surrounding words are passed as input to a log-linear classifier which is used to classify the flanked word. The Skip-gram model can be thought of as the mirror image of CBOW. It uses the input word to predict its surrounding words. In figure 4 both models are depicted with a window of 2. In CBOW the window dictates how many preceeding and succeeding words will be used to classify the flanked word i.e the output. In Skip-gram the window dictates how many preceeding and succeeding words will be predicted by the flanked word i.e how many outputs will be formed based on the flanked word. Contrary

to CBOW, the order of words does matter in Skip-gram since words far apart are usually less related. The Skip-gram model uses a range $[1; C]$ where $C = window$. A random number R is drawn from the range which becomes the effective window size. Hence R words from the past, and R words from the future are being predicted by the input word. This effectively down weights predictions of words far apart.

Figure 4: NNLM models



Source: Mikolov et. al(2013) p5

4.3.2 Training of the Models

Rong (2014) provides a thorough explanation of the parameter updates and training of the models presented by Mikolov et al. (2013). Both the Skip-gram and CBOW uses gradient descent to train a set of weights from the input to the hidden layer and from the hidden layer to the output. Due to the limited scope of this project we will not elaborate further on the weight adjustment.

4.4 Model application

Our objective with the application of a NNLP model is to maximize semantic performance. We apply both the CBOW and the Skip-gram model to our data and contradictory to the results in Mikolov et al. (2013), we do not see any indications that the Skip-gram is outperforming CBOW. Model performance is evaluated by comparing results from vector operations and printing of most similar words. Our final model is a CBOW consisting of an embedding vector with 50 values,

and which only includes words with a higher total frequency than 10. We choose a window of 5, hence, the model evaluates up to 5 preceding and succeeding words to classify the output word. We performed 100 iterations over the corpus i.e 100 epochs. Below we print some of our model accuracy tests with corresponding scores:

$$Muslim = islam(0,76)$$

$$King - Man + Woman = Exception(0.55)$$

$$Trump - America + Russia = Putin(0.78)$$

The first test prints the most similar words to muslim according to the model. Results where Islam, Islamic, Sunni, Arab and invader. The second equation is a classic example of vector operations, however the words man, woman, king and queen are used in different contexts in our dataset than usually, hence results were poor. The model delivers accurate results in third operation and when using racist terms.

5 Analysis

5.1 Descriptive analysis

By looking at the monthly activity on 4chan, shown in figure 1, we find that the overall number of posts have increased from 2013 to 2020. Furthermore, we find that there are several interesting spikes. One of the more prominent spikes is in November-December 2016. This large increase in posts coincide with the election of Trump as president. Furthermore, we find an increased activity in 2020, which coincide with several societal events as the corona epidemic and Black Lives Matter.

Table 1 depicts the top 20 most frequently used words in our prepossessed data. Trump is the second most used word indicating a relative high interest in the current American president between 4chan users. In general we find that there is a lot of hateful and race-related words included in the list of most frequently used words ie. white, fuck, jew, nigger and black.

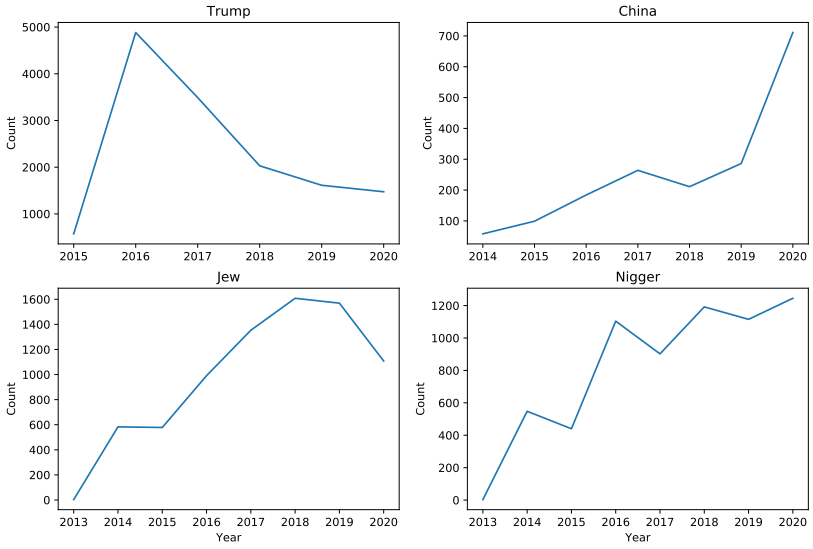
Table 1: Top 20 most frequently used words

Words	Count	Words	Count	Words	Count	Words	Count
people	17,934	time	8,279	year	6,934	world	6,033
trump	14,065	fucking	8,249	country	6,802	guy	5,685
white	13,855	jew	7,793	good	6,636	really	4,983
fuck	8,942	right	7,472	nigger	6,549	never	4,970
shit	8,352	woman	7,394	black	6,299	new	4,890

Source: own calculations on 4chan data collected from 4plebs

Figure 5 plots the development of selected words, without taking the general activity development into consideration nor adjusting for the fact that we have only 8 months of data in 2020. Trump was first mentioned in 2015, and was the most mentioned word on 4chan during the election period in 2016 - 2017. In 2020 the word trump were mentioned 1,474 times, indicating a high interest in Trump from 4chan's users. The words nigger and jew have been increasing for the entire period. This could indicate an increasing tendency of racism and anti-sematism posts on 4chan. The word China have increased dramatically from 2019 to 2020. We expect that this is due to discussions on Covid-19 specifically Trump addressing it as the China-virus.

Figure 5: Number of times chosen words is mentioned



Source: own calculations on 4chan data collected from 4plebs

5.2 Categorical analysis

In order to categorize the topics on 4chan's board /pol/, we predict the cluster of each post for 18 clusters. Table 2 shows the top seven most frequently used words within each cluster. Based on these words and by investigating the actual content of a sub sample of comments within each cluster, we assign topics to the clusters. The clusters range in size from 1,065 to 99,093 posts. Most clusters cover more than 2 pct and a single cluster, cluster 6, covers 64 pct. of all posts.

Even though most of the clusters make intuitively sense, we merge some of them into larger groups in order to investigate how these groups have evolved over time. Cluster 1 and cluster 12 constitute a group about warfare, history and politics. Cluster 2 clearly contains racist and hateful views. The same applies to cluster 10, though this racism is formulated a bit differently,

Table 2: Clusters, top words

Clusters	1: War	2: Racism	3: Gaming	4: Anti-Semitism	5: US politics	6: Other
	world	nigger	right	jew	trump	year
	country	fuck	guy	jewish	donald	realy
	war	white	kek_2	white	donald trump	country
	people	fucking	mean	israel	pres	never
	world war	hate	$warningmarquee_1$	hate	president	new
	america	kill	$marqueetrigger_1$	people	pres trump	back
Size	3,027	2,319	6,037	3,060	4,566	99,093

Clusters	7: Hate	8: Hate	9: Hate	10: White supremacy	11: Friendly	12: Historical politics
	fucking	fuck	shit	white	good	time
	fuck	$kike_3$	holy shit	white people	pretty good	every
	shit	guy	holy	people	good luck	every time
	stupid	wrong	fuck	black	luck	year
	retard	holy	fucking	race	goy_4	people
	hate	shut	people	white man	feel	last
Size	3,065	2,629	3,681	4,874	2,708	3,787

Clusters	13: Random	14: USA	15: Women	16: Race	17: Hate	18: Youtube
	love	state	woman	people	faggot	based
	wtf	united	men	black	fucking	$redpilled_5$
	much	united state	white woman	black people	fuck	fuck
	hate	absolute	white	country	op	man
	people	america	black	really	nigger	pretty
	guy	country	man	many	shit	fucking
Size	1,131	2,011	3,103	7,228	1,817	1,065

Note: 1: annoying HTML code, 2:Laughter, 3: degrading term for jew, 4:term for a non-Jewish person, 5:enlightened

Source: own calculations on 4chan data collected from 4plebs

best described as white supremacy. These posts state, in hateful way, how superior white people are to black people and people of color. Together cluster 2 and 10 form a new group called racism. We combine cluster 3, 11 and 18 which are about gaming and computer issues, friendly chatting and Youtube based discussions. We call this group Youtube and gaming. Cluster 4 contains anti-Semitic opinions. Although anti-Semitism is also racism, we leave this category on its own to isolate the racism related to the 8th most frequently occurring word in our data, jew. Cluster 5 is about American politics with Donald Trump as the most frequently used word. We combine cluster 5 with cluster 14 about the United States of America into a group called *USA*. It is worth mentioning that when we dig deeper into the posts cluster 14 contain quite a lot of Alt-Right posts. Cluster 7, 8, 9 and 17 add up to one group with mixed hateful comments. Many of these comments do not have any real content but are short insults and frequently use discriminating and racists words like faggot, kike and nigger, cf. table 2 for translations. Cluster 15 is about women, mostly either objectification or hateful comments against women. Cluster 16 is a broader category on race that contains racism but also anti-racist views and discussions on

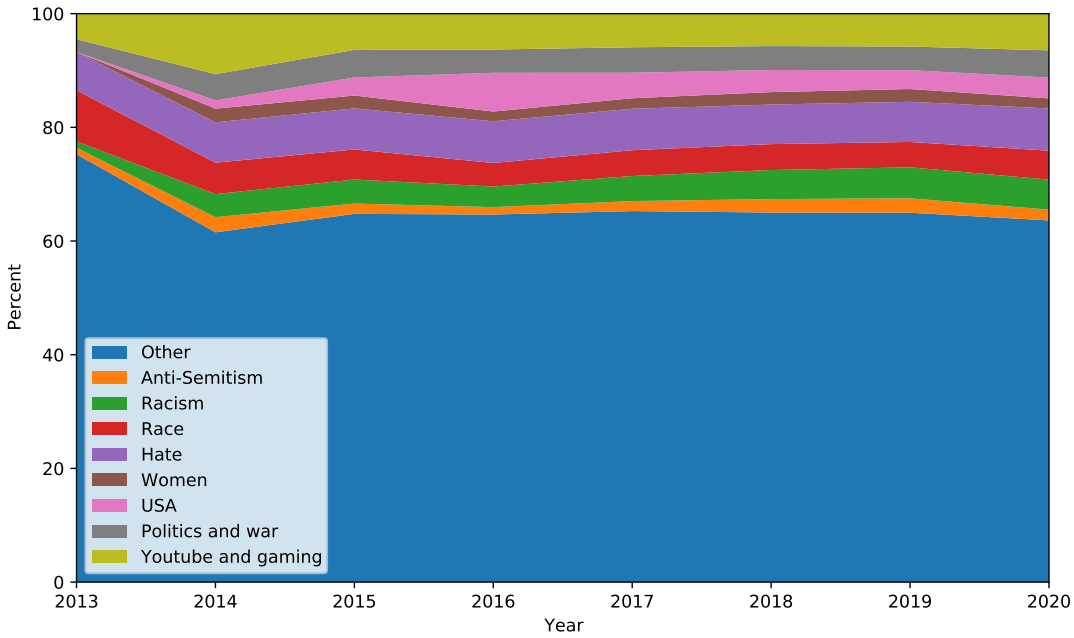
cultures. The remaining two clusters, 6 and 13, form a group of posts characterised as *Other*. To provide insight into the *Other* group, we randomly pick 3 relatively short posts below. Whereas the first two comments are somewhat hateful and written in a computer games lingo, the third comment is basically a discussion on whether or not longboard-skateboarding is cool. This shows how mixed the *Other* group is.

Post 1: *"Christians wont fight back so lets attack them, better not offend muslims or they might kill us - maybe we should become extremist"*

Post 2: *"Yank talking about religious extremist terrorists. HELL!HELL IS REAL! HELL IS WHERE WE FIGHT THEM! Hes sounds like a radical muslim..."*

Post 3: *"not voting LDP. Really, ausfags?"*

Figure 6: The share of posts in each group of clusters, year 2013-2020



Source: own calculations on 4chan data collected from 4plebs

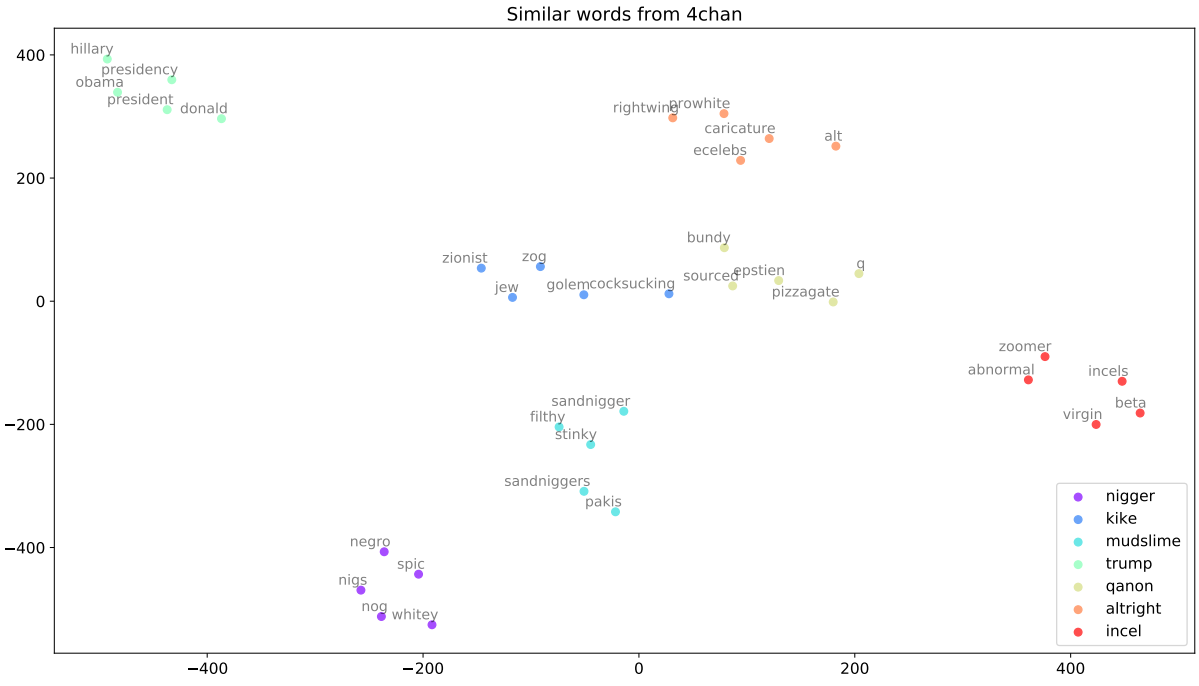
Figure 6 depicts the topics created based on topic modelling and how their shares evolve over time. Despite in 2013 where our data is limited, the topic *Other* is stable at around 64 pct. *Anti-Semitism* has an average of 2 pct. with spikes in 2014 and 2019. Similar for *Women*. *Racism* has slowly grown from 4 pct. in 2014 to 5.3 pct. in 2020. *Race*, *Hate* and *Politics and*

war have all been rather stable since 2014. *Youtube and gaming* took up 10 pct. of the posts in 2014 but on 6.5 in 2020. The share was actually smallest in the in 2016, where *USA* spiked with 6.8 pct. of all posts. This spike is correlated with the year that Trump ran for office and won the presidential election. /pol/ on 4chan is know for hosting many Trump-supporters, which the posts from *USA* to a large degree confirms (Ingraham (2017)). This is also in line with the large activity on 4chan in the month of the election, cf. section 2.

5.3 Coherent analysis

In this section we perform coherent analysis using a CBOW model. Coherent analysis can be used to gain insights about the meaning and context of the vocabulary and topics on 4chan. We use word embeddings to demonstrate words that have the same semantic associations.

Figure 7: Word embetting



Source: own calculations on 4chan data collected from 4plebs

In figure 7 we have selected seven words of interest and plotted their word associations. The word trump is associated with obama, donald, president, presidency and hillary. This is our baseline example to illustrate how the word association turns out in our model. The following words include frequently used insults and concepts from 4chan.

The word nigger is associated with negro, nigs, spic, nog and whitey. Negro, nigs and nog

are racist slangs for black people, while spic is an offensive term for spanish speaking people. Whitey is a slang for white people. The patronizing word for muslim, mudslime, is associated with sandnigger, filthy, stinky and pakis. The word kike, which is an offensive slang for jew, is associated with zog, zionist, golem, jew and cocksucking. The word associations for niggers, mudslime and kike, which are all discriminating against religions and ethnic miniorities, are relatively closely located in the plot.

The word qanon is a conspiracy theory, and is associated with q, pizzagate, epstein, bundy and sourced. Q is the anonymous poster of the Qanon conspiracy theory, and pizzagate is another closely linked conspiracy theory that states, that the power elite runs a paedophilia ring (Wong (2020)). Jeffery Epstein was a successful banker and recently convicted sex offender. Bundy and sourced can mean several things according to the urban dictionary, however, both can be used to describe sexual deviancy. In our plot Qanon is relatively close located to the word kike. The reasoning behind this might be, that Qanon belives that the power elite originates from the Protocols of the Elders of Zion, a fake document that claims to expose a Jewish plot to control the world. (ibid.).

The word altright, refers to a far-right, white supremacist movement supporting Trump called alternative right (Ingraham (2017)). Words connected to altright is rightwing, prowhite, alt, caricature and ecelebs. The first three associations indicates their political view, whereas caricature and ecelebs could be related in several ways. Ecelebs refer to a persons who have become famous on Youtube, a medium forum for alt-right.

The word incel stands for involuntary celibate and is a misogynistic online subculture, linked to alt-right and several violent attacks (Hern (2018)). Word associations form our model is beta, virgin, abnormal and zoomer. Beta refers to the incels themselves as opposed to the alpha-male who are dominating and gets to have sex with women. Abnormal can both refer to how the incels see themselvs, but also how the surrounding see them. Zoomer is the generation born late in the 90's to early 00's, indicating the typical age of this online subculture.

The fact that the model performs so well on discriminating words reflect how much they actually use these words on /pol/. On our data, CBOW is apparently also well suited to define term like *Qanon*.

6 Discussion

It is worth to consider how the data collection could affect the results. We only sample the last five comments to each OP, which could question the external validity of this analysis if the last five comments are systematically different from other comments. For instance, the last five

comments might be less controversial than the first five ones, since they these comments closed the debate. Furthermore, conducting a text analysis on our data from /pol/ might be challenged by the extensive use of slang and misspelling.

The clusters are rather informative about the topics on /pol/, in the sense that they roughly cover the topics we saw on /pol/ by eyeballing the posts ourselves, with the exception of an independent category of conspiracies. In general, it seems to be difficult for k-means to categorize the posts well into these clusters. We spend a lot of time trying to improve the clusters by adjusting the k-means and tf-idf parameters, as well as the numbers of clusters but we never managed to get rid of the large *other* topic. Even the elbow method was only vaguely informative about the optimal cluster numbers, which might indicate the type of data we work is not well suited for k-means clustering, because of the many comments containing mixed topics but only few words. This also leads to an underestimation of racism by the categories *racism* and *anti-Semitism*. Theoretically, k-means is efficient and good all around clustering algorithm, but k-means can be challenged when applied on high-dimensional data because the Euclidean distances can become inflated (https://scikit-learn.org/stable/modules/clustering.htmlk-means). This could challenge our clustering. We also tried applying Latent Dirichlet Allocation (LDA) but without any improvement in topics. However, we get similar topics to Papasavva et al. (2020), who applies LDA on a dataset with 134.5M posts from /pol/. For further analysis, the problem could be alleviated by either reducing the dimensions of the feature vectors before applying k-means clustering, or by simply using an algorithm developed for the purpose of topic modelling.

The word embeddings reflect how closely connected topics are in context, eg. that qanon and jew's are somewhat related on /pol/. This feature makes it possible for outsiders to understand the topics and language used on 4chan. A drawback to this is that the model is trained on a relatively small data set. Meaning that the connections created in the model can be based on few posts. We find that our model performs better when tested on insults compared to the usual man/woman example. This reflects a rather harsh language on the forum, where these insults are frequently used. We did not find a big difference between using the Skip-gram model and the CBOW model. Initially we expected the skip-gram model to preform better as it reaches more precise semantics according to (Mikolov et al. (2013)). However, they found the Skip-gram model particularly successful when using large windows of words. Our posts from 4chan are relatively short, and this might be the reason that the skip-gram model performs worse on our dataset.

7 Conclusion

This paper investigates the debate on 4chan's board /pol/. We find that the frequently discussed topics are anti-Semitism, racism, race, women, America, politics, history and war, Youtube and gaming as well as small hateful comments and a big *other* category. These topics have been rather constant over time but impacted by outside events like the presidential election in November 2016. The spillover from 4chan to the rest of society, and vice versa, is in accordance with most literature regarding 4chan. Our word embedding analysis confirms a discriminating language on the board /pol/. It also works surprisingly well to define users understanding of terms and concepts.

References

4plebs. (n.d.). *4plebs » FAQ.* Retrieved from `http://archive.4plebs.org/_/articles/faq/`

Culliford, E., & Paul, K. (2020). *With fact-checks, Twitter takes on a new kind of task.* Retrieved from `https://www.reuters.com/article/us-twitter-factcheck/with-fact-checks-twitter-takes-on-a-new-kind-of-task-idUSKBN2360U0`

Dewey, C. (2014). *Absolutely everything you need to know to understand 4chan, the Internet's own bogeyman - The Washington Post.* Retrieved from `https://www.washingtonpost.com/news/the-intersect/wp/2014/09/25/absolutely-everything-you-need-to-know-to-understand-4chan-the-internets-own-bogeyman/?fbclid=IwAR1qeinTY_i8vuSxy8aG7668SmqqpqDYpwVU3sGdVckVjGLUt1yneP2ZTyU`

Hern, A. (2018). *Who are the 'incels' and how do they relate to Toronto van attack?* Retrieved from `https://www.theguardian.com/technology/2018/apr/25/what-is-incel-movement-toronto-van-attack-suspect`

Ian M. Smith., Cook, D., & Smith., B. P. (2001). *Python Machine Learning* (Second Edi ed.) (No. June).

Ingraham, C. (2017). *The 'alt-right' is just another word for white supremacy, study finds - The Washington Post.* Retrieved from `https://www.washingtonpost.com/news/wonk/wp/2017/08/16/the-alt-right-is-just-another-word-for-white-supremacy-study-finds/`

Kassam, A., & Cecco, L. (2018). *Toronto man charged in 'horrific' van attack that killed 10 people | World news | The Guardian.* Retrieved from `https://www.theguardian.com/world/2018/apr/24/alek-minassian-toronto-van-attack-latest-news-suspect-charged`

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, 1–12.

Papasavva, A., Zannettou, S., De Cristofaro, E., Stringhini, G., & Blackburn, J. (2020). Raiders of the Lost Kek: 3.5 Years of Augmented 4chan Posts from the Politically Incorrect Board. (Icwsm). Retrieved from http://arxiv.org/abs/2001.07487

Reuters. (2019). *Christchurch mosque attack suspect pleads not guilty, trial set for next year.* Retrieved from https://www.nbcnews.com/news/world/christchurch-mosque-attack-suspect-pleads-not-guilty-trial-set-next-n1017466

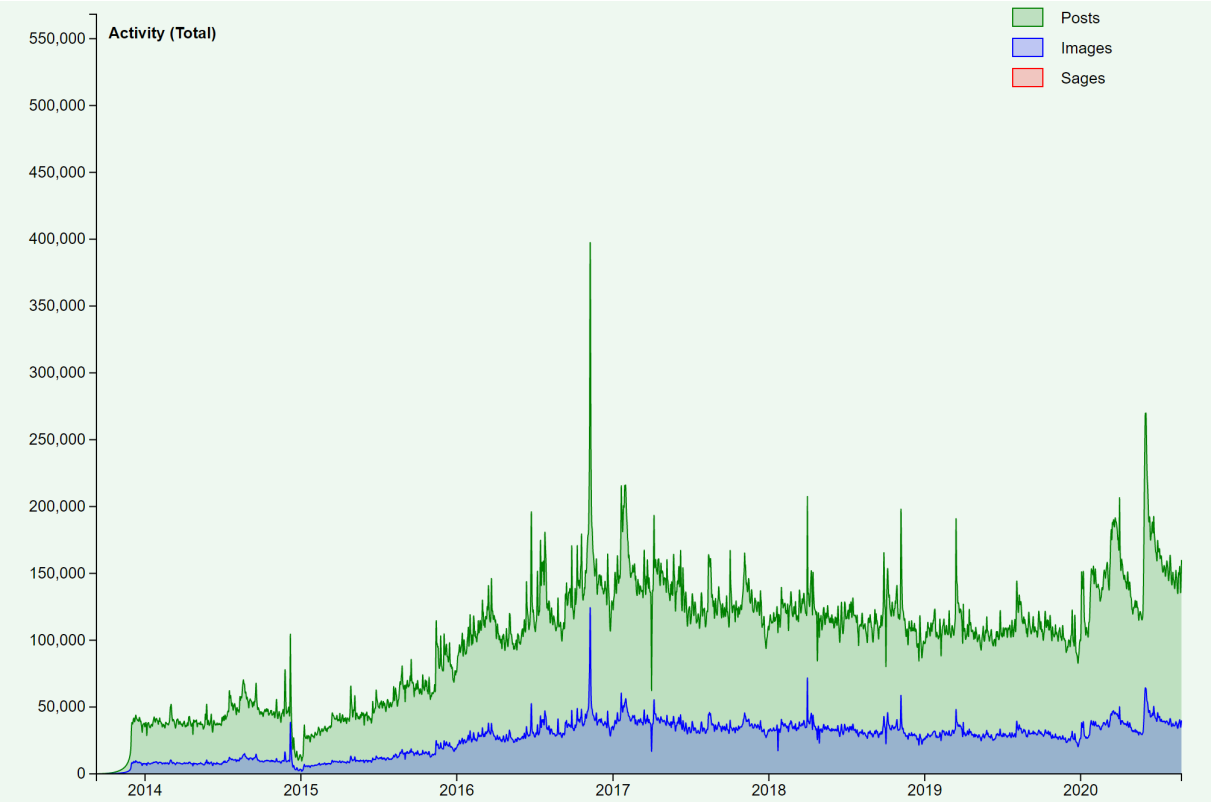
Rong, X. (2014). word2vec Parameter Learning Explained. , 1–21. Retrieved from http://arxiv.org/abs/1411.2738

Roose, K. (2020). *qanon: All about QAnon, the pro-Donald Trump conspiracy theory that is going viral as US polls near - The Economic Times.* Retrieved from https://economictimes.indiatimes.com/news/international/world-news/all-about-qanon-the-pro-donald-trump-conspiracy-theory-that-is-going-viral-as-us-polls-near/articleshow/77625603.cms

Wikipedia. (n.d.). *Personal data - Wikipedia.* Retrieved from https://en.wikipedia.org/wiki/Personal_data#NIST_definition

Wong, J. C. (2020). *QAnon explained: the antisemitic conspiracy theory gaining traction around the world | US news | The Guardian.* Retrieved from https://www.theguardian.com/us-news/2020/aug/25/qanon-conspiracy-theory-explained-trump-what-is

A Monthly activity on 4chan



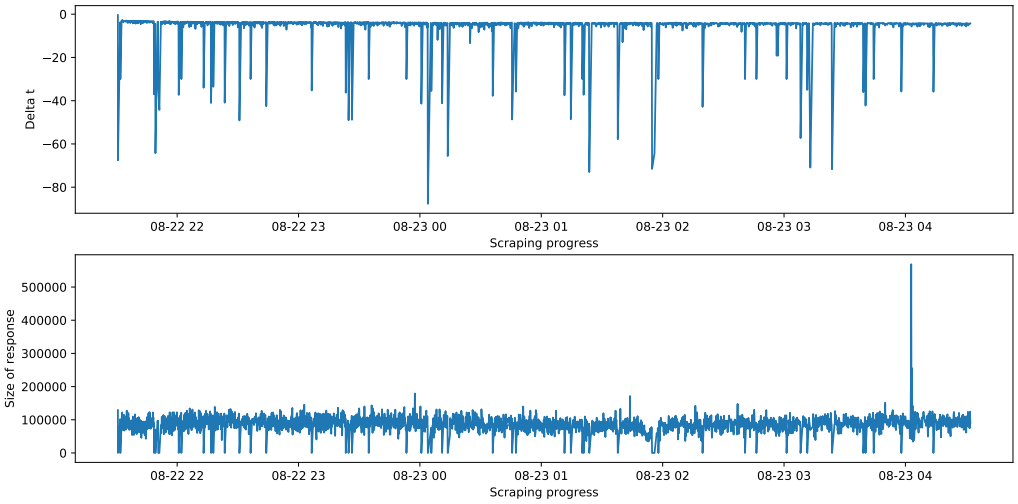
(a) Number of posts in total

Source: 4pleps, /pol/ activity statistics (*https* : *//archive.4plebs.org/pol/statistics/activity/*)

B Logfile

Figure 9: Log file

(a) Data gathering of Comments



(b) Data gathering of OPs