Political Orientation on Reddit

Jan Bachmann, Jan Philipp Hafer, Xiaoqi Ma, and Zain Selman

RWTH Aachen University,
52074 Aachen, Germany
{jan.bachmann,jan.hafer,xiaoqi.ma,zain.selman}@rwth-aachen.de

Abstract. When traversing the internet, e.g. on a search for information, one might come across content, which contains political bias from other humans. While news outlets are more or less trying to remain neutral, as part of their mission, a lot of content is created by private individuals, which have no or few concerns about their bias. One of those sites, where users can anonymously publish their opinion, is reddit, which is in the top 10 of Alexa's ranking system. Reddit features a large, publicly accessible dataset that suits this work well. In this work we want to analyse the political orientation of users on reddit using the two most notorious political boards. These boards, due to their large volume, offer the best possibility to analyse users' political orientation. We analysed language behaviour, cross subreddit activity, user behaviour, and temporal behaviour. Our results suggest that there are differences in certain behavioural patterns, and that its possible to predict political orientation with good accuracy using frequently used words.

Keywords: reddit, politics, The_Donald, analysis, orientation

1 Introduction

Reddit is among today's most used websites of the internet [1]. It describes itself as a collection of thousands of communities, allowing its users to socialize and discuss about various topics, be it sports, movies or politics. It claims that it has over 330 million active monthly users on average, surfing through more than 138 thousand different subreddits [2]. There is a specific subreddit (a community specialized on a certain topic) for nearly every topic one could imagine. Users can create and share content in forms of texts, pictures, links and comments, participate in discussion or just come and leave as a spectator.

The result of the 2016 presidential election in the U.S.A. left a substantial part of the mainstream media and population in shock. A subreddit called *The_Donald*, dedicated to the current president of the United States, Donald Trump, creates a space for discussions and content for his supporters [3]. As the subreddit is restricted to Trump supporters by rule, politically opposing content is non-existent. While *The_Donald* serves as a home for conservative leaning Trump supporters, the subreddit *politics* does the opposite [4]. Officially, it deals with all kinds of political news regarding the U.S. Realistically, most of its content is based around progressive/liberal leaning content, even though conservative posts are not banned explicitly. However, if a users decides to post Trump supporting content, the submission will most likely get downvoted rather quickly. This causes the post to disappear from the subreddit's front page, while progressive leaning posts will get upvoted more often, resulting in being viewed by more users, as it will appear on the subreddit's landing page.

In this analysis, we take users of *politics* and *The_Donald* as representative for a left- or right- leaning political orientations. We work on the comment data of reddit to find interesting similarities and deviations between both user groups in terms of language and behavioural patterns.

The data is openly available through an interface provided by reddit. However, we use an pre-packaged, archived dataset, uploaded by the user *Dewarim* on reddit itself [5], in order to reduce traffic, memory and time constraints. The whole available dataset ranges from 2005 up to 2017 and contains all comments, including meta information, by reddit. As the US presidential election, the election battle, debates, and the aftermath mainly took place from 2016 to early 2017, we restricted our analysis to those years in hope of high political activity across the platform. Additionally, older and more recent data distinguish due to API changes, requiring additional effort for synchronization. The restricted dataset contain around about 410 million comments in total, from which 41 million were issued in *The_Donald* and *politics*. The data itself is structured as a list of *JSON* objects, making it easy to process.

Analysis is sectioned into several fields, each focusing on other aspects. The first one will mainly focus on other subreddits aside from *politics* and *The_Donald*. Thereafter we study possible correlations between the user bases of the original subreddits and foreign ones. The second Section is about investigation of the comment body for analysing the language used in both subreddits. Thirdly we get into voting and replying behaviour of the user bases and analyse the language of comments user-wise. Fourthly, in the temporal analysis, we look for correlations between real life events and posting sessions. The fifth and last content part is about prediction of political orientation using information from prior analysis. After discussion of findings and methods, we conclude with summaries of the most important implications.

2 Related Work

This Section gives a brief overview of the related field of studies in social networks, reddit and political discourse. In the paper "Online Political Discourse in the Trump Era" Nithyanand et al. identified the factors that they assume must be the cause for the declining quality of political discourse in America [6]. In their paper "The Rise and Fall of Network Stars" Fire et al. tried to find the mechanism the emergence of new trends, using a graph representation of dynamic systems. They look at the change of the graph structure over time and identify which characteristics of the graph correlate to rise of a new trend [7]. The paper "Researching Social News" Mills analysed the front page of reddit, where the users decide what is popular and what is not, by analysing the ranking of popular news stories on the site [8].

3 Cross Subreddit Analysis

Reddit does not only consist of *The_Donald* and *politics* alone, but rather contains well over $100,000$ other subreddits [2]. This Section focuses on the behaviour of users from the two source boards in other subreddits, including the respective other. First, we check how popular *The_Donald* or *politics* are in general, proving the statistical relevance of our findings in Section 3.2. We continue by comparing the behaviour of the two users bases across the platform towards each other and the common reddit user, regarding their commenting frequency in Section 3.3. The question whether reddit in general is politically active, or maybe even biased, will then move into focus in Section 3.4, before we close out by revealing potential connections between the two source boards and other highly biased subreddits in Section 3.5.

3.1 Methods

Each user is assigned a group based on the boards he is posting to: He is called ***l**eft leaning*, if he posted to *politics*, but **not** in *The_Donald*, ***m**utual*, if he posted in both political source boards, ***r**ight leaning*, if he posted in *The_Donald*, but **not** in *politics* or ***n**eutral*, if he posted in neither of the two source boards. Additionally, if he posted in at least one of the two source boards, we refer to him as a *political* user. Given a comment in subreddit s, we then keep track of the following counts:

$$user\_count_{s,o} \text{ with } o \in \{l, r, m, n\}$$

describes the number of users that posted in s. The second index depends on the assigned orientation of the author. Similarly, $comment\_count_{s,o}$ describes the amount of comments that were found in s. For simplicity, we also define

$$total\_user\_count_s = \sum_{o \in \{l,r,m,n\}} user\_count_{s,o}$$

as the total number of users. By limiting our dataset to subreddits with more than 1500 users, we consider only boards with a certain statistical relevance.

3.2 General Questions

Comparing the *user_counts* of *The_Donald* and *politics* to other subreddits allows us to see how popular they are within reddit. We rank the most popular subreddits based on their *user_count* results in Figure 2.

Table 1 shows the same ranking using the *comment_count* instead.

Table 1: Most commented Subreddits:

Subreddit	Number of comments	Rank
AskReddit	40M	1
politics	23M	2
The_Donald	14M	3
leagueoflegends	10M	4
...	...	...

Results & Implications Figure 2 reveals that *politics*, with a user base of more than 570,000, ranks as the subreddit with the 15th most users across whole reddit, while *The_Donald* ranks as 26th most popular subreddit with over 358,000 unique users. When considering the amount of comments instead, Table 1 shows that *politics* now ranks as the subreddit with the second highest amount of comments, while *The_Donald* is the third most commented on board, with an amount of 23 and 14 million respectively.

Both *politics* and *The_Donald* rank among reddit's most popular subreddits, when it comes to the size of user bases or amount of comments, proving that they are very relevant to the daily discussion on the platform. The higher rank in number of comments suggests that the discussion in the forums seem to be very active.

3.3 Commenting Frequency

Figure 2 and Table 1 imply that users from *The_Donald* and *politics* post more often than neutral users. We calculate the mean of comments per user for all user groups over all subreddits of reddit in Table 2a. We want to find out, whether the prejudice that Trump supporters tend to spam more holds. Additionally, we want to check whether the commenting frequency of political users differs between their activity on the source boards and other subreddits. In Table 2b we only look into *politics* and *The_Donald* and calculate the comments per user for exclusive and mutual users.

(a) Comments per User

User Group	Mean	Standard Deviation
left-leaning	4.89	88.02
mutual	4.73	32.57
right-leaning	4.04	19.24
neutral	3.66	67.55

(b) Comments per User in Source Boards

	Subreddits	
User Origin	politics	The_Donald
exclusive	29.19	22.26
mutual	70.88	62.20

Results & Implications Table 2a highlights that exclusive users of *politics* post most frequent, while the mean for mutual users is close to it. Right-leaning users tend to post less frequent, followed by neutral users. The deviations are quite high for all user bases, with the left-leaning group being the most extreme one, showing that there must be a high variance in the values.

In Table 2b we find that users from *politics* and *The_Donald* do in fact post more often in their respective origin board compared to the mean. The value for left-leaning exclusive users is again higher than for right-leaning users and the counts for mutual users are significantly higher in both boards.

While Table 2a proved that exclusive users of *politics* post more often than their counterparts in *The_Donald*, they only differ by one comment per user on average. Thus, the hypothesis that right-leaning users spam more, does **not** hold. However, neutral reddit users indeed seem to post less frequent than political users. The extreme deviation for left-leaning users could be explained by very extreme users. Alternatively, bots could be responsible for the high deviation as well. In comparison, the distribution of comments per right-leaning user seems to be more stable. The fact that the means for mutual and left-leaning users are so close to each other indicates that both groups seem to share characteristics. The difference in deviations suggests that more very active users seem to isolate themselves in *politics*. However, this also strengthens the bot hypothesis, as active bots would most likely only comment in one previously specified board, thus be assigned as left-leaning.

In Table 2b it appears that political users indeed post more frequent in their source subreddits. However, this also correlates with the already active discussion space we showed in Section 3.2. More interestingly, the number of comments per mutual user is more than twice as high as for users that post exclusively in *politics* or *The_Donald*, indicating that mutual users seem to take a very active part in the discussions in both boards. Combined with their high share in the number of total users, shown in Figure 2, this leads to the fact that mutual users are responsible for approximately 69% of all comments in *The_Donald* and 48% in *politics*, indicating that the discussion in *politics* is more isolated. Either a lot of conservative users also post rarely in *politics* or otherwise a lot of liberal users also take part in the discussions in *The_Donald*.

In order to further investigate on their share of content in both boards, we compute every users posting ratio using following formula:

$$ratio = \frac{don\_comments_u}{don\_comments_u + pol\_comments_u} \in \ [0,1], \ \forall u \in \texttt{mutualUsers}$$

Where $x\_comments_u$ is the amount of comments on x by user u. Additionally we substract 0.5 from the *ratio* in order to center the ratio around 0. We can then analyse the distribution of these ratios and try to derive the effect of mutual users on both subreddits.

Figure 14 supports our finding that mutual users are responsible for a high share of comments on the two source boards, as it is mostly linear. When going further and filtering out users without at least 5 (Figure 15), 15 (Figure 16), and 25 (Figure 17) comments in total, the linear line shape changes to a rather S-shaped line, which polarizes the sides. This indicates that most of the users that are active on both boards, only account for few posts in total, and high active users are mostly active on their own subreddit, and only have a few posts in the respective other subreddit. This shows that political users tend to stay in their origin boards and post much less across the opposing political board.

3.4 Political Subreddits

Using the user bases of *politics* and *The_Donald*, we try to find other political subreddits by introducing the share of political users pol_share_s given subreddit s by

$$pol\_share_s = \frac{\sum_{o \in l,m,r} user\_count_{s,o}}{total\_user\_count_s}$$

Extracting a list of subreddit with the highest shares of political users leads us to Table 3.

Table 3: Political Subreddits

Subreddit	Political user share	Rank
The_Donald	1	1
politics	1	1
Mr_Trump	0.94	2
tucker_carlson	0.93	3
...	...	...
StillSandersForPres	0.89	8
TrumpForPrison	0.89	9
...	...	...

We also calculate the mean of pol_share_s over all of reddit to see if reddit itself is political.

Results & Implications The mean evaluates to approximately 23% political users on average over all subreddits. Table 3 is dominated by politically biased boards. Additionally, the first left-leaning subreddit only appears on position 8. It is hard to tell whether reddit itself is political, without taking the semantics of users' comments into account. However, it is clear that reddit seems to attract people with a certain interest in politics, as more than a fifth of every subreddit's user base also posts in one of the two source boards on average.

The dominance of politically biased boards in Table 3 could potentially be explained by *The_Donald* being better connected to other conservative boards across reddit. Neutral poltical boards are nearly missing entirely, with *PoliticalDiscussion* being the first one at rank 37. Taking users from two highly biased boards as source of political users is a possible explanation that we see so few neutral boards. It appears that these users tend to post specifically more likely in other biased subreddits. Another reason could be that subreddits like *news* also contain a lot of neutral users, further decreasing the share of political users.

3.5 Political Bias

Lastly, we try to find whether reddit in general tends towards a political orientation and which subreddits drastically drift towards the left or right. The first question can be answered by considering the share of users from *politics* in other subreddits, comparing it to the same value for *The_Donald*.

In order to find the subreddits with the most extreme user bases, regarding political bias, we introduce $bias_s \in [-1, 1]$ for a given subreddit s by:

$$bias_s = \frac{user\_count_{s,r} - user\_count_{s,l}}{\sum_{o \in \{l,m,r\}} user\_count_{s,o}}$$

If the share of users from *The_Donald* is high compared to the share of left-leaning users in subreddit s, $bias_s$ will tend towards 1 and to -1 for the opposite case. Table 4 provides a selected set of subreddits with high *bias* values.

Table 4: Biased Subreddits

Subreddit	bias
The_Donald	0.55
TheRightBoycott	0.32
the_frauke	0.31
The_Wilders	0.29
The_Farage	0.28
...	...
BernieSanders	-0.40
occupywallstreet	-0.41
HillaryForAmerica	-0.42
progressive	-0.46
BlueMidterm2018	-0.48
politics	-0.72

Results & Implications It turns out that, on average, 9% of each subreddit's user base also post in *politics*. In comparison, only 6% post exclusively in *The_Donald* and 9% in both. These values are not significant enough in their own to say that reddit as a whole is a left- or right-leaning platform, but they indicate a trend.

In Table 4, the absolute *bias*-values are higher for the left leaning subreddits, which might be due to the size advantage of *politics*. However, it is interesting to notice that subreddits with the highest positive score are about European leaders of modern right-wing parties, while on the other side of the spectrum we find typical liberal subreddits, mostly about U.S. related topics.

The focus on U.S related topics is due to the fact that *politics* only deals with news regarding the U.S. itself. Finding mostly progressive topics on the list also further strengthens our ground truth that *politics* can be considered liberal. All in all the *bias*-investigations arguably indicate that the right-wing movement is better connected internationally than its opposition. It would be interesting to see if we would find different results if we also considered general liberal subreddits like *progressive* as a source board.

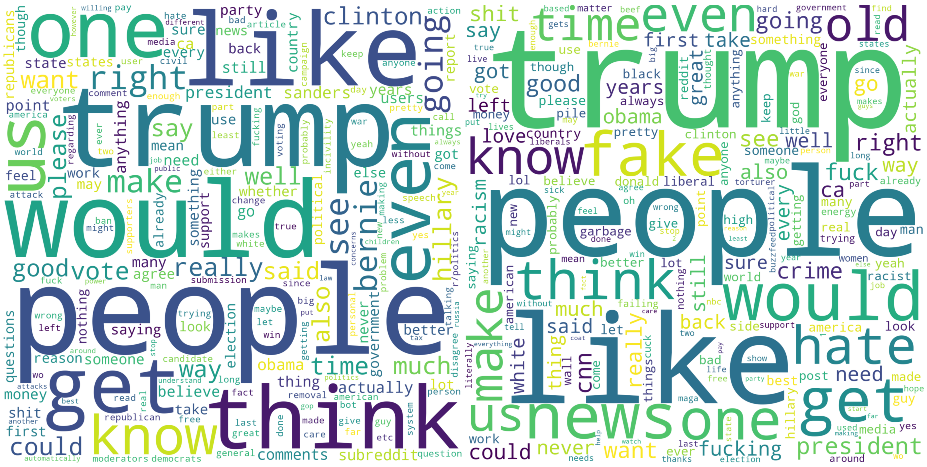
4 Language Analysis

In this part we want to retrieve information from the used language in the comments. To this regard we check **lexemes** in Section 4.1 and **semantic** in Section 4.2 of the given comments with the following *analyses*: The first is **lexemes** by usage of *words* which includes *website links*(and attributed meaning), *frequency of words* to identify `key words` for group identity. The second **semantic**, simplified as *sentimental*, which returns `polarity`(affirmation or negation) and `subjectivity`(as contrast to neutrality of statement from a persons viewpoint).

4.1 Lexemes

Word analysis As wordings and phrases are used to define group identities [9, p.65], we want to identify those. *Remark* that this may also be context or semantic dependent, but analyzing overall characteristics is a first step into that direction.

Results In Figure 1 the word cloud as frequency of most common words relative to another is plotted using word_cloud [10]. Frequently used English words (called *stop words*) and often used abbreviations like *it's* are hereby not considered.



(a) politics

(b) The_Donald

Fig. 1: Word map of subreddits

Relative common phrases are `trump`, `like`, `think`, `people`. Comparing left to right, more frequently used words by *politics* in Figure 1a are `please`, `going`, `hillary`, `bernie`, `really`. However there are not many overall identification keywords with notable frequency difference in use. Much more frequent words in *The_Donald*, shown in Figure 1b compared to *politics* are `fake`, `news`, `hate`, and hostile words. Notable mention here is `cnn` and other usually negative connoted words like `garbage` and `old`.

Table 5: comment news website bias on reddit

(a) politics

	websites	share	score
1	washingtonpost	9.42%	-1
2	nytimes	8.45%	-1
3	politico	5.58%	-1
4	cnn	4.82%	-1
5	politifact	4.60%	-1
6	realclearpolitics	3.64%	0
7	theguardian	3.07%	-1
8	fivethirtyeight	2.95%	0
9	thehill	2.79%	0
10	huffingtonpost	2.70%	-1
⋮	⋮	⋮	
17	breitbart	1.41%	1
⋮	⋮	⋮	
25	foxnews	1.19%	1

(b) The_Donald

	websites	share	score
1	breitbart	13.94%	1
2	dailymail.co.uk	7.57%	1
3	nytimes	7.45%	-1
4	washingtonpost	7.32%	-1
5	foxnews	7.02%	1
6	cnn	5.94%	-1
7	theguardian	5.51%	-1
8	politico	5.03%	-1
9	dailycaller	4.27%	1
10	thehill	3.06%	0
⋮	⋮	⋮	⋮
13	huffingtonpost	2.25%	-1
⋮	⋮	⋮	⋮
17	nbcnews	2.03%	0

Links analysis News websites, as advertisement platforms, target certain groups, often with certain political bias. For English speaking news websites we found 2 bigger collection websites that do justified categorization of these:
mediabiasfactchec(`mbfc`) [11] and allsides(`allsides`) [12]. `mbfc` provides a bias categorization with weighting of different aspects. Therefore they first choose the left or right leaning and then use different categories for which they use the mean of the weights in $\{0, 1, \ldots, 10\}$. However they do not provide the underlying decision reasons for the weights (meaning analysed articles). Additionally they claim to do fact checking, but do not provide exact methodology on what base. More accurately they do not explain how they can recognize wrong facts (and which wrong facts were found) and being vague about the time scale they are updated. User feedback is not provided.

`allsides` however provides user feedback and includes that point for a confidence level in the decision, but does not do fact checking. The granularity of bias is with 5 steps $\{-2, -1, \ldots, 2\}$ from left to right a more broad description.

Since we do not concentrate on fact checking and want to have simple results, we use `allsides`. Especially to correct the websites and our own bias, we want to use feedback which is not provided in `mbfc`.

Results Extracting the first mentioned website of each comment and assigning according types and bias $\{-1, 0, 1\}$ for left, neutral and right, we obtain the following results for the data set as shown in Table 5. Having in *politics* 57465 news website links and in *The_Donald* 7812, we see in Table 5a roughly 2.5% red-leaning website, as might be expected from the left-leaning subreddit. Comparing to Table 5a, Table 5b shows only 5% neutral websites and all the most relevant websites from *politics*.

Implications The common phrases are with the controversial character of Trump not very surprising, but reflect the behavior of the debate and the election outcome. The language phrasing is typical for left and right leaning opinions: The left tends to more inclusive wording, whereas the right uses more frequently hostile language as already analysed in [6]. More interestingly are the common used phrases `fake`, `news` for where a further analysis should be very interesting and which news sources on what context were specifically aimed. To this regard the keyword `cnn` and the used strategy of repeatedly accusations against the political enemy (including the media) and effect analysis may be of interest.

Linking on most relevant left-leaning websites does not indicate, as one might expect, left bias of a user. However frequent linking on right-leaning websites or neutral websites may give the political orientation more accurately. Several hypothesis for this behavior are made in Section 8.2.

4.2 Semantics analysis

> "Semantic change asks how words change meaning over time, and questions both the processes involved and the causes." [9, p.2]

Thus for a certain time point the assigned meaning of a word might be wrong and analysis must somehow reflect this process. Additionally adding to this hardness is the use in human subgroups having additionally or different meaning [9, p.].

Possible use of irony, per definition using unexpected situations (or here meaning of words) and sometimes even for humans hard to understand, adds to this problem.

The underlying idea of minimizing ironic statements is, that rarely people use highly subjective jokes, i.e. express subjective emotions within a joke. Irony works best, when one lets the audience identify with the situation, so the reader can draw conclusions later introduced as faulty.

For semantic analysis of the comments we choose the language processing library *TextBlob* [13]. Processing the polarity sentimental score for each comment, we inspect the sentimental score distribution and combine it with chosen keywords.

Results In Table 6 characteristics of the sentimental score of each post are shown. Although there are much more comments posted in *politics*, the mean and standard deviation are lower. A way to explain this is, that most comments behave more neutral rather than semantically extreme ones. The huge difference lies in the first quartile value. In *The_Donald*, more than a quarter of all posts are negative, while in *politics* the proportion of negative comments is less than a quarter. Additionally, the third quartile value in *The_Donald* is higher.

4.3 Comment Keywords Comparison

Additionally, we can analyse certain keywords and show whether those keywords might influence the sentimental score distribution.

Table 6: User Post Behaviour Statistics

	The_Donald	politics
Subjectivity > 0.6		
count	3475323	5578708
mean	0.07	0.04
std	0.47	0.40
25%	-0.25	0.2
75%	0.40	0.32

Results Before plotting the keywords sentimental score distribution, we shall utilize a list of most frequently appeared words from language analysis stated previously. When choosing some sensitive keywords, like *Trump* or *Hillary*, we can obtain the corresponding results, which can be found in Figure 5 and Figure 6. For the keyword *Trump*, we can observe the differences that those comments in *The_Donald* is more positive and less negative than the other subreddits, while the keywords *Hillary* is slightly more negative than the others.

5 User Behaviour Analysis

In this chapter, we focus on analyzing user's behaviour to figure out whether there are very active users and how they behave in different subreddits.

We take the following steps for preprocessing:

1. Identify mutual users and the subreddits users from *The_Donald* and *politics*,
2. extract all related comments for the three user types (*mutual*, *The_Donald*, *politics*),
3. Extract *author*, *score*, *body*, *link_id*, *created_utc* of these comments,
4. Handle special characters or escape comments and rewrite data to csv.

5.1 Power User Analysis

In order to be able to influence other's political orientation, one has to be very active in the subreddit by frequent and attentional comments (being either aggressive or inspiring), which is so called power user.Since reddit has a voting mechanism, it is assumed a high score reflects high attention of the community. Usually one would need information from whom these comments are and in what kind of debate they occur, but we do not have these information and simplify the situation.

We define **two** criteria for the importance of a power user:

1. Net score of the user's comments,
2. Number of times that the user's comment is linked by others.

The second criteria hereby presents the weight of the user's comment. For extraction of power users, we traverse the data and sort the users by the total scores of all comments. Further we assume that the author, who has the smallest *created_utc* under the specific *link_id*, is the comment submitter. Based on this idea, we additionally get the users, whose comment has the highest references.

Results & Implications Following Table shows the upvote behaviour in both subreddits: As shown from Table 7, there are more users in *politics*, almost twice the number of users in *The_Donald*. Correspondingly, the sum of the total upvotes is larger in *politics*. However, the average upvote for each user in *The_Donald* is higher, which implies that users in *The_Donald* interact with each other more frequently. The standard deviation of upvotes in *The_Donald* is also larger than in *politics*, which indicates that the users' upvote distribution in *The_Donald* is more sparse. Similarly, there are about 6.3% of users in *politics* and about 8% users in *The_Donald* who contributes 80% of total upvotes, which means about 6-8% users dominating the commenting behaviour in each subreddit. This suggests that there are indeed some power users with high activity. The higher mean in *The_Donald* suggests that discussions are more interactive in *politics*.

Table 7: Users' Upvote Behaviour Analysis

Measurements	politics	The_Donald
Count of users	570,349	358,165
Sum of upvote	29,206,680	23,966,048
Mean of upvote	116.29	164.40
Std of upvote	696.03	776.84
% of users contribute 80 % upvotes	6.26 %	7.97 %

5.2 User Comment Behavior

In this Section we analyse the sentimental score distribution for users. There are some users who are active in both subreddits. Therefore we inspect their commenting behaviour in other subreddits, regarding subjectivity and sentiment.

Results & Implications After filtering out those too objective comments with the subjectivity cut value 0.4, we gather all the comments sentimental score in each subreddit and plot them as shown in Figure 3. In order get a better comparison for the user commenting behaviour, we choose two other popular subreddits: *AskReddit* and *news*, which also include a large portion of the political subreddits mutual users. The x-axis represents the sentimental scores, ranging from -1 to 1, and the y-axis shows the total number of comments. Similarly, we derive another plot with the subjectivity cut value 0.6, shown below as Figure 4. In Figure 3, those lines are relatively smooth on the left part, however, there is a huge bump on Figure 4. There is also a higher peak value in Figure 4 on the positive region, than in Figure 3. This suggests that with the increase in subjectivity value, the comments become less neutral, having the tendency to become more negative or more positive. However, in both plots the red line shows relatively extreme commenting behaviour, indicating users in *The_Donald* tends to post more offensive or more positive comments. The different commenting behaviour is presented from the statistical overview in Table 8. First, we observe that users with subjectivity above the threshold post most comments in *The_Donald*. Therefore the user comment sentimental score in *The_Donald* also has the largest standard deviation, and inclines to have more negative and more positive scores regarding the first quartile value and third quartile value respectively. With the increase of the subjectivity value, and by filtering out more objective comments, we observe that the standard deviation becomes larger and first quartile value and third quartile becomes more negative and more positive respectively.

Table 8: User Commenting Behaviour Statistics Comparison

	The_donald	politics	AskReddit	news
Subjectivity > 0.4				
count	63411	60572	31759	18864
mean	0.081	0.076	0.083	0.051
std	0.387	0.329	0.352	0.313
25%	-0.164	-0.100	-0.118	-0.122
75%	0.344	0.274	0.300	0.250
Subjectivity > 0.6				
count	32050	27121	15143	7928
mean	0.052	0.046	0.053	0.010
std	0.462	0.406	0.430	0.387
25%	-0.267	-0.212	-0.225	-0.225
75%	0.390	0.333	0.350	0.262

5.3 Summary

Through user behaviour analysis, we can observe a list of power users in each subreddit and find out that about 6-8% of users dominating the comments upvote behaviour. Those users in *The_Donald* tend to post more aggressive comments. Besides, we notice that those mutual authors have varied posting behaviour in diverse subreddits, in addition, there are differences in keywords sentimental behaviour cross various subreddits.

6 Time Analysis

Additionally we analyse whether there are significant differences between temporal behaviour of different political orientations. To this end, we analyse temporal behaviour of different users. We use the field `created_utc` (Coordinated Universal Time) as timestamp.

First, we analyse the general activity on both boards and observe whether we can identify differences.

We also have a look into a user session, which are the posts of a user which happen with no more than 1 hour difference, to see if there are any specific orders in which the user decides to post. In order to do this, we collect the timestamps, subreddit names of a user and give them a session identifier i.e. a number starting at 1. If the next post by this user is within a period of one hour, it belongs to the same session. Otherwise it is the start of a new session.

Furthermore we analyse whether peaks in overall activity on the respective subreddits are correlated to certain events that occured in a similar time frame. For that we analyse the overall activity of the boards over the months and identify peaks and lookup events that happened roughly in the same timespan in the news. We continue to see, if there are patterns between certain reoccurring events and peaks in activity.

6.1 General Activity

In the first step to this, we extract all users which are active on *politics* and *The_Donald* and save them to a file, for easier access. Then we extract all timestamps (`created_utc`) related to these users and both subreddits, and resample them into *hours*, *days*, *weeks*, and *months*. The remainder of the anomaly analysis focuses on the daily sampling, because it is not too coarse but also not to detailed, such that its possible to have a good overview over differences. For the general activity we look at different kinds of activity: Amount of comments posted: by users, in a subreddit, per hour, per day, per week, and per month.

Table 9: General Activity: Across all subreddits, complete time-frame

Comments	Min	Max	Mean	Std Dev
by author	1	73843110	532.805	84653.19
in subreddit	1	27705040	1861.384	98806.78
by hour	267	97447	37521.76	14070.59
by day	606819	1578528	900522.18	110459.64
by week	2796798	7836740	6221789.58	706254.13
by month	961460	31018069	25664882.00	6687706.49

Results & Implications In Table 9 one can see the min, max, mean, and standard deviation of each of the categories. Notable results are e.g. the maximum value for comments by author. This high value is given due to the fact, that deleted posts get marked "[deleted]" in the author-field, such that this value represents the amount of deleted posts. The mean indicates that most users have posted fewer comments in the observed time. Its noteworthy that the mean values for hour/day/week/month relate to each other roughly like hour/-day/week/month relate to each other. The mean amount of posts per day are around 24-times the amount per hour, the mean of posts per week is around

7-times the amount per day, and the mean of posts per month is roughly 4-times the amount per week.

This suggests that overall the activity on the website does not behave significantly abnormal. As visible in Figure 7, 8, and 9 that there is a increasing trend, especially for *The_Donald*, where there was almost no activity in the first month. Future work includes more recent content. Also additionally to the comments, analysing submissions is of further interest.

6.2 User Sessions

In order to compute the user sessions to perform a comparison, we extract the political users again, and remove mutual users from both sets, such that we only have exclusively the users from each subreddit, who do not post on the respective other board. For each of these sets, we collect all their activity on the whole website and subdivide it into sessions. A session consists of all consecutive posts which happen with no greater time difference than 3600 seconds. For every of these sessions, for every user we then compute the total amount of sessions of that particular user, the amount of comments for each session, and the duration in seconds. We remove all sessions with only one comment, because this indicates that the user has not been active more than once within 1 hours. The respective duration of that session, which is zero, is also not included in the statistics.

Results & Implications For the user sessions we compute the mean and the standard deviation of the amount of sessions per user, comments per session and duration of the session for exclusive *politics* and *The_Donald* users in Table 10. Its noteworthy, that the data available in Table 10 for *politics* is inherently larger than *The_Donald*, which makes it hard to compare them. Nevertheless its visible that the standard deviation is much higher relative to the mean for *The_Donald*. Its around four times more compared to *politics*, where its only up to around two times as high.

Table 10: Session Statistics: *politics* vs. *The_Donald*

	Mean	Std Dev	Mean	Std Dev
Sessions per User	192.44	280.09	9.787	40.659
Comments per Session	81.850	154.42	9.850	35.919
Duration of Session (s)	81.837	154.40	9.854	35.930

This could be an indication to that exclusive *The_Donald* users, are less active than their counterparts. Its interesting to find, that the duration of a session in seconds is very close to the amount of comments per session, which suggests that on average comments are rather short replies which could be done within a second. Further work includes testing several session duration ranges, including information about human attention span.

6.3 Activity Anomaly

To identify anomalies in commenting activity, we manually look for peaks in the activity representation mentioned above. We decide for a cut-off at 100K comments per day, which is high enough to eliminate noise, but also low enough to identify multiple peaks which are high enough. For each of these peaks, we look up the dates at the website `archive.org` with the search term "US". This returns multiple results, some entries of the website have higher amounts of views compared to others. Using this information we try to find political events happening within a one-day-tolerance of the peaks.

Results & Implications When analysing the daily subreddit activity we find notable peaks, which reach far over 100K comments per day as seen in Figure 8. A closer look at these specific peaks reveals that they are always tied to a major political event in the US, which are the Democratic National Convention (DNC) in Figure 10, the three dates of presidential debate in Figure 11, the night of the presidential election in Figure 12, and the inauguration of Donald Trump as the president of the US (and other, rather minor events) in Figure 13.

The most likely cause for these peaks is a increase in overall activity when certain events happen. In those events users do voice their opinion or provide further information they might have gathered. This analysis can be extended to compare, if for all subreddits relevant events cause increase activity. Or, whether this is a purely political phenomenon.

7 Prediction of Political Orientation

In this Section we try to predict the political orientation using the most frequently used words on both subreddits as a feature. For a user we want to predict his orientation, based on one of the two main subreddits, *politics* and *The_Donald*. In order to prepare for this, we extract the whole text-body of both political subreddits and identify the top 100 most used words. We then choose those 100 words as our features, such that every comment-body is represented by a vector $v \in \mathbb{R}^{100}$, where each entry is the amount of times that specific feature-word is used in the comment-body. If the comment is posted in *politics* we assign it the label $+1$ and if it is commented in *The_Donald* we assign -1. We repeat this for every comment in both boards. Afterwards we select the first half of the comment body (due to limitations in RAM) and split the set 70/30 into training and testing. For the training we perform Logistic-Regression [14] with 10-fold cross validation with a 70/30 training/validation split and 10 different values for the hyper-parameters, which results in 100 runs for fitting. Afterwards we use the remaining test-set, which was not seen by the model before and compute the accuracy of the model.

Results & Implications Lastly this part contains the results of the prediction of political orientation using most frequently used words. The total amount of all positive labels is 25535696, the total amount of negative labels is 16108468 (1.6:1 ratio). Before attempting any sophisticated prediction, we try simple Logistic-Regression [14], and achieve a precision of 96%. This was most likely due to overfitting to the data. Therefore we apply cross-fold-validation together with different hyper-parameter values, after which the accuracy of the prediction drops to about 90%. This might suggest that the models generalization improves, and is less likely to overfit. The advantage of Logistic-Regression is that each of the features is assigned a weight, such that we evaluate the importance of each of feature-words in this context. Figure 18 shows the plot of each of the words including their respective weight. The red arrows indicate words of interest, which are (from left to right): 'She', 'Clinton', 'Vote', 'I', 'Trump', 'Hillary', 'I'm', 'He'. This suggests that some words are more likely to be used on one side of the political spectrum, than on the other.

To further improve these results it is viable to put more effort into choosing the feature words, e.g. only consider nouns and use more advanced methods for prediction, e.g. neuronal networks.

8 Discussion & Limitations

This Section discusses the results of our analysis and present the limitations thereof.

8.1 Cross Subreddit Analysis

A problem in the Cross Subreddit Analysis is the group of mutual users. It's hard to tell whether *politics* or *The_Donald* are more or less isolated. If we knew whether the average user of the mutual group leans towards one of the two political orientations, we could better explain which of the both groups' users post more in the respective other subreddit and which user base prefers to stay isolated. In addition, it would be interesting to see what kind of comments the user groups post in other subreddits. One could then see whether they are influencing reddit on a political basis or whether they stay neutral in other subreddits.

8.2 Language Analysis

Sanitizing content took a lot of work, which has not yet be done by other groups for reddit. This strained results presented in here. The phrased keywords would be interesting to analyse over several time in their change and with regard to combinations of words they are used. The political narrative of Trump's presidency is reflected in the used words, whereas it is vaguely for *politics*. Differences in Obama's presidency and in other campaigns and their success with computing influence of wordings to success would be of interest.

The link analysis leaves certain hypotheses to analyse for *The_Donald*: 1. Right-leaning commenters on reddit are more interested in ideology and less in facts, 2. Right-leaning posts could be of higher quality and good sources, 3. Homogeneity of user base for right-leaning commenters is higher with regard to the topics, 4. *The_Donald* commenters are more inhomogeneous (”imposters”, ”angry folks”,. . .), 5. More consumer-oriented posts with fewer debate intention and linking. For *politics* the opposite hypotheses may be true.

8.3 User Behavior Analysis

There was a problem when analyzing comment semantics because there are plenty of available libraries, however, we have no idea about identifying which library is the most reliable. Since we only adopt one popular text processing library to conduct all semantic analysis, more experiments are still remained to be done by alternatives. Besides, though observed evidence that users behave differently in various subreddits, it is still an open question to identify how a user gradually change their posting behaviour, e.g. posting frequency, comment length, comment semantics, and etc. There do exist better approaches like for identification of irony [15], but we wanted to keep things simple.

8.4 Temporal Analysis

When attempting to analyse temporal behaviour an issue that occurs are bots, which are highly active and thus should be excluded from analysis. The large difference in user-base for the analysed time also poses an issue, when trying to compare activity on both political spectrums, as one side is not represented strongly enough. Also there are certain rules that apply on each of the subreddits which might influence certain aspects of activity or use of words. This might cause some artificial differences/similarities, which do not exists naturally. Further work includes more data and maybe analysis of different social networks besides reddit.

8.5 Prediction of Political Orientation

Using the most frequent words to predict the political orientation poses the issue, that certain commonly used words need to be filtered out, e.g. “and”, “or”, “to”, “in”, “the”. This will likely provide better results, as these words are frequently used across all political orientations. Another issue is the ground truth selection: There is a lot of content from mutual users, shown in Section 3, which should be filtered out in order to get a higher contrast between the political groups. Also the number of text-bodies available from *The_Donald* is 60% of the number of text-bodies of *politics*, which could cause a bias as simply always guessing +1 gives an accuracy of 60%. Other future work includes adjusting the training data such that the amount of positive and negative examples would be the same.

9 Conclusion

This Section briefly summarizes the implications of the other Sections.

politics and *The_Donald* users are more active than average reddit users and on average *politics* users are even more active than *The_Donald* users. However the standard deviation of comments in *politics* compared to *The_Donald* is very high and thus *The_Donald* users tend to post far more consistent. High active users on *politics* and *The_Donald* have only few comments in the respective other board. Thus the discussion in both boards is very isolated from another. The top 7 subreddits by amount of political users(being *politics* or *The_Donald*) aside from *politics* are conservative. This may indicate better international connection by many European users or concern about European politics. Slightly more users seem to be left-leaning on reddit. The left- and right-leaning of the boards reflects also in the amount of users being in other boards by topic.

The controversial character of Trump is reflected in the frequency of his mentions in both boards. Trump's campaign narrative `fake, news`, as being anti-establishment, is reflected in the wording, whereas in *politics* no such keywords occur in that frequency. For reddit linking on left websites indicates no political bias, whereas linking to neutral news websites indicates being more liberal. Linking to right websites indicates being more conservative. Further investigation is needed to derive more information.

Discussions in *The_Donald* have on average more upvotes. Only 6-8% in *The_Donald* and *politics* users are responsible for 80% of total upvotes. This indicates a small share of highly active users. *The_Donald* users tend to post more polarized comments, meaning to be more affirmative or negative, on higher subjective comments. The keywords *Trump* is more uniformly positive to the positive end, whereas *Hillary* has more only slightly affirmative comments for high subjective comments. Therefore *Trump* is more supported in *The_Donald* than *Hillary* in *politics* on reddit.

The overall activity on the website *reddit* did not behave abnormal in the chosen time frame of the data set. Most comments are rather short replies, which indicates creation time of few seconds. The temporal analysis showed that there is correlation between activity and political events. Further work is needed in order to derive more definitive information from user sessions.

Regarding political orientation, our results suggest that its possible to predict political orientation using frequently used words. More work can be put into sophisticated classifiers and improved feature selection.

References

1. A. I. Inc., "reddit.com traffic statistics," accessed: 2018-07-23. [Online]. Available: https://www.alexa.com/siteinfo/reddit.com
2. R. Inc., "The conversation starts here," accessed: 2018-07-23. [Online]. Available: https://www.redditinc.com/
3. Users of reddit.com, "The_Donald," accessed: 2018-07-23. [Online]. Available: https://www.reddit.com/r/The_Donald/

4. ——, “politics,” accessed: 2018-07-23. [Online]. Available: https://www.reddit.com/r/politics/
5. u/Dewarim on reddit.com, “Updated reddit comment dataset as torrents,” 2017. [Online]. Available: https://www.reddit.com/r/datasets/comments/65o7py/updated_reddit_comment_dataset_as_torrents
6. R. Nithyanand, B. Schaffner, and P. Gill, “Online political discourse in the trump era,” *arXiv preprint arXiv:1711.05303*, 2017.
7. M. Fire and C. Guestrin, “The rise and fall of network stars: Analyzing 2.5 million graphs to reveal how high-degree vertices emerge over time,” *arXiv preprint arXiv:1706.06690*, 2017.
8. R. Mills, “Researching social news – is reddit.com a mouthpiece for the ‘hive mind’, or a collective intelligence approach to information overload?” in *ETHICOMP 2011 Conference Proceedings: The Social Impact of Social Computing*, S. Rogerson, A. Bissett, T. Bynum, A. Light, and A. Lauener, Eds., 09 2011, pp. 300 – 310.
9. A. Boussidan, “Dynamics of semantic change : Detecting, analyzing and modeling semantic change in corpus in short diachrony,” Ph.D. dissertation, Universite Lumiere - Lyon 2, 2013, thèse de doctorat dirigée par Ploux, Sabine Sciences cognitives Lyon 2 2013. [Online]. Available: http://www.theses.fr/2013LYO20039
10. A. Mueller, J.-C. Fillion-Robin, and F. Tian. (2018) word_cloud. [Online]. Available: https://github.com/amueller/word_cloud
11. M. team & community. (2018) Media bias/fact check. [Online]. Available: https://mediabiasfactcheck.com/
12. A. team & community. (2018) Media bias ratings. [Online]. Available: https://www.allsides.com/media-bias/media-bias-ratings
13. S. Loria and community. (2018) Textblob: Simplified text processing. [Online]. Available: https://github.com/sloria/TextBlob
14. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
15. J. M. Chenlo and D. E. Losada, “A machine learning approach for subjectivity classification based on positional and discourse features,” in *Multidisciplinary Information Retrieval*, M. Lupu, E. Kanoulas, and F. Loizides, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 17–28.
16. P. T. Inc. (2015) Collaborative data science. Montreal, QC. [Online]. Available: https://plot.ly
17. W. McKinney, “Data structures for statistical computing in python,” in *Proceedings of the 9th Python in Science Conference*, S. van der Walt and J. Millman, Eds., 2010, pp. 51 – 56.
18. J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin, “Learning subjective language,” *Comput. Linguist.*, vol. 30, no. 3, pp. 277–308, Sep. 2004. [Online]. Available: http://dx.doi.org/10.1162/0891201041850885

Appendix

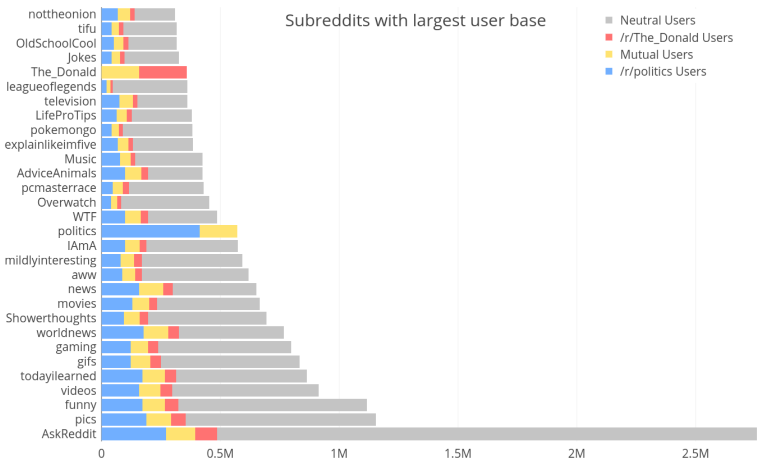


Fig. 2: Top 30 subreddits with the highes number of users

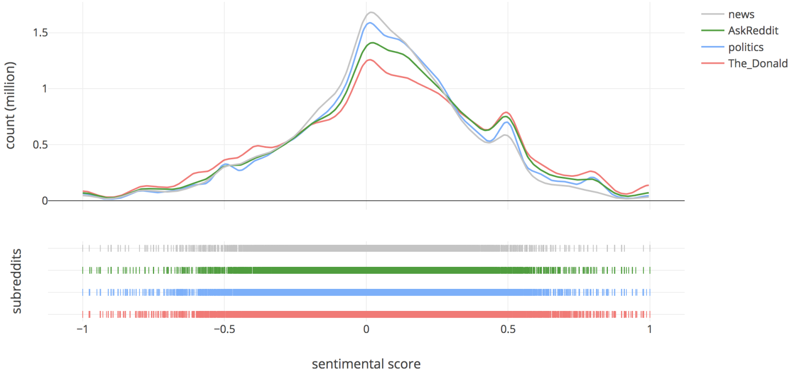


Fig. 3: Comment Sentimental Behaviour Comparison (subjectivity > 0.4)

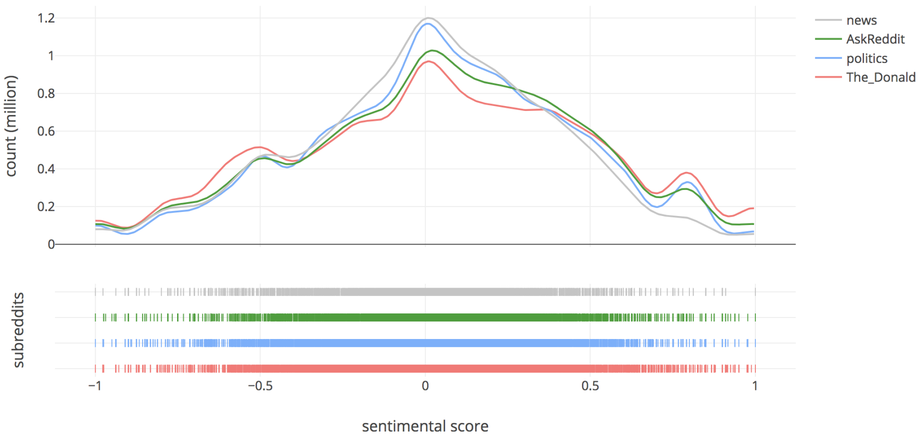


Fig. 4: Comment Sentimental Behaviour Comparison (subjectivity > 0.6)

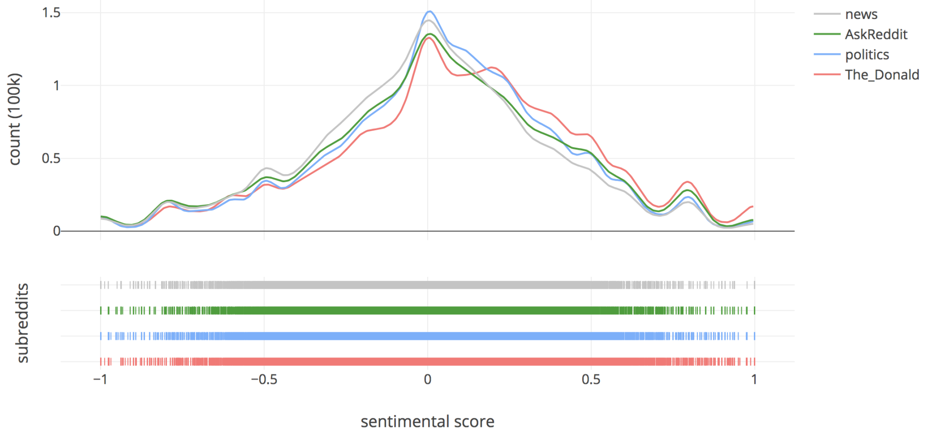


Fig. 5: Keywords–*trump* Sentimental Behaviour Comparison

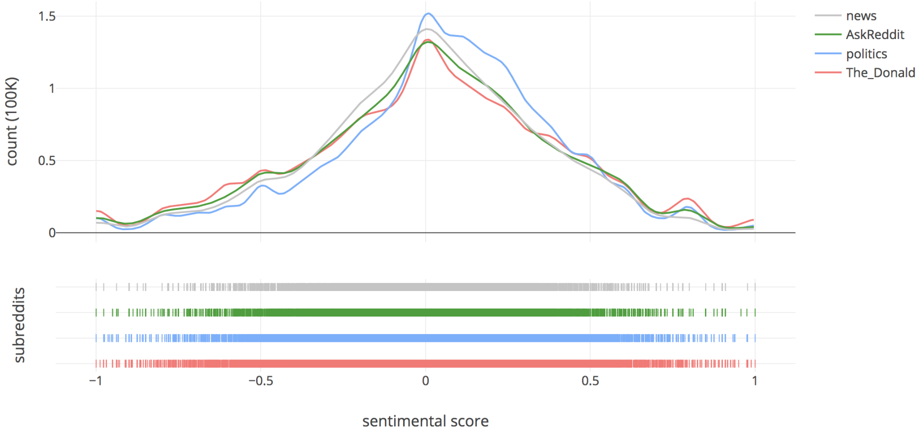


Fig. 6: Keywords–*hillary* Sentimental Behaviour Comparison

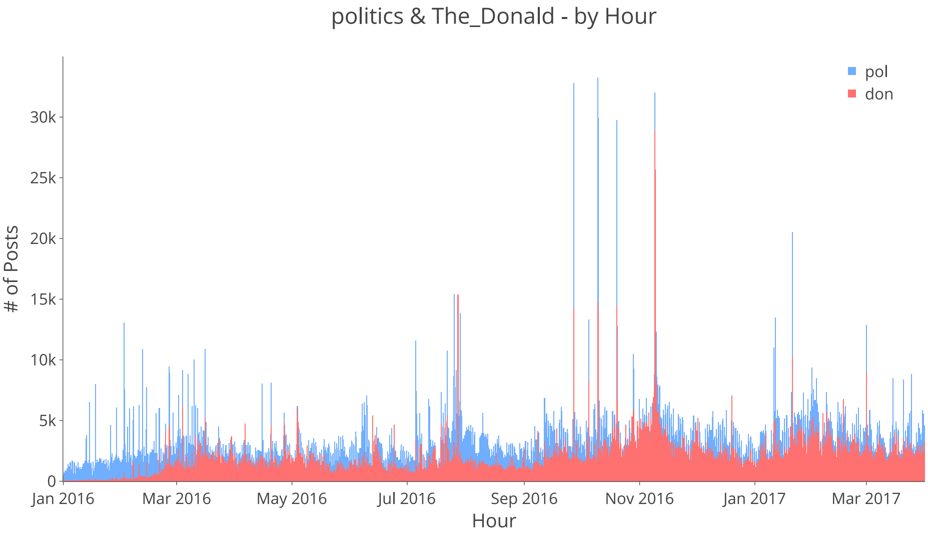


Fig. 7: Overview of the data in resolution of hours.

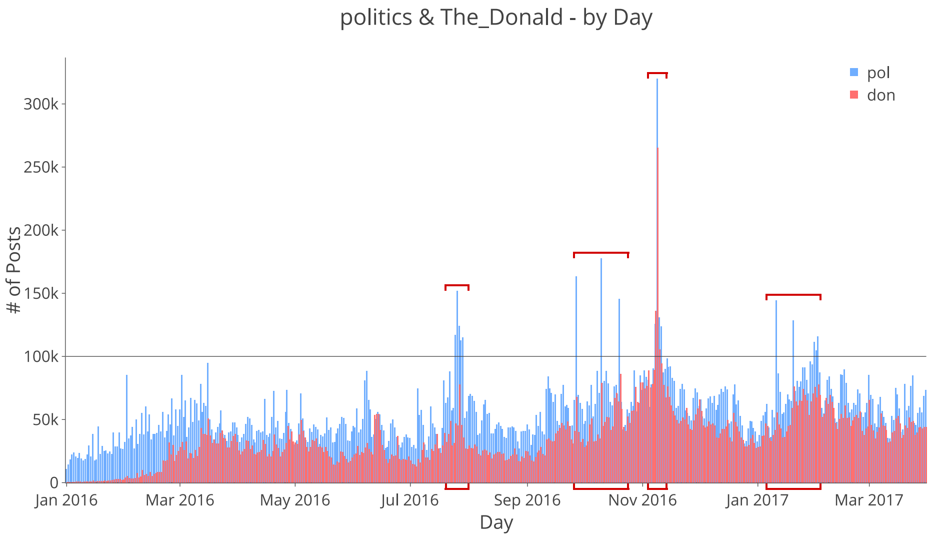


Fig. 8: Overview of the data in resolution of days. The horizontal line marks the 100K-comments-per-day-mark, above which the points of interest are marked using a red bracket. Those are analysed in more detail.

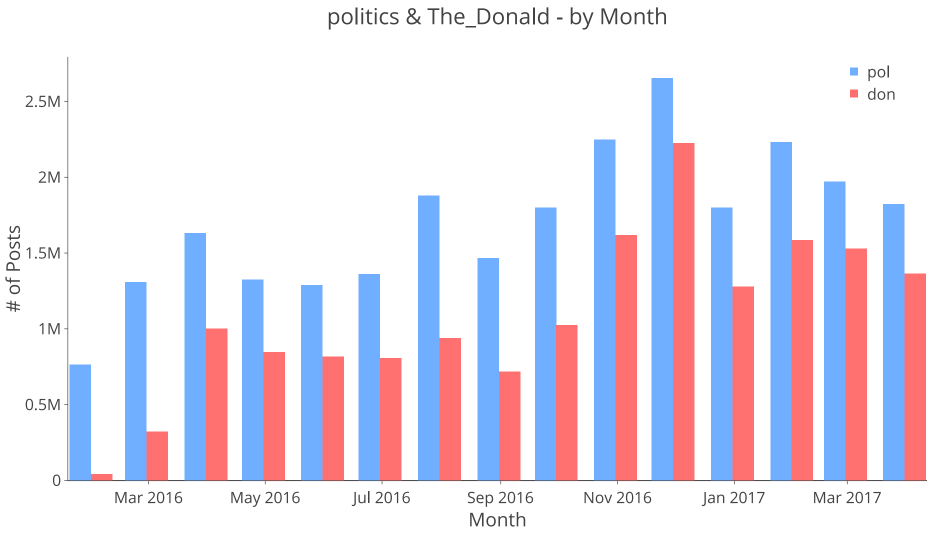


Fig. 9: Overview of the data in resolution of months.

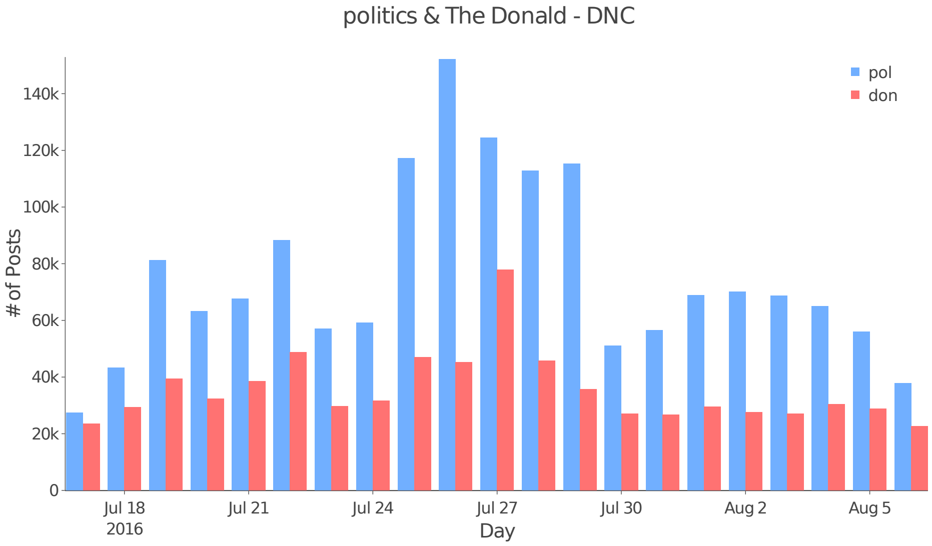


Fig. 10: Zoom in on the first peak. A lookup revealed, that the DNC happened at that date.

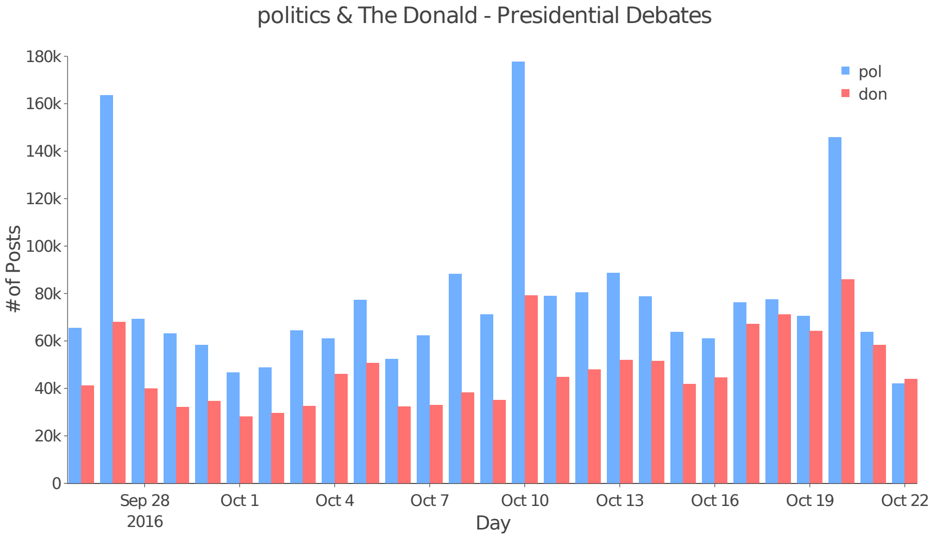


Fig. 11: Zoom in on the next three peaks. It turned out to be each date of each presidential debate in the autumn of 2016.

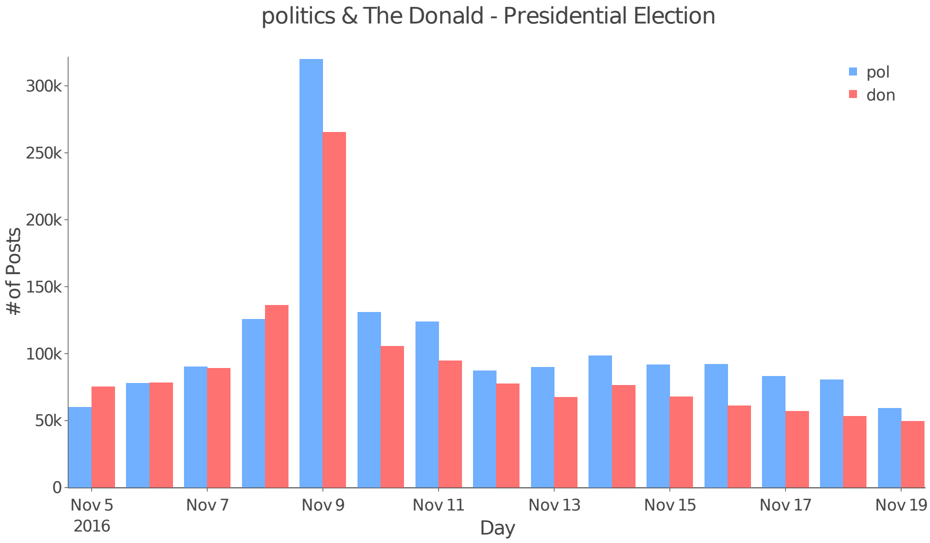


Fig. 12: By far the largest peak of activity is caused by the election victory of Donald Trump. Notice the height of the peak, which is almost twice as high as every peak so far.

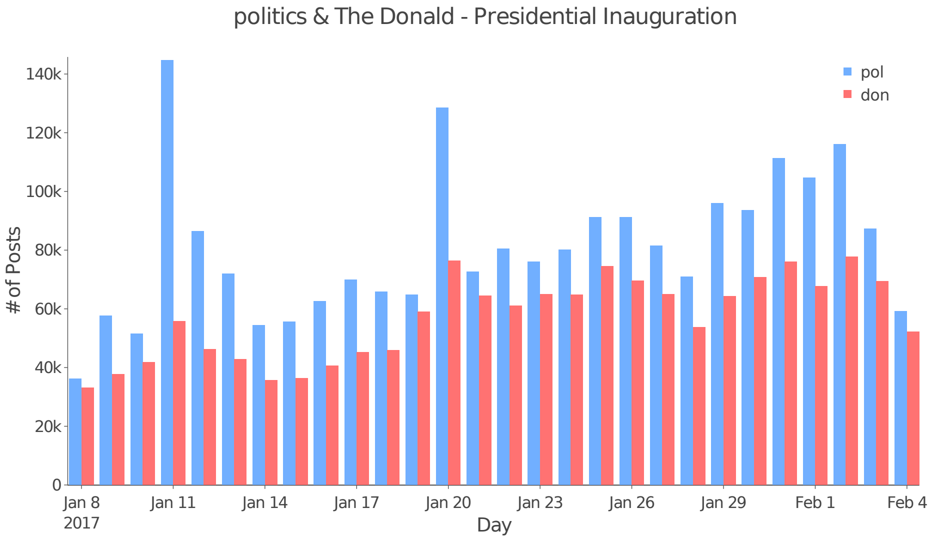


Fig. 13: The last three peaks in the observed data: Trump accuses Russia of the DNC hack; Trump is inaugurated as president of the US; somewhat unclear: Trump fires his attorney general and reports about military strikes in Yemen arise.

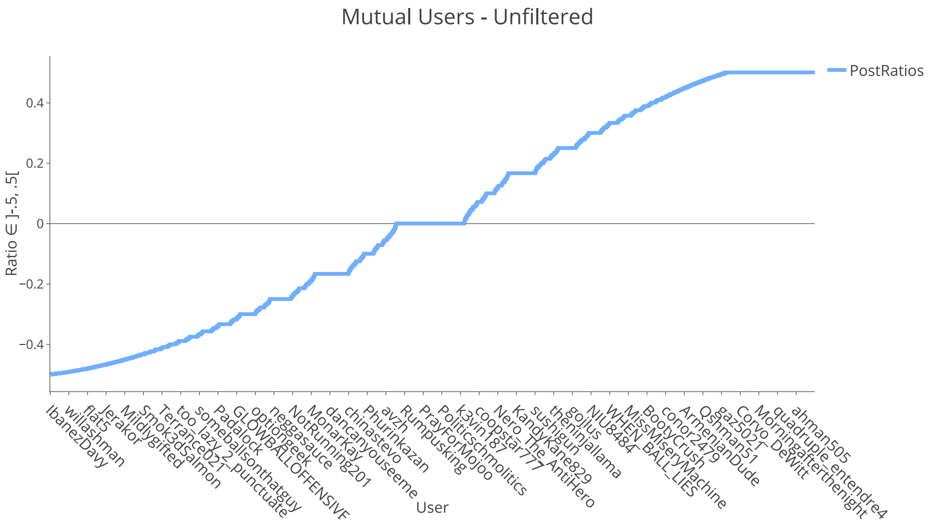


Fig. 14: Mutual Users commenting ratio, without filtering users. Appears almost linear. Many points lie at 0. 112349 Users in total.

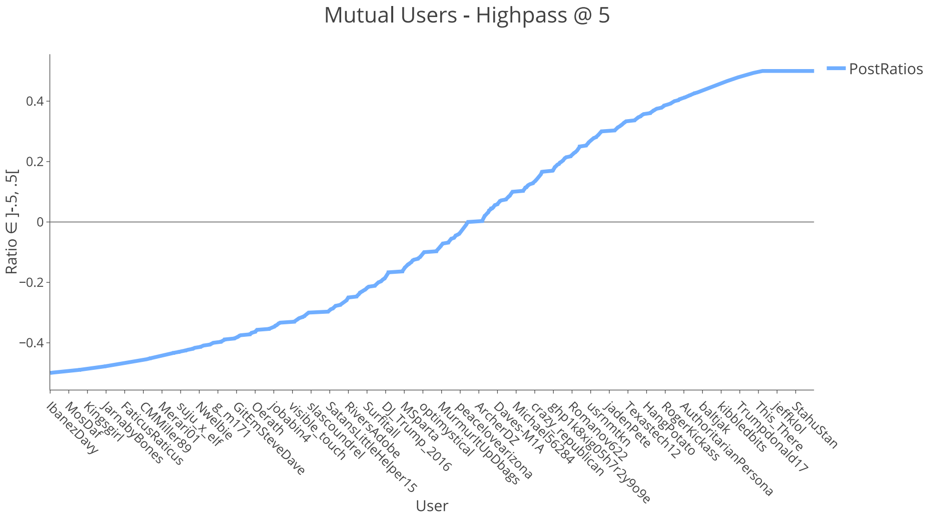


Fig. 15: Mutual Users commenting ratio, after filtering out all below 5 comments. More smooth than before, rather S-shaped. 81231 Users in total.

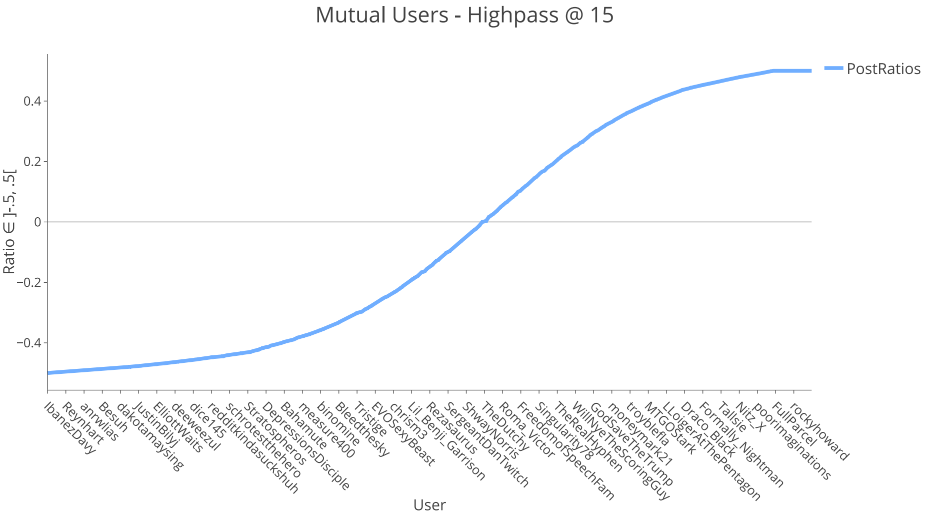


Fig. 16: Mutual Users commenting ratio, after filtering out all below 15 comments. Even smooth than before, distinctive S-shape. 51635 Users in total.

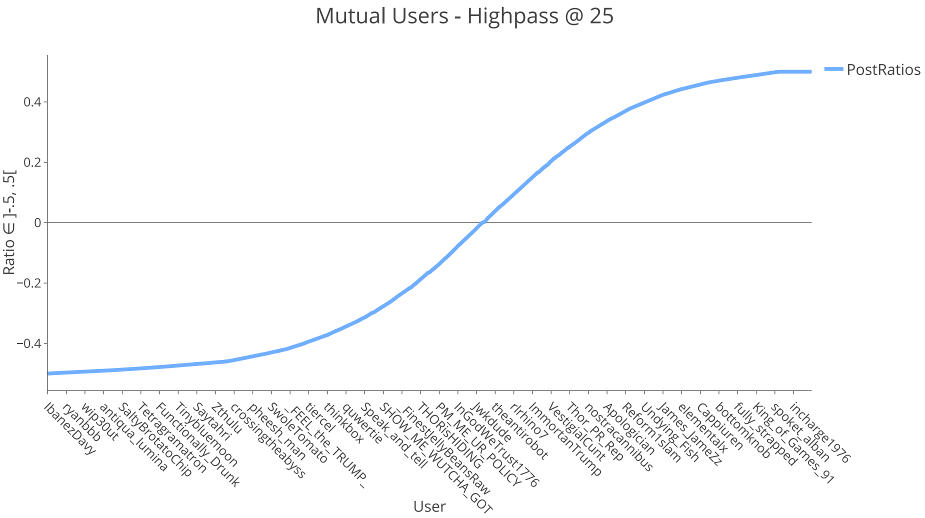


Fig. 17: Mutual Users commenting ratio, after filtering out all below 25 comments. Almost perfectly smooth S-shape. 39866 Users in total.

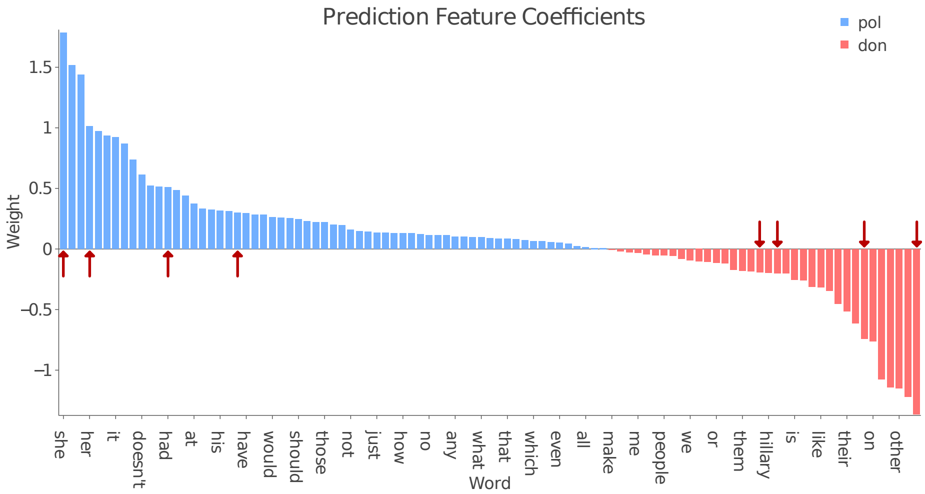


Fig. 18: Top 100 (99) words, including their respective weight assigned by the Regression. Interesting words are (from left to right): 'She', 'Clinton', 'Vote', 'I', 'Trump', 'Hillary', 'I'm', 'He'