Københavns Universitet

Topics in Social Data Science, Spring 2019

Game of Twitter

Who of the members of the Danish Parliament are connected, what are they talking about, how are they communicating and what drives engagement/popularity?

Tags: Network, Natural Language Processing, Sentiment Analysis

Lise Bohr, Nicklas Johansen Anna Beck Thelin

24/05/2019

Contents

Contributions

Joint: 1, 5, 6

wrz304: 2.1, 2.2, 3.1, 4.1, 4.2, 4.3

tzp831: 2.2, 2.3, 3.2, 4.3, 4.4, 4.5

jvs499: 2.1, 2.3, 3.1, 4.1, 4.2, 4.3, 4.5

1 Introduction

Voter participation has always been high in Denmark, with 80-90 percent of the population entitled to vote, voting. However, since the election for Parliament in 2007, a decline in voter participation has emerged (Folketingets Oplysning (2015)). According to DUF (2019) (Dansk Ungdoms Fællesråd) the participation among the younger population is especially low and this despite the fact that they do seem to have a genuine interest in politics and engagement in the society. They further argue that the low voter participation may be due to the young not believing that they can actually influence political system. Hence, politicians are faced with the task of closing the gap between the voters and politics, and this is where social media offers a platform for interactions between voters and politicians. Hopefully the political engagement created through the use of social media platforms increase voter participation going forward. And as argued by Schneider et al. (2007) politicians do increasingly seek to reach voters by means of social media.

Twitter has become an extremely popular communication tool among politicians and 161 out of the 179 Danish Members of Parliament ("Folketinget") have a Twitter account. Twitter works as a "microblog" where users get the opportunity to share opinions and debate different subjects every day. Hence, for politicians Twitter is a excellent platform for reaching their voters and express political opinions. Furthermore, as users can "like" or "dis-like" a tweet, politicians can also use tweets to detect voters' support.

However, politicians use Twitter very differently and especially the president of the United States, Donald Trump, has moved the boundaries for how members of state communicate on social media platforms. Trump is notoriously known for communicating very aggressively and distribute "fake news". Likewise Ott (2016) argues that Twitter privileges discourse that is simple, impulsive and uncivil, also among politicians. On the contrary, according to a survey study carried out in Korea by Hwang (2013) young voters generally have a positive attitude towards politicians' use of Twitter and thus, may be used to reach these voters for whom social media is a big part of everyday life.

Several papers has investigates the use of Twitter by politicians especially during elections. Vergeer et al. (2011) analyze the use of twitter during the 2009 EU election. At

this time only 13 percent of the candidates used Twitter, and hence the results should be interpreted carefully. The authors do however find a positive effect of the use on twitter on the number of votes. But whether this is merely because the candidates' twitter followers are better informed which will lead to endogeneity problems, cannot be told from the analysis. In a study by Tumasjan et al. (2010) the authors analyze a tweets sample consisting of 100.000 tweets from German politicians. Carrying out a text analysis, they find that politicians talk about the same subject on Twitter as they do off-line. Hence, they were able to pare a politicians to his/her party based on the words they use. Also interesting is how politicians communicate with each other on Twitter. Several studies on behavior on social media has found behavior is often homophilistic, and hence, politicians and citizens often only engage with people with who they have political agreement. This finding is also supported by Guerrero-Solé (2018) who, by using Twitter data from the Spanish election in 2016, finds that a re-tweet network strongly ensemble a echo-chamber.

In this study we investigate how Danish politicians use Twitter as a means of communication, prior to the Danish election for Parliament 2019. Creating a network using re-tweets with and without annotations we find that politicians mostly share content from members of their own parties, supporting the existence of an echo-chamber. However, re-tweets with annotation may be less endorsing, as the network here is less clustered. Next we create a network of the top 170 bigrams in the datasset, which show a clear pattern of the most popular subjects among the members of Parliament of twitter during that time period. Of these are for instance climate, retirement and health. Next, carrying out a sentiment analysis we establish in what sentiment the political parties communicate in general and about specific topics. A pattern here is that the parties closest to center the political spectrum communicates more positively than the extremist parties. Furthermore, the liberal parties that are currently in government has a more positive sentiment towards subjects such as climate, health and the EU relative to the social parties. This may be due to the fact that the liberal parties has an incentive to speak positively about political issues when in government. Lastly, we investigate which features predicts the number of likes a tweet gets. Here we find that the number of followers, re-tweets and sentiment are of most importance.

The paper is structured as followed: Section 2 lays of the theoretical framework of the analysis, hence, text analysis, model prediction and networks. In section 3 the data is presented along with some relevant descriptive analysis. Section 4 presents the results from the different analysis. Section 5 discuss and section 6 concludes.

2 Theory

This sections contains the different theories used in the analysis. All the computations and implementations of algorithms is carried out in python. The theories covered in this section are networks, sentiment analysis and predictive modelling.

2.1 Network

Networks are a useful tool used to model complex systems of relationships between entities. These networks can be constructed of to basic concepts: *nodes* and *edges*. Nodes are the components of a network and edges are the interactions between them. The edges of a network can either be directed or in-directed. Network theory can be applied on all sorts of systems in the real world and can be be used in any context where a meaningful definition of nodes and edges can be determined e.g. social networks, biological networks, web networks etc. (Barabási (2015))

2.1.1 Adjacency Matrix

A network can be represented mathematically by an adjacency matrix, A_{ij}. An adjacency matrix of a directed network consisting of N nodes has N rows and N columns, where the entries of the matrix are:

$A_{ij} = 1$ if there is a link pointing from node j to node i*
$A_{ij} = 0$ if nodes i and j are not connected

* In many applications networks are weighted, meaning that each link (i, j) has a unique weight, w_{ij}. For weighted networks the entries of the adjacency matrix states the weight, w_{ij}, instead of 1.(Barabási (2015))

2.1.2 Community Detection

Community detection is a method used to find highly connected groups of nodes in a Network. The communities are found by using the Louvain algorithm which finds the partition of nodes that maximizes the modularity. The modularity function measures how good a partition is. The modularity is defined as:

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \tag{1}$$

where A_{ij} represents the adjencency matrix; k_i, k_i are the sum of the weights of the edges attached to nodes i and j, respectively; $2m$ is the sum of all of the edge weights in the graph; c_i, c_i and are the communities of the nodes; and δ is a simple delta function.

The modularity takes values between -1 and 1, where a higher Modulairty implies a better partition. To maximize the value the Louvain algorithm repeats two phases, iteratively until the maximum of modularity is attained. Firstly, the algorithm search for "small" communities by optimizing modularity locally. Secondly, it aggregates nodes in the community and creates a new network whose nodes are the communities. This procedure can be performed by the best_partition() function in the python package named 'community'. (Barabási (2015))

2.1.3 Degree

The degree of a node is the number of edges it has to other nodes(Barabási (2015)). The degree of the i^{th}'th node in a network can be denoted k_i. The degree of nodes are used to produce two important properties when describing networks: *Degree distribution* and *average degree*. The *degree histogram* illustrates the distribution of degrees, by providing the frequency of nodes in the network. The *average degree*, $\langle k \rangle$, for undirected networks is:

$$\langle k \rangle = \frac{1}{N} \sum_{i=1} k_j \tag{2}$$

2.1.4 Clustering Coefficient

The degree to which the neighbors of a given node link to each other is captured in the *clustering coefficient*, C_i. E.g. if none of the neighbors of a node link to each other C_i=0, and oppositely if all the neighbors are connected C_i=1. For node i with degree k_i. The clustering coefficient is defined as:

$$C_i = \frac{2T_i}{k_i(k_i - 1)}$$

This measure is computed by the clustering() function in the python package 'networkx'.

2.1.5 Betweenness Centrality

Centrality measures are used to identify "important" nodes in a network. One of many ways to measure importance, is *betweenness centrality*, which computes the shortest-path betweennes centrality for nodes and determine the number of times a nodes acts is a part of the shortest path between to other nodes. Betweenness centrality of a node i is the sum of the fraction of all-pairs shortest paths that pass through v:

$$c_B(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}$$

where V is the set of nodes, $\sigma(s,t)$ is the number of shortest (s,t)-paths, and $\sigma(s,t|v)$ is the number of those paths passing through some node v other than s, t. If $s = t$, $\sigma(s,t) = 1$, and if $v \in s,t$, $\sigma(s,t|v) = 0$. This measure can be computed by the betweenness_centrality() function in the python package 'networkx'.

2.2 Sentiment Analysis

Sentiment analysis refers to the use of natural language processing to systematically identify, extract, quantify, and study subjective information. Sentiment analysis is also known as opinion mining and is often based on a text corpus from reviews and survey responses and social media.

The primary tools used in this analysis is the python library VADER[1] which is considered to be one of the best models for understanding the sentiment of the English language Hutto & Gilbert (2014). It is based on a Lexical approach which maps words to sentiment by building dictionary of sentiment negative, neutral, positive. VADER look at the sentiment score of each word in the sentence and decide what the sentiment score of the whole sentence is. The power of using VADER lies in the fact that is does not require training a model on labeled data Calderon (2017).

The sentiment score for a word is measured on a scale from -4 to +4, where -4 is the most negative and +4 is the most positive. The midpoint 0 represents a neutral sentiment. Sample entries in the dictionary are "horrible" and "okay," which get mapped to -2.5 and 0.9, respectively. In addition, the emoticons "/-:" and "0:-3" get mapped to -1.3 and 1.5. VADER is able to handle emoticons like ":-)", acronyms like "LOL", and slang like "meh" because it has been labeled by humans, though it is based on a average assessment score to mitigate the bias of humans individual perception.

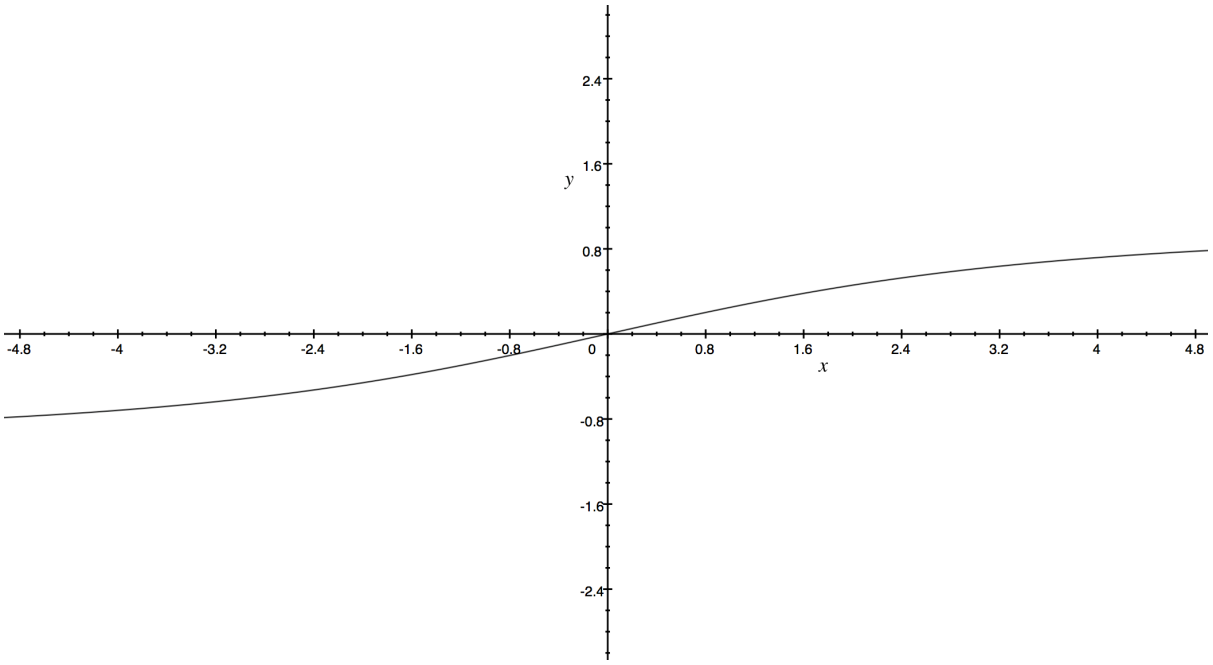
VADER returns a sentiment score in the range -1 to 1, from most negative to most positive. The sentiment score of a sentence is calculated by summing up the sentiment scores of each VADER-dictionary-listed word in the sentence and later uses the normalization:

$$\frac{x}{\sqrt{x^2 + \alpha}}$$

where x is the sum of the sentiment scores of the constituent words of the sentence and α is a normalization parameter. When x grows larger, the sentence score gets closer to -1 or 1, whereas modifying α can normalize more/less depending on how many words the sentence consist of. The normalization is shown on figure 1.

[1]GitHub repository https://github.com/cjhutto/vaderSentiment

Figure 1: Normalization of Sentiment Score



2.3 Predictive Modelling

In this section we will lay out the theory behind predictive modelling using machine learning. Specifically, the Random Forest model is used in this analysis.

2.3.1 Random Forest

Hyperparameters are used in machine learning model to increase performance. Random Forest is based on Decision Tree regression. Each tree has a node, which is the feature with the best predictive power. From this node the tree is split into a series of child nodes and ends with a leaf node. Hence, the tree is build by iteratively splitting up the internal nodes, until all the leaf nodes observations have been allocated, or until a specific criterion has been met. The regression seeks to minimize the mean squared errors (MSE) of each node. The performance of the prediction is evaluated by the MSE, which is measured:

$$I(t) = MSE(t) = \frac{1}{N_t} \sum_{i \in D_i} (y_i - \hat{y}_t)^2$$

where N is the number of training samples at each node t, D is the training subset at node t, y_i is the true value from data and $\hat{y}$ is the predicted target value samples mean,

which is defined as:

$$\hat{y_t} = \frac{1}{N} \sum_{i \in D_i} y_i$$

Random Forest has the advantages of having multiple trees which increase the predictive performance by capturing more of the model variance. This is though also one of main drawback of Random Forest, as a high number of trees may lead to over-fitting. Hence, tuning the number of trees in the model and the number of features to be considered by each tree when splitting a node, is important.

3 Data

Twitter data is retrieved from Crowdtangle[2] which is a platform for managing social media. Crowdtangle can be used for tracking social media activity and popular content as well as bench-marking of your own social media account. For this analysis we created a list of all the Danish members of Parliament with a Twitter account for whom Crowdtangle collects twitter activities going back three months. Data on each account is downloaded as csv files and merged together.

3.1 Descriptive Statistics

The final data set consists of 26,389 tweets from 141 members of the Danish Parliament. The tweets are from the period 29/01-02/05 2019 and hence approximately 3 month prior the announcement of the election. The descriptive statistics for each of the political parties are shown in table 1.

[2] https://www.crowdtangle.com/features

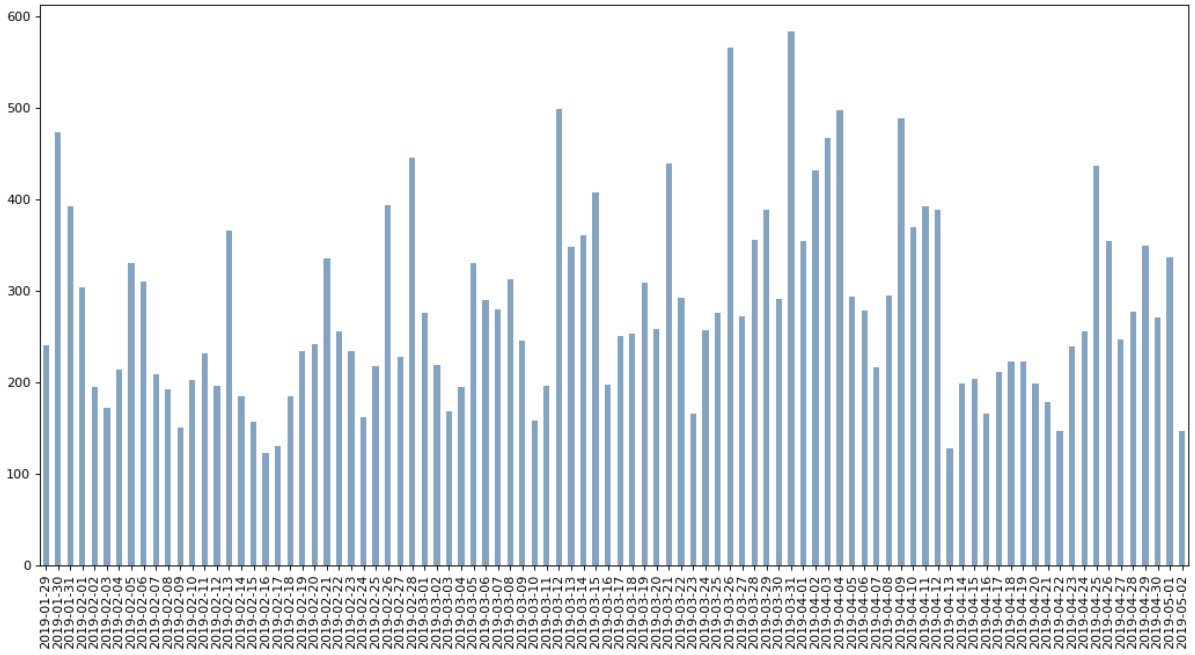
Table 1: Descriptive statistics of Political Parties

		Accounts	Total Activities	Avg. Account Activities
Social Parties				
A	Socialdemokraterne	34	5,700	168
B	Radikale Venstre	9	3,556	202
F	Socialistisk Folkeparti	8	2,881	314
Å	Alternativet	10	1,820	288
Ø	Enhedslisten	13	1,571	259
Liberal Parties				
C	Konservative	8	3,365	196
I	Liberal Alliance	12	2,599	199
O	Dansk Folkeparti	20	2,514	130
V	Venstre	27	2,382	132

Note: The descriptive statistics is calculated on the final dataset of 26,389 tweets.

Socialdemokraterne is the partly with the highest total number of activities in the given period. However, considering the average number of activities per party member, Socialistisk Folkeparty takes the lead. Both of these parties are social parties to which 5 of the 9 parties in the Parliament belongs. In the dataset, the social parties in general are the most active in Twitter considering their average activities. These parties are also opposition parties as the current government is liberal. A study by Vergeer et al. (2011) also finds that during the 2009 European Parliament election campaign in the Netherlands, the opponents parties were the most active on Twitter.

Figure 2: Tweets per Day



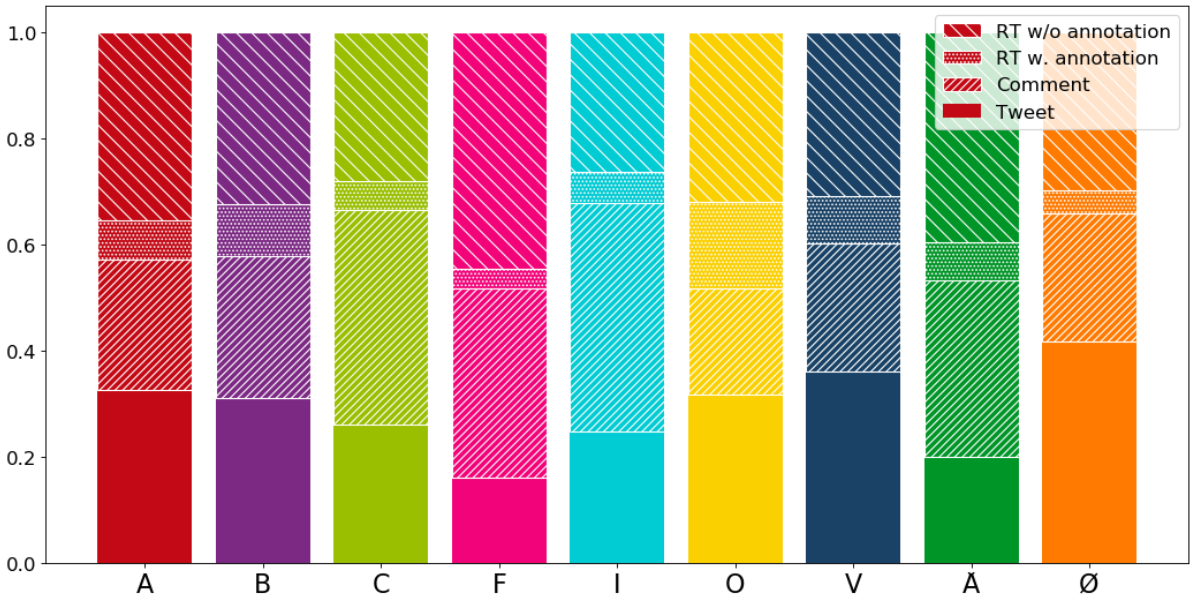
Next in figure 2 we will consider the number of tweets during the given time period. There is high variations in the daily number of activities throughout the period with some days as many as +550 tweets and other a few as 100 tweets. In general, low activity is due to weekends or bank holidays.

In the dataset, there are four different kinds of activities. These are tweets, comments, re-tweets with annotations and re-tweets without annotations. We distinguish between re-tweets with and without annotations, as these have been found to have different characteristics (see Appendix figure A1). Formerly, a re-tweet has been perceived as a sign of endorsement. However, studies now suggest that re-tweets to which the user has written an annotation are often more negative and hence, disagrees with the re-tweeted content. This distinction will be used later to create networks.

In figure 3 each party's share of the four different kinds of activities is depicted. The distribution between parties varies significantly, and it clear that parties use Twitter very differently. For instance, whereas *F* (Radikale Ventre) has a relatively small share of tweets, but a large share of re-tweets, Ø (Enhedslisten) has almost 50 percent tweets. We also notice here, that the party with the largest amount of re-tweets with annotations is

O (Dansk Folkeparti). Given that this kind re-tweet may be more negative, it could be expected that they will be found to have a negative sentiment in the analysis.

Figure 3: Activity Shares



Common for all of the parties is that they all how a fair share of re-tweeted content: 26-45% w/o annotations plus 5-20% with annotations. The sample contains a total of 9,221 re-tweets w/o annotations and of these, 3,620 are between the members of Parliament. Likewise, the sample contains 1685 re-tweets with annotations where 408 are between members of Parliament. Using this data, we will create networks describing how politicians are connected via these, both within and across parties. Here we can also detect whether the re-tweet network ensembles an echo-chamber in line with existing literature. If we do find evidence of this homophilic behavior among politicians in our network, we expect it to mainly reflect a strategic choices of primarily promoting politicians within their own party. We could expect to find such evidence in the re-tweet w/o annotations network.

3.2 Data Cleaning

Natural language processing often requires various kinds of data wrangling to become cleaned and structured, and ready for feature engineering and modeling. In our analysis we have used regular expressions (regex) to perform various string manipulations. It is

a technique developed in theoretical computer science and formal language theory which has allowed us to find tags, links and emojicons in our tweets by leveraging different string searching algorithms (Barnett (2019)).

During our data preparation we have not only used regex to clean the data but have also choosen to translate all of our tweets from Danish to English using Microsoft Translation API[3]. Since the average tweet is only 140 characters long and often consist of tags and urls, the tweets has not lost its meaning during the translation. Microsoft API is a large artificial neural network that predict the likelihood of a sequence of words. It is considered to be among the best translation models because of the a huge amount of high quality data it has been trained on.

To create a network of the most tweetet political topics we have manipulated our tweets with *NLTK*[4] a well known python library for string manipulation. It allows us to tokenize every tweet, to remove stop words and later create bigrams (a sequence of two adjacent elements from a string of tokens).

4 Results

In this section we will go through the results of our analysis. First we present the two networks of re-tweets without and with annotations followed by a network of bigrams. Next we analyze the sentiment of the political parties in general and in the context of specific political bozz-words. Lastly we seek to predict "Likes" using machine learning modelling.

4.1 Network on Re-Tweets without Annotation

Using the 3,620 re-tweets without annotations between the accounts in the sample, we have created two networks which are depicted in figure 4. The nodes represents accounts and edges re-tweeets. Weighing is done in accordance to how many re-tweets there is between two accounts.

[3]Microsoft Translation API: https://azure.microsoft.com/en-us/pricing/details/cognitive-services/translator-text-api/

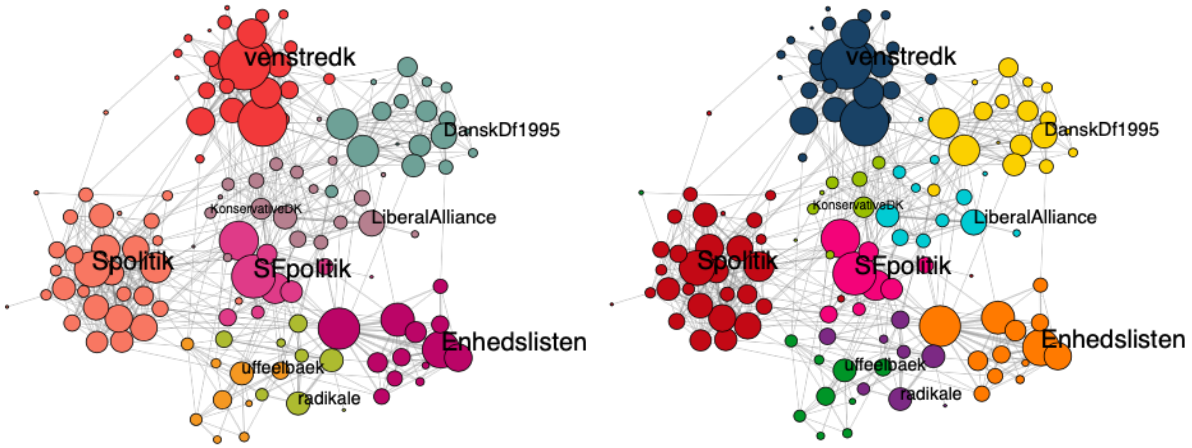
[4]NLTK: https://www.nltk.org/

The network dataset consists of:

- Nodes: 133
- Edges: 676

There are more than 6 times the number of edges than nodes, which suggests that the network is densely clustered. Looking at the networks we do see groups of nodes, which are common in social networks as these represent some underlying mechanisms. This suggests that that politicians mostly re-tweet fellow party members, which we see from the clearly separated groups. The difference in the two networks depicted, is that 4a is grouped by best community partition and 4b by actual political parties. The only difference between the groupings in the two networks, is that Konservative and Liberal Alliance are grouped together in the partition network. Likewise, the modularity score are almost equal in the two networks.

Figure 4: Network: Re-Tweets w/o Annotation



(a) Partition Network
Modularity Score: 0.758

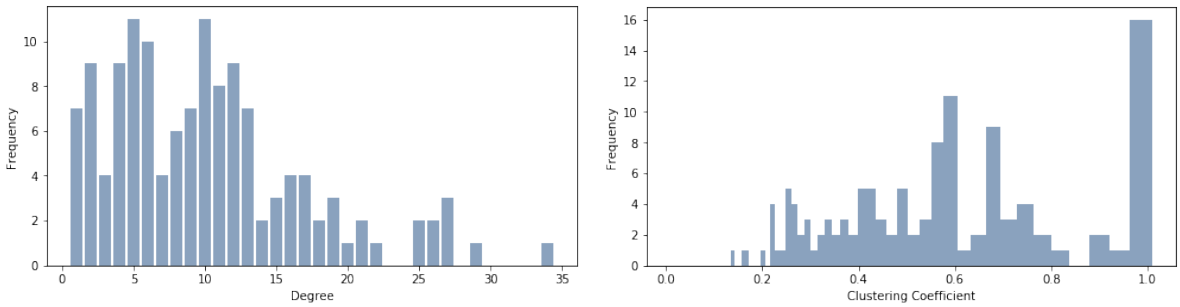
(b) Party Network
Modularity Score: 0.756

Furthermore, the network can be split into the two types of parties (Liberal and Social), where the Liberal parties are in the top-right of the network, and the Social parties in the bottom-left, which again indicate that the connections are strongest within the political ideologies. These observations all support the existence of a political echo-chamber on social media.

In figure 5a the degree histogram of the network is depicted. Degrees represent the

number of re-tweets connected to each account. Most accounts have a degree between 1 and 13, however a few have up to 25-34 degree. The average degree is 10.17 which suggests that the network is fairly connected, which is also visible in the figures 4.

Figure 5



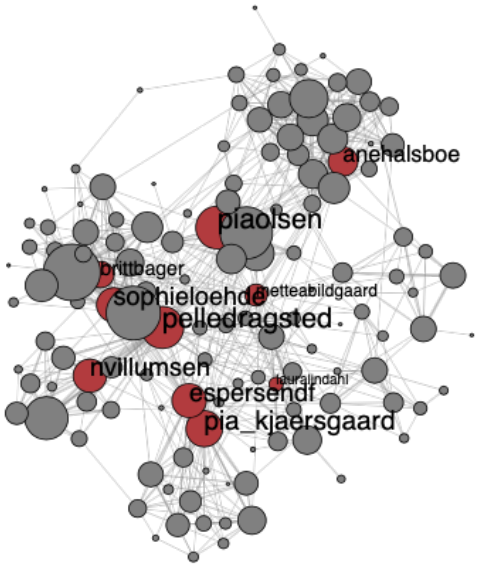
(a) Degree Histogram
Average: 10.17

(b) Clustering Coefficient
Average: 0.51

In figure 5b a distribution of the clustering coefficients is shown. The bins are log-transformed in order to get a clearer distribution. In general, the coefficient are high with an average of 0.505. This finding supports that the network is densely clustered.

We expect some accounts in the network to be more "important" than others. Here importance is measured by betweenness centrality given by the number of times a node acts as a bridge along the shortest path between two other nodes. The top ten "important" nodes are depicted in figure 6.

Figure 6: Betweenness Centrality



Common for these nodes is that they are situated in the center of the network, and hence, are well connected to groups other than their own. The top 10 accounts are represented by politicians from both Dansk Folkeparti, Enhedslisten, Venstre, Socialdemokraterne, Socialistisk Folkeparti and Liberal Alliance.

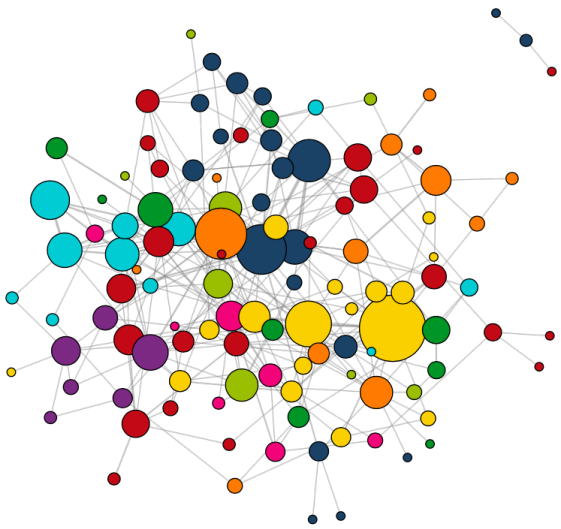
4.2 Network on Re-Tweets with Annotation

In figure 7 a network using observations on 408 re-tweets with annotations is depicted. The nodes represents accounts and edges re-tweeets. Weighing is done in accordance to how many re-tweets there is between two accounts. The network is grouped and colored according the tweeters' political parties. The network dataset consists of:

- Nodes: 114
- Edges: 295

The ration between nodes and edges in this network in significantly smaller than in the former, with suggests that this network is significantly more cluttered. This is clearly visible in figure 7, where accounts within political parties are spread across the network with more connections across political ideologies. This suggests that this type of re-tweet may indeed be less of an endorsement.

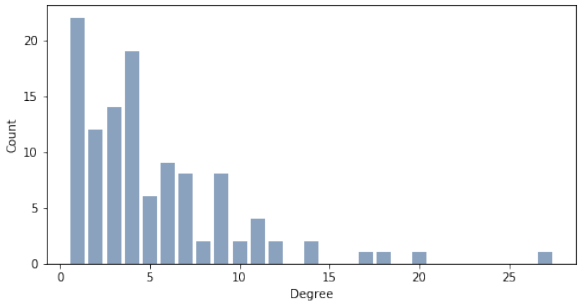
Figure 7: Network: Re-Tweets with Annotations



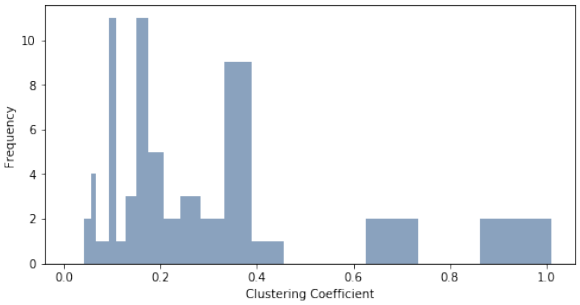
The degree distribution is depicted in figure 8a. The level of degree in this network mostly lies between 1-7, with a single observation with a degree of 27. The average

degree is 5.18, which is only half of the average degree in the re-tweet network without annotations (5a). This may be due to the fact that this network is a lot smaller, however it may also suggest that it is less connected.

Figure 8



(a) Degree Distribution
Average: 5.18

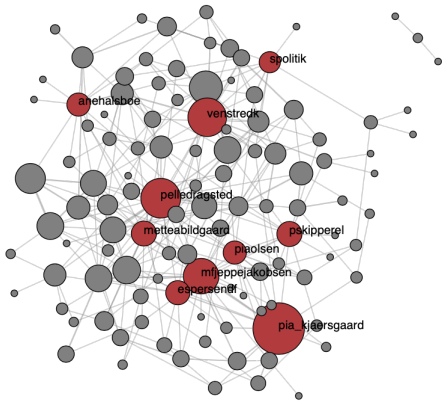


(b) Clustering Coefficient
Average: 0.12

Clustering coefficients are depicted in figure 8b. The ratios of the coefficients likewise suggest that the network is less clustered, as the coefficients have a mean of only 0.12.

We once again investigate who are "important" in the network by betweenness centrality. The top 10 "important" accounts are depicted in figure 9.

Figure 9: Betweenness Centrality



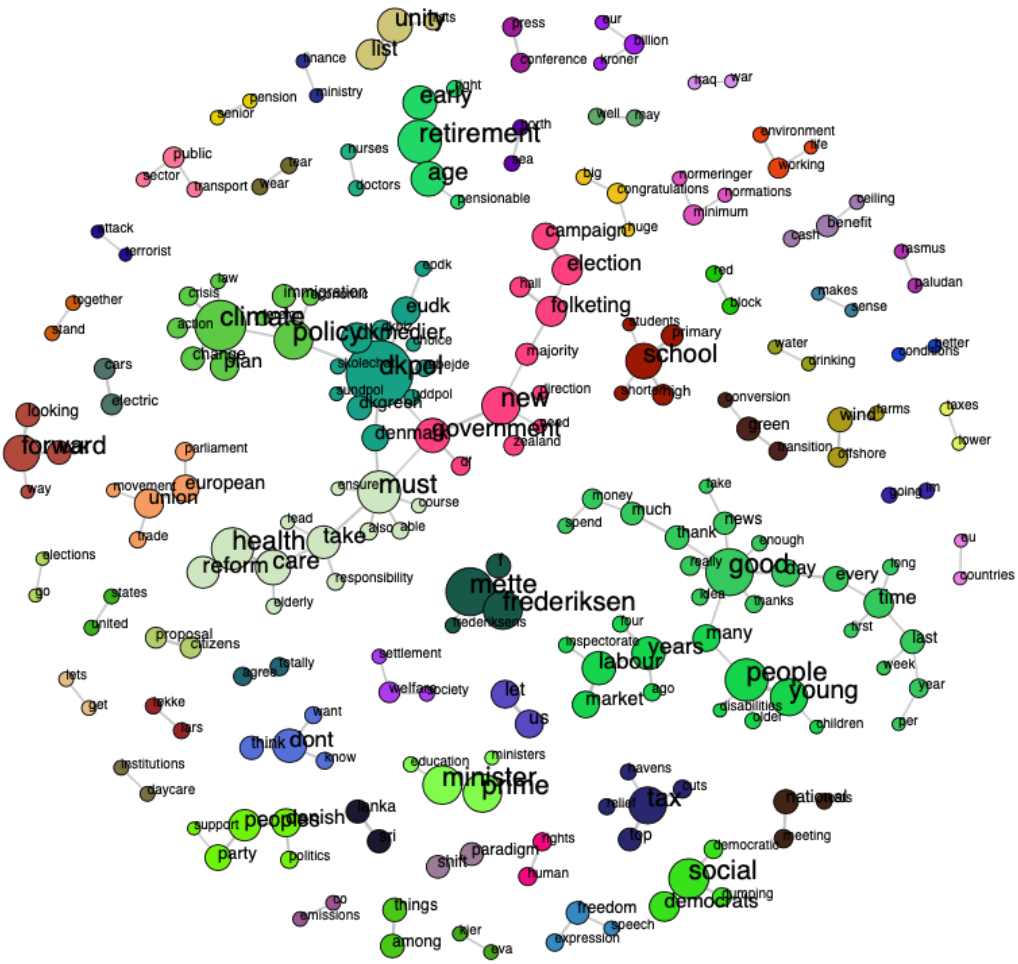
The "important" accounts are more randomly distributed in the network, but still represents a variety of political parties (Ventre, Socialdemokraterne, Dansk Folkeparti, Enhedslisten, Konservative and Social Folkeparti). Likewise, some of "important" accounts in figure 6 are also found to be important in this network. These are the accounts

of pia_kjaersgaard, espersendf, pelledragsted and metteabildgaard.

4.3 Network on Bigrams

In figure 10 a network of the top 170 bigrams is depicted. The nodes are words and the edges indicate that the two words are used sequentially. The words are grouped by colors using community detection. The bigrams are constructed on our entire text corpus including tweets, re-tweets and comments.

Figure 10: Network: Bigrams of Words



From the network of bigrams above the most common combinations of words used sequential in tweets are visualised. When inspecting the network, it is easily observed that the center consists of larger communities, which appears to be some of the main topics politicians discuss on twitter. To mention some, the largest node is representing

the word 'dkpol', which is a commonly used hashtag when posting political content on Twitter. This node is well connected with other large communities in the network, such as climate, health and election which are political topics that have been heavily discussed by politicians during the period covered by the data. Secondly, a couple of isolated larger communities outline other important political topics such as retirement and the European Union. Lastly, pairs or triplets are observed in the periphery of the circle that form common short phrases, which are either less discussed political topics (e.g. 'public'+'transport'), combinations of first and last name (e.g. 'rasmus'+'paludan') or just combinations of two words forming common expressions (e.g. 'makes'+'sense').

4.4 Sentiment Analysis

Table 2 illustrates the average sentiment scores of each party. The positive, neutral and negative scores are ratios for proportions of text that fall in each category and together they sum to 1. The compound score is computed by summing the scores of each word in the lexicon and is the most useful metric if you want a single uni-dimensional measure of sentiment for a given sentence.

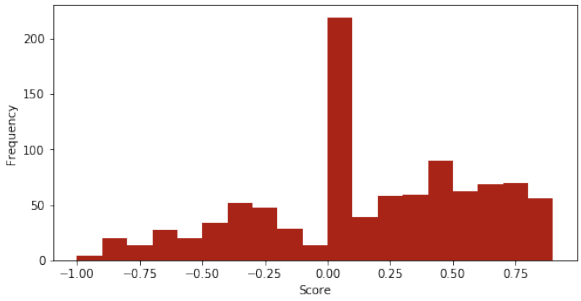
Table 2: Political Party Sentiment Analysis

		Negative	Neutral	Positive	Compound
Social Parties					
A	Socialdemokraterne	0.06	0.82	0.12	0.15
B	Radikale Venstre	0.05	0.82	0.13	0.20
F	Socialistisk Folkeparti	0.05	0.82	0.13	0.17
Å	Alternativet	0.05	0.84	0.10	0.14
Ø	Enhedslisten	0.07	0.83	0.10	0.10
Liberal Parties					
C	Konservative	0.06	0.81	0.13	0.16
I	Liberal Alliance	0.06	0.82	0.12	0.16
O	Dansk Folkeparti	0.07	0.82	0.11	0.10
V	Venstre	0.04	0.83	0.13	0.24

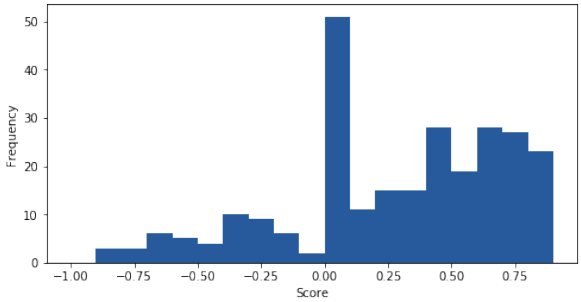
Note: Sentiment analysis per political party is measured as the mean sentiment of all types of activities in the sample.

The levels of the negative, neutral, positive scores represented in Table 2 are very similar, but there are some minor differences between the parties: Enhedslisten and Dansk Folkeparti has the highest negative scores, Alternativet has the highest neutral score, Radikale Venstre Socialistisk Folkeparti and Konservative and Venstre has the highest positive score. Instead, when considering the compound score, a larger variation is observed. Venstre is the party with the highest compound score overall, indicating that the party accounts for the most positive content on Twitter. In general, it is discovered that the parties closest to the center of the political spectrum, post more positive content, whereas the most right wing together with the most left wing parties post more negative content. This is at least the pattern when considering all the tweets in the data. In the following, four political topics from the network of bigrams have been selected to inspect differences in sentiment distributions for liberal versus social parties.

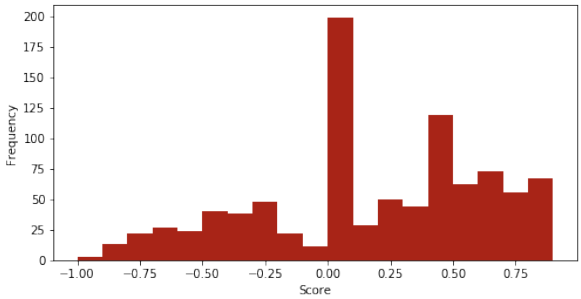
Figure 11: Sentiment Analysis



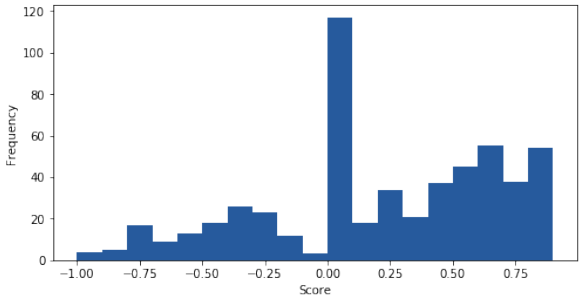
(a) Climate: Social
Average: 0.17 - Count: 997



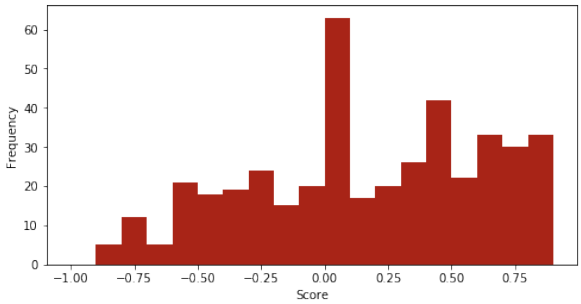
(b) Climate: Liberal
Average: 0.30 - Count: 279



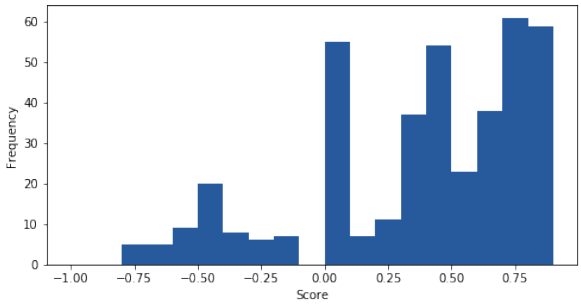
(c) EU: Social
Average: 0.18 - Count: 967



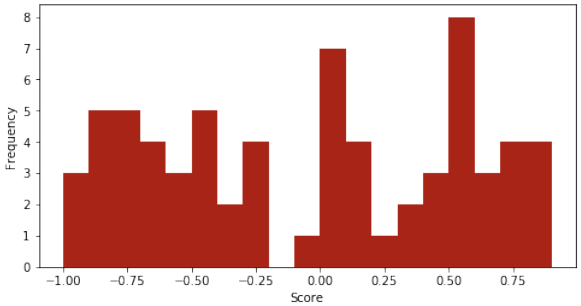
(d) EU: Liberal
Average: 0.23 - Count: 577



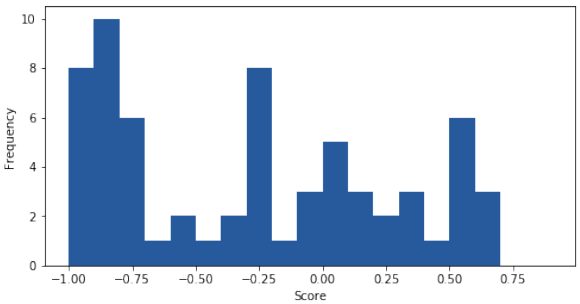
(e) Health: Social
Average: 0.17 - Count: 433



(f) Health: Liberal
Average: 0.39 - Count: 420



(g) Paludan: Social
Average: -0.02 - Count: 70



(h) Paludan: Liberal
Average: -0.22 - Count: 68

In Figure 11(a)-(d) the distribution of sentiments for activities containing the words

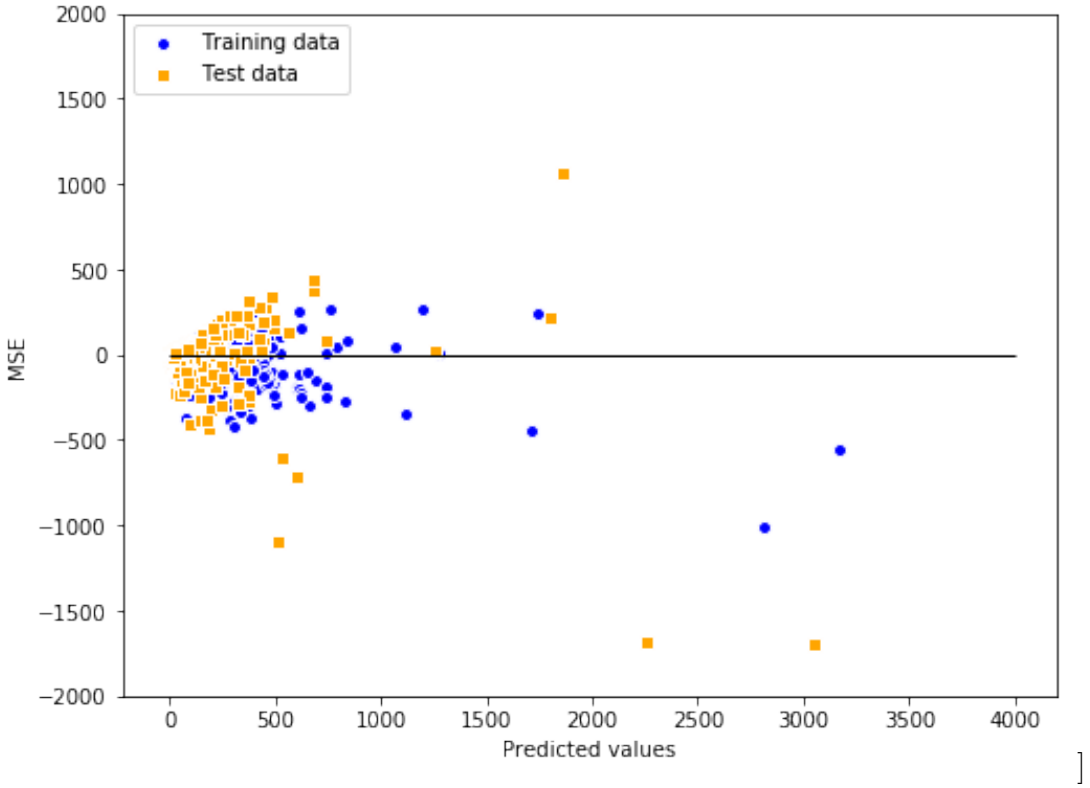
climate and EU, respectively, divided on liberal and social parties are depicted. The four distributions are very similar, indicating that the topics EU and Climate do not differentiate when considering the sentiment. Both sides has a subtle overweight of positive tweets and the liberal parties has more positive tweets than the social parties which is also reflected in higher average scores. In figure 11(e)-(f) sentiment distributions for activities containing containing the word 'health' are shown. Again, the liberal parties are more positive and here the differences in both distribution and average score are more clear.

Lastly, figure 11(g)-(h) show sentiment distributions for tweets containing containing the word 'paludan'. This is the last name of a highly controversial politician which is also reflected in the histogram. The previous topics had a large proportion of 'neutral' scores (compounds score close to 0). But for this keyword, sentiment score are more divided and located away from the centre, however with an overweight to the negative side, which is supported by the averages scores being negative for both social and liberal parties.

4.5 Predicting Likes on Twitter

The aim this analysis is to predict the number of "Likes" a twitter post has, given a number of features (variables). The descriptive statistics of these are shown in table 1 in Appendix. By tuning the hyperparameters with cross-validation (grid search), we find that the optimal number of trees is 6 and the depth 8. In figure 12 the residual from the Random Forest are shown. The test data seems to be performing worse than the training data. However, none of them are randomly distributed around the mean of 0, which suggests that there is heteroskedasticity in the model.

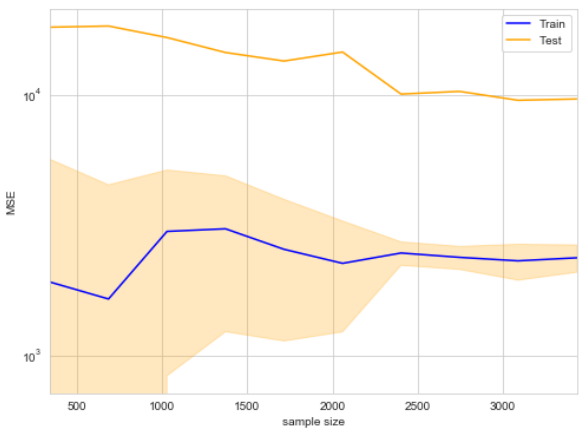
Figure 12: MSE on train and test



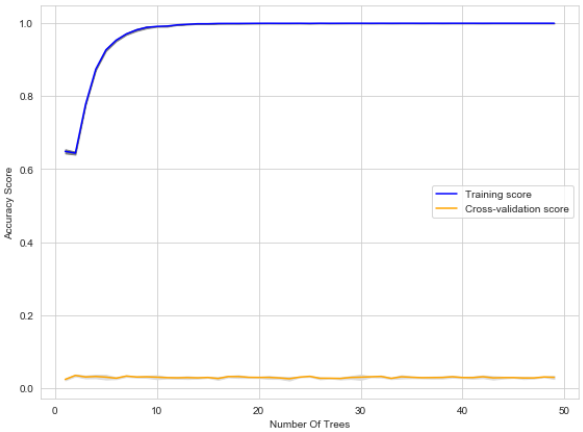
]

In figure 13 the learning and validation curves are shown. In the learning curve, we first notice that the distance between the curves is quite big, which indicated that the model is over-fitting. The same is true for the validation curve, which show the model accuracy as the number of trees increase. After only 1 tree, the training score increases significantly. The cross-validation score does not seem to be changing when including more trees.

Figure 13



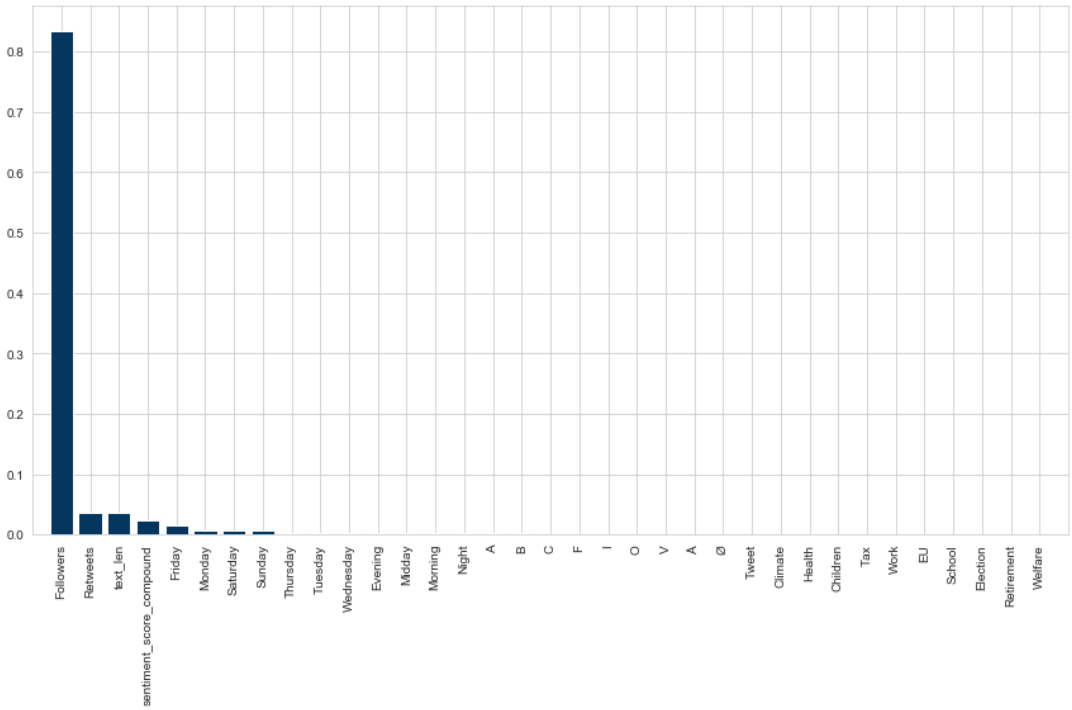
(a) Learning Curve



(b) Validation Curve

The ranking of the most important features is depicted in figure 14. The by far most important feature of the number of Followers, which explains more than 80 percent of the variation in the number of Likes. Next are Re-tweets, length of Tweet and Sentiment Score. From here the rating decrease, and the features becomes more or less insignificant for the prediction.

Figure 14: Feature Importance



5 Discussion

This analysis is carried out using Twitter data. The data was obtained through CrowdTangle which included tweets 3 month back in contrast to Twitter who only offers users to go 14 days back through their free API. Our dataset is restricted to only include twitter activities from Danish politicians that are members of Parliament. Hence, the conclusions drawn from the analysis may not be generalized to all Danish politicians. We could expect that there does exist a selection bias as members of Parliament may communicate in a different way from non-members. Likewise, it may also be that communication on Twitter is different from the real-world communication. However, as argued Tumasjan et al. (2010) there does seem to be a connection between online of offline subject discussions though, as argued by Ott (2016), it may be done more aggressively on Twitter. One could argue that the limit on 270 characters in a tweet, differentiates the communication on twitter from traditional media. This force politicians to cram potentially complicated ideas into small messages, which potentially leads to a different form of communicating that is fast, easy to digest, but encourage a black-and-white view of the political scene.

The distinction of re-tweets with and without annotations showed that there is a noticeable difference. The result implies that re-tweets without annotations added can be perceived as an endorsement of the content shared, which also supports why clear communities within parties was easily identified. This behavior encourage the existence of an 'echo chamber', when no annotations are used. Oppositely, it can be inferred that when an annotation is added before re-tweeting, it is more often an indication of a disagreement. This is encouraged by cluttered network, with a high amount of links between parties and ideologies. The finding is also supported by the fact that the party with the largest share of such re-tweets, Dansk Folkeparti, was found to have an on average, negative sentiment.

We have made a network of bigrams that describes what the Danish politicians are tweeting about. The network shows simple patterns of the most "talk-about" subject during the sample period, however the network could have been more sophisticated if we instead build it using n-grams. With more available data, we could have trained a Word2Vec model leveraging word embeddings to create clusters, which would lead to an

even greater impression of what the politicians are talking about.

We decided to translate all of our tweets before using VADER to score the sentiment of every tweet. Peculiarities of the Danish language might have been lost in translation. VADER recognized both the positive, neutral and negative words and form a quick and simple compound score for every tweet. Even though VADER is a well known library it is yet to improve on understanding humor, ambiguity, irony and other forms of complexity that are found in the natural languages.

The number of tweets was sufficient for building networks but may be insufficient for training a great machine learning model is this requires substantial amounts of data, where we only had 5,722 observations. The relatively little amount of data may be part of the reason for the detected over-fitting, which could be decreased by including more data. Also, including features such as clicks and impressions, which are commonly known for being important for tweet-performance, are not included. Including such variables could potentially increase the predictive power of the Random Forest model.

6 Conclusion

In this paper, we investigated how members of Danish Parliament communicate on Twitter, using Twitter activities from the period 29/01-02/05 2019. From the analysis we find that politicians endorse fellow party members by re-tweeting their content. As this is done in a relatively great extent, the network of politicians come to look like an echo-chamber and hence, Twitter may breed homophilistic behavior. However, the analysis do though also suggest that re-tweets with annotation may not be endorsing, but a sign of disagreement.

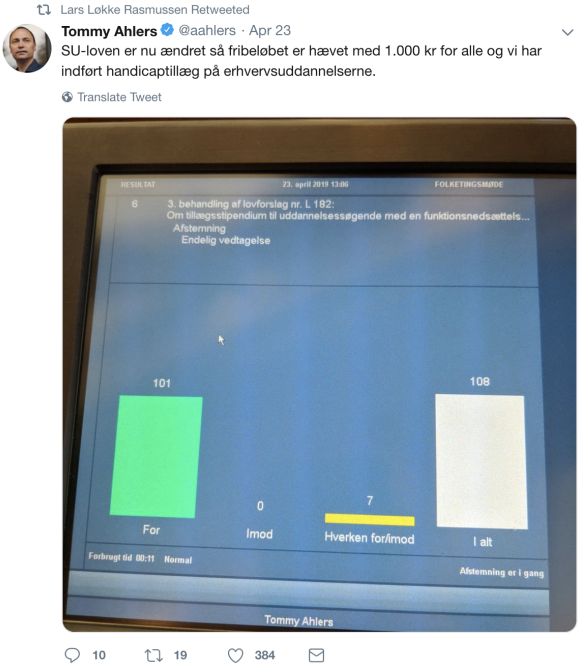
Investigating which subjects the politicians talk about, we find that especially climate, health and EU are high on the list. However, the sentiment by which the subjects are communicated differs between the Liberal and Social parties. In general the sentiment analysis revealed interesting patterns of difference in sentiments across parties. It was found that the parties close the the centre of the political spectrum communicate more positively, whereas the extremist parties communicate more positively.

Lastly, using Random Forest modelling we predicted "Likes" on tweets given a number of features. Here we find that the most important features are the number of followers, re-tweets, the length of the tweet and the sentiment. However, we also find evidence of heteroskedasticity and hence, there is variation in the number of likes that cannot be explained by our model.

7 Appendix

Figure A1

(a) Re-Tweet with annotation



(b) Re-Tweet w/o annotation

Table 1: Random Forest Feature Statistics

	Mean	**Std**	**Min**	**Max**
Followers	20,724.05	35,125.99	0	247,811.00
Retweets	9.81	55.72	0	2,584
Tweet Length	188.31	74.58	4	390
Sentiment Score	0.19	0.53	-0.98	0.98
Day of the week				
Monday	0.12	0.33	0	1
Tuesday	0.18	0.39	0	1
Wednesday	0.16	0.37	0	1
Thursday	0.17	0.38	0	1
Friday	0.15	0.36	0	1
Saturday	0.10	0.30	0	1
Sunday	0.11	0.31	0	1
Time of the day				
Evening	0.29	0.46	0	1
Midday	0.37	0.48	0	1
Morning	0.31	0.46	0	1
Night	0.03	0.16	0	1
Political Party				
A	0.20	0.40	0	1
B	0.07	0.25	0	1
C	0.05	0.21	0	1
F	0.06	0.24	0	1
I	0.09	0.28	0	1
O	0.15	0.36	0	1
V	0.14	0.35	0	1
Å	0.09	0.29	0	1
Ø	0.15	0.36	0	1
Frequent Words				
Climate	0.06	0.25	0	1
Health	0.04	0.20	0	1
Children	0.04	0.20	0	1
Tax	0.03	0.17	0	1
Work	0.03	0.18	0	1
EU	0.03	0.17	0	1
School	0.02	0.15	0	1
Election	0.03	0.16	0	1
Retirement	0.02	0.13	0	1
Welfare	0.02	0.14	0	1

References

Barabási, Albert-László. 2015. Network Science. Cambridge University Press.

Barnett, Matthew. 2019. regex. https://pypi.org/project/regex/.

Calderon, Pio. 2017. VADER Sentiment Analysis Explained. http://datameetsmedia.com/vader-sentiment-analysis-explained/.

DUF. 2019. Unge er politisk aktive - men stemmer bare ikke.

Folketingets Oplysning. 2015. Folketingsvalgene 1953-2015.

Guerrero-Solé, Frederic. 2018. Interactive Behavior in Political Discussions on Twitter: Politicians, Media, and Citizens' Patterns of Interaction in the 2015 and 2016 Electoral Campaigns in Spain. Sage.

Hutto, Clayton, & Gilbert, Eric. 2014. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. AAAI Publications.

Hwang, Sungwook. 2013. The Effect of Twitter Use on Politicians' Credibility and Attitudes toward Politicians. Journal of Public Relations Research.

Ott, Brian L. 2016. The age of Twitter: Donald J. Trump and the politics of debasement. Critical Studies in Media Communication.

Schneider, Steven M., Kluver, Randolph A, & Foot, Kirsten A. 2007. The Internet and national elections: A comparative study of Web campaigning.

Tumasjan, Andranik, Sprenger, Timm O., Sandner, Philipp G., & Welpe, Isabell M. 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment.

Vergeer, Maurice, Hermans, Liesbeth, & Sams, Steven. 2011. Is the voter only a tweet away? Micro–blogging during the 2009 European Parliament election campaign in the Netherlands. First Monday, **16**(8).