Measuring Political Polarization in a Social Media Environment

Student details

Name: C. van den Hurk
Student number: U885201

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

Thesis committee

Supervisor: drs. C.D. Emmery
Second reader: dr. M. Alimardani

Tilburg University
School of Humanities & Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands
25 June, 2021

Measuring Political Polarization in a Social Media Environment

C. van den Hurk

Abstract

With the recent perceived political polarization trend in the US, the negative consequences of political polarization become an increasing danger for society. Social media seems to play a role in polarizing society, although this is not very well understood. In order to understand political polarization on social media, it must be clearly defined and measured. Using existing studies on polarization, a definition of political polarization is constructed and polarization indicators are defined. Previous work that measured political polarization in social media did this surrounding a central topic or in partisan terms. This study attempted to measure political polarization in a social media environment in terms of political ideologies to gain a more direct understanding of political polarization in an online environment. Using distilBERT word-embeddings with a Logistic Regression layer for text classification, two models trained on the liberal and conservative and on the communism and conservative ideologies are trained with data taken from online forum Reddit. To measure polarization, comments were collected from the subreddit page r/Politics. The model trained on communism and conservatism showed promising results in the evaluation metrics. However, the results obtained from both models in the classification dataset do not support the perceived political polarization trend in the US and thus do not provide evidence that political polarization can be measured in terms of political ideologies in a social media environment using this method. This study however does offer advice and lessons that can be valuable for future studies on this topic.

1. Introduction

On January 6th 2021, the US capitol building was subject to a violent attack. On this day, the electoral college votes were being officiated. Trump supporters stormed the capitol building believing that the election was stolen and that Trump was the rightful winner. In the aftermath of this attack, the narrative surrounding this attack seemed to be concerning terms like polarization and social media, often in relation to each other (Brownstein, 2021; Smith, 2021; Cassese et al., 2021). This perceived polarization and division in American society is mentioned as one of the main points in the inauguration speech by new president Biden, further emphasizing the narrative of the US as a divided and polarizing nation.

According to the Pew Research Center, America has rarely been more divided than in recent times. Pew Research Center also conducted a survey with around 5000 US adults to measure political sentiments. The survey's results show a clear shift towards a more consistent liberal/conservative view of the public. In particular, this polarizing shift is visible between 2011 and 2017 (2017 was the last year results for the survey are available)(Pew Research Center, 2017).

The concept of polarization in a political context is not new, and the possible consequences of polarization have therefore also been studied. Polarization can act as an exacerbator for intolerance and discrimination and a catalyst for increasing violence in society (Carothers & O'Donohue, 2019). Additionally, polarization can be harmful to democracy as it can have negative consequences for its efficiency and operation (Körösényi, 2013). The recent polarizing trend in the US and its possible negative consequences make it a currently relevant issue worth researching in order to limit these negative consequences.

The relation between social media and polarization seems to be worth researching in particular as polarization and social media are often talked about in reference with each other. An example of a practical connection between social media and political polarization is the Russian interference in the 2016 US election between Hillary Clinton and Donald Trump. Russia used many social media accounts on sites like Facebook, Instagram and Twitter in order to infiltrate and influence the political discussions in several communities in the US (Howard et al., 2019). The IRA (Internet Research Agency based in Sint-Petersburg, Russia) created thousands of social media accounts in order to boost Trump's campaign. Social media was thus being used as a tool to steer elections and ultimately influence social discourse in the US.

One of the theories that might explain the connection between polarization and social media lies within the term 'echo chamber'. Echo chambers are environments where someone is only exposed to opinions that agree with their own. Algorithms on social media that show people only things they like, lent themself to form such environments. For example, exchanges on twitter seem to be mostly between users with similar viewpoints (Barbera, 2013). These echo chambers are shown to be reinforcing polarization on political topics (Garimella et al., 2018). Also, echo chambers make one more likely to believe in things that are not true (e.g. fake news) (Sunstein, 2018).

Alternatively, in a study done by Bail et al. (2019), evidence was found for an effect that is contrary to the 'echo chamber' effect. While exposing Republican voters to a liberal twitter bot, the Republican voters held significantly more conservative views afterwards. Similar results were found for Democrats being exposed to a conservative twitter bot, however this effect was not statistically significant. These findings offer an interesting contrast to the theory of echo chambers explained earlier.

A study by Hong & Kim (2016) examined both these theories in a political polarization context. This study researched the role of social media in partisan polarization using politician's twitter activities, like the number of tweets and followers. This approach to measuring polarization is more specifically aimed at the role of social media and is not directly related to a public perspective of polarization. Hong & Kim (2016) found that social media might play a role in 'heightened levels of extremism', thereby further increasing online political partisan polarization. However, results as to what mechanism caused this political polarization were not conclusive. Although there seems to be a link between social media and political polarization, the exact nature of this link is still unclear.

Polarization is clearly perceived to be an important issue in US society today. Polarization in a political context can have serious consequences for society as a whole. Social media offers both a potential explanation as an influence for political polarization as well as an opportunity to sample the thoughts of millions of people when it comes to political

ideas. Being able to measure political polarization directly in a certain social media environment could offer a way to further understand, contextualize and monitor this political polarization. Developing such a tool could prove to be a valuable asset from both a scientific and a public perspective.

A scientific interest for this tool would be centered around looking for relations between certain policy implementations, political elections and other, also non-political, events and political polarization. A way to quantify and measure political polarization in a social media context might additionally help in the research towards the effects of echo chambers and fake news in social media. It is about understanding the phenomenon of political polarization and its causes on a societal scale.

Public interest is directly related to the benefits of understanding political polarization and its underlying mechanisms. A further understanding of the causes of political polarization could help recognize signs of polarization early in order to limit its negative societal consequences. Understanding political polarization in relation to social media especially could help recognize signs of political polarization online and in real time. In order to realize such a tool, the use of machine learning seems to be essential. This would allow big amounts of data to be analyzed without having many labor hours invested in reading social media comments. Especially, the field of NLP (natural language processing) seems to be right for such a tool. NLP is specialized in the analysis of human language by machines. Following this, the main research question for this thesis is formulated as;

To what extent can NLP be used to detect political polarization in a social media environment?

This research question is further divided into 4 sub-question;

What exactly is political polarization? What are its different forms and how can it be defined?

What are indicators that can be measured that suggest political polarization?

The two above formulated sub questions are aimed to understand political polarization as a concept. Both in terms of the different forms it comes in and in terms of particular measurements or characteristics that would indicate political polarization. Such clearly defined characteristics (grounded in literature) are needed to be able to measure polarization.

Which social media environment would be suitable to measure polarization?

This sub question is aimed to look for the right type of data to both train the model and use the model to measure the characteristics of political polarization specified by the first two sub questions. Having the right kind of data for training and measuring is vital for answering the main research question.

What is a suitable algorithm to measure these political polarization indicators in a social media environment?

There are many different NLP algorithms used in different circumstances and for different purposes. Finding the right algorithm for the specific purpose to answer the main research question is central in attempting to answer the question.

2. Related work

In this section, extant work is searched in order to identify the underlying factors and define political polarization. Existing measurements of political polarization are explored and assessed. Finally, an overview of the existing models and algorithms that are available for text representations is given in order to motivate research choices.

2.1 Defining Political Polarization

In order to measure political polarization, the concept itself and its components must be defined. Political polarization can be referred to as both a state and a process (DiMaggio, 1996). However, this thesis is interested in the increase in political polarization over time, so political polarization will be seen as a process.

Political polarization seems to be divided into two main sub categories: elite polarization and mass polarization. Elite polarization is associated with political parties and partisanship versus mass polarization which is centered around the Populus (Hetherington, 2009). In this thesis, the focus is on ideologies (e.g. liberal vs conservative) and not political parties (e.g. Republicans vs Democrats), therefore political polarization will be looked at from a mass polarization perspective.

Unfortunately however, there does not seem to be a consensus about the exact components that are used to measure mass political polarization. In a literature review on this topic, Lelkes (2016) looks at a debate between Abramowitz & Saunders (2005) on one side and Fiorina et al. (2005) on the other. Lelkes (2016) points to the difference in definition of political polarization as a cause for this debate. Abramowitz & Saunders (2005) seem to look at political polarization from a perspective of alignment and ideological consistency. Political polarization can be seen as increasingly sorting the population into parties (e.g. ideologies like liberal and conservative) and to what extent this sorting becomes internally consistent. Fiorina et al. (2005) see political polarization as the divergence of ideologies within a Populus.

Because there does not seem to be a consensus about the definition of political polarization and its components, a working definition is compiled based on these works. The definition for this thesis should contain the four elements that have been introduced above. Political polarization should be a process, looked at from a perspective of mass polarization and should take into account both polarization as the sorting of this mass into groups and the ideological distance between those groups. Before constructing the working definition for this thesis, it is useful to understand what is meant by an ideology and especially what is referred to by the ‘distance’ between these ideologies.

Ideologies can be seen as an understanding of the meaning of values and how they relate to each other (Blattberg, 2001). ‘What does liberty entail?’ and ‘Is liberty more important than safety?’ are examples of questions different ideologies might have different answers to. For the purpose of this thesis, it is more relevant to see how these ideologies themselves relate to each other in order to have a meaning for the distance between them. In popular speech, they are often referred to as ‘left’, ‘moderate’ or ‘right’ with variations like ‘far-left’ and ‘moderate-right’ for example. These words suggest that political ideologies can

be placed on a linear spectrum. This spectrum ranges from communism on the far-left and fascism on the far-right (Heywood, 2017). Such a linear relation between ideologies seems a fairly simple and straightforward representation of the relation between ideologies, but useful for the purposes of this thesis.

Taking all these components into account while compiling a working definition for political polarization in this thesis results in a definition that can be formulated as follows;

Political polarization is the process of the division of the Populus into clearly divided and internally consistent ideological groups on a linear ideological spectrum.

Polarization is thus looked at as the ideological divergence and ideological alignment on a linear spectrum, with communism on the far-left and fascism on the far right. However, the most dominant ideologies in the US are less extreme. According to a recent study by Saad (2021) published by Gallup, conservatives and moderates are the most predominant ideological groups in the US. The other group that was measured was the liberal ideology. Studies concerning political polarization like Saad (2021) and Pew Research Center (2014) both look at polarization in terms of the conservative vs liberal ideologies. For the goal of this thesis, it will be interesting to not only look at polarization in terms of liberal vs conservative, but also take into account the more extreme variants to have a bigger spectrum to examine and to compare.

2.2 Measuring Political Polarization

Several studies have looked at and tried to measure political polarization. One of these is a paper by Klinker and Hapanowicz (2015). This paper analyzed the 2004 US election and aimed to see if America was more polarized during this election than in the 2000 election. This study used voter data to measure polarization. The downside of this is that polarization can only be measured on times where voter data is available, which limits the frequency of the measurement.

Another study that tried to measure political polarization is a paper by Azzimonti (2013). This study aimed to create a 'high-frequency measure of polarization' and used the frequency of newspaper coverage of political disagreements of government policy. This resulted in a '*political polarization index*' that was available on a monthly and quarterly basis. The frequency with which political polarization can be measured is an important issue that was brought forward, as frequent measuring allows for spotting trends and placing polarization in a societal context. Both these papers however, focus on political polarization from a partisan perspective and not from a societal perspective. The divide between the Democratic and Republican party is not necessarily the same as a societal divide between groups with different political ideas and does not necessarily reflect political polarization in broader society.

A different approach to measuring polarization is with the use of social media data, which offers a way to have a more frequent or even constant measure. A recent paper by Morini, Pollacci and Rossetti (2020) used data from social media platform Reddit to measure polarization. In this paper, polarization is measured using text classification. Polarization is

this study was in terms of pro- and anti-Trump sentiments. Using reddit pages that display strong pro and anti-Trump sentiments to train the classifier.

Earlier studies on political polarization are using data that limits the frequency to which political polarization can be measured (Klinker & Hapanowicz, 2015; Azzimonti, 2013; Iyyer et al., 2014; Yu et al., 2008), with later studies correcting this by using social media data to look for political polarization online. However, these newer studies are often focused on polarization on specific topics, central entities or from a partisan perspective (Morini et al., 2020; Demszky et al., 2019; Gupta et al., 2020; Rao & Spasojevic, 2016). Examples of studies that are focused on polarization from ideological perspective as opposed to a partisan perspective often don't use social media data or are centered around a certain topic (Farrell, 2015; Melki & Pickering, 2014; Clark, 2009). A measure that looks for political polarization in a broader ideological sense through social media seems to be lacking.

The research proposed in this thesis could offer a way to have a more direct way to measure political polarization online. The word ‘direct’ is interpreted to have two main components. The first component is that political polarization is measured through social media comments posted by regular people, not necessarily on a specific topic. The second component refers to the perspective in which political polarization is viewed (ideologically and not partisan), essentially removing political parties as the intermediary between societal political polarization and the point of measurement. This new measuring approach could offer a way to acquire more direct and much needed insight into political polarization on social media and the nature of their underlying connection. The measures for ideological polarization in a social media environment are expected to reflect the perceived polarizing trend in the US, but are expected to differ depending on the online environment.

2.3 Numerical Text Representation

Social media is a rich data source consisting of ever increasing potential of information. In order to analyze large amounts of text data from social media, tools in the field of machine learning should be looked for. Especially in the domain of NLP (natural language processing). In order to detect political sentiments from text data, text classification seems to be widely used (Rao and Spasojevic (2016); Iyyer et al., (2014); Yu et al., (2008); Gupta et al., 2020)), but for machines to read text, this text first must be represented in a numerical manner. This numerical representation of text contains the information fed to the model. In public political discourse, expressions of sarcasm or irony are often used as a response (Musolff, 2017). Such expressions require a deeper understanding of language like context along with the content (Mehndiratta & Soni, 2019).

A way to incorporate such information into the numerical representation of text is through contextualized word-embeddings. Contextualized word-embeddings assign word representations based on the context in which the word is used, as one word can have multiple meanings based on the way it is used. Initially, this was done using a RNN (recurrent neural network) such as LSTM (long-short term memory) (Hochreiter & Schmidhuber, 1997). The above mentioned paper by Morini, Pollacci and Rossetti (2020) used word embeddings based on LSTM networks to classify comments in a political context.

LSTM forms the basis for ELMo (Embeddings from Language Models), introduced by Peters et al. (2018).

In 2017, Viswani et al. (2017) introduced the Transformer. Transformers deal with long-dependencies better as they are based on the concept of attention and do not need a RNN, which allows transformers to process sequential data like text in parallel. Unlike a RNN, which processes sequential data serially. This reduces training time and memory use, and allows for contextual word-embedding models to be trained on bigger data sets.

Two examples of contextualized word-embedding models based on transformers are OpenAi's GPT-2 (Radford et al.,2019) and Google's BERT (Bi-directional Encoder Representations from Transformers) (Devlin et al., 2018). The main difference is that BERT is bi-directional and GPT-2 is unidirectional (left-to-right context only). Bi-directionality allows BERT to learn context based on its surrounding words and not just the words that follow it. This allows BERT to better understand the meaning and context of a word as opposed to GPT-2. The tradeoff that made BERT able to use bi-directionality is that GPT-2 uses the concept of 'auto-regression', which BERT does not. Auto-regression based models generate tokens in sequence and add this token to the next model input. Because BERT does not use this auto-regression, BERT is able to take into account the context on both sides of the word (bi-directionality).

For BERT, a smaller and computationally less expensive variant exists called distilBERT. DistilBERT was introduced by Sahn et al. (2019) and is 40% smaller than BERT. Despite its smaller size, it was shown to have retained 97% of its language understanding capabilities, while also being 60% faster to train.

2.4 Text Classification

For the purpose of this thesis, the output of this layer should be a probability that a certain comment belongs to the right ideology. Therefore, the classification model should be of a probabilistic nature. The candidate classification models for this task are Logistic Regression and Naive Bayes. Naive Bayes is however not a good estimator (Zhang, 2005), meaning that probabilities are not well calibrated. The probabilities outputted by Logistic Regression are calibrated, meaning that they are interpretable as the probability some instance belongs to a class. Additionally, Logistic Regression has accomplished good results in comment classification tasks for other domains (Indra et al., 2016; Pranckevičius & Marcinkevičius, 2017; Shah et al., 2020). Logistic Regression also showed good results in text classification with BERT text embeddings for medical comments (Roitero et al., 2020).

3. Method Description

This section describes the datasets for model training and for the classification task. The method used to answer the central research question is described using the algorithms and models introduced in the related work section. The evaluation for model performance and classification results is described.

3.1 Datasets

The training dataset is taken from Reddit using pushshift.io. Reddit was chosen because it offers separate forums where people with similar interests or viewpoints gather. In order to look for polarization, the dataset is labeled as the respective central ideologies the comments on the subreddits are coming from. The data will be coming from both the ideological extremes as well as the most predominant conflicting ideologies in the US according to a linear relation between the ideologies. However, obtaining data that is far-right is difficult, as subreddits that entertain these ideologies are often removed and/or banned. Therefore, the conservative dataset is used for both. The training dataset for the dominant ideologies consists of subreddits r/Liberal and r/Conservative. The training dataset containing the extreme opposites is coming from the subreddits r/Communism and r/Conservative.

The individual datasets will be filtered by the total score per comment. A comment's 'score' refers to the balance between up- and downvotes. A comment with 5 upvotes and 3 downvotes has a score of 2 for example. This is done to ensure that the comments are of sufficient quality and to balance the number of comments somewhat. A score of greater than 5 was used for r/Conservative and a score of greater than 0 was used for both r/Liberal and r/Communism. Also, comments that have less than 10 characters are discarded to disregard one word replies.

Table 1
Amount of comments per subreddit after filtering on score and sentence length.

Subreddit	Number of Comments	Score (<)
r/Conservative	322187	5
r/Liberal	33354	0
r/Communism	12315	0

As classification data, a similarly moderated dataset must be used. This in order to ensure that the model is trained on data that is similar in tone as the classification data. This is checked for in the subreddit rules. Also, it should be taken from a relatively large environment which does not have an apparent political tone. Reddit is also used to find such an environment.

The trained model is used to detect polarization on a big political subreddit 'r/Politics' (7.5 million members as of April, 2021). This is a subreddit where only American political news articles are shared and commented upon. Because polarization is defined containing a time factor, data from 2016 and 2019 is used and compared. This subreddit is picked because

it is a rich source of data, it is a relatively large subreddit with many users and is similarly monitored as the training data (according to the rules and regulations formulated on the subreddit). The data for classification is filtered on a minimum score of 5 to ensure quality of the comment and to control the total number of comments needed to be processed. Additionally, comments consisting of under 10 characters are discarded.

Table 2
Amount of comment per year used for classification.

r/Politics	Number of Comments	Score (<)
2016	3.894.148	2
2019	4.717.554	2

3.2 Text representation

In the related work section, the concept of word-embedding and contextualized word-embedding was introduced and the importance of context in political text classification. Therefore, contextual word-embeddings are used to retain as much contextual information as possible in order to answer the research question.
In this thesis, it is opted to make use of a transformer based model, mainly for the computation efficiency they provide. Additionally, BERT seemed to be the best option for this classification task, thanks to its bi-directionally nature to maximize contextual information. To further reduce computationally expensiveness, a smaller version of BERT, distilBERT is used. The distilBERT model is provided by the Transformer library on Hugging Face.

The comments are changed to lowercase and tokenized by a tokenizer especially for BERT. The tokenized comments are all padded to size 512 and masked in order for distilBERT to not interpret these extra tokens. DistilBERT generates embeddings for each word but also for each comment, which is used for classification. Only the embeddings for classifications are kept. This data-preparation is done for both the training and classification dataset.

3.3 Train Classification layer

The imbalance of the dataset is dealt with in part by a combination of over- and under sampling, using SMOTE (Synthetic Minority Over-sampling Technique) (Chawla et al., 2002). SMOTE is an oversampling technique that creates ‘synthetic’ examples rather than using replacements for oversampling. According to Chawla et al., 2002, this technique of oversampling in combination with under sampling showed good results in dealing with imbalanced datasets. An oversampling proportion of 50% and an under sampling proportion of 80% is found to balance the bias and model performance well. The resulting dataset is split into a train and test dataset of proportions 75% and 25%, respectively.

The Logistic Regression algorithm is used as the classification layer. Logistic Regression provides calibrated probabilities as output, which make them interpretable as the probability some comment belongs to a certain class. This Logistic Regression algorithm with SGD (stochastic gradient descent) solver is used to train the text classification layer on top of distilBERT. The SGD solver was opted for to facilitate incremental learning, meaning that the dataset does not have to be loaded in memory all at once and weights are updated incrementally (Manogaran & Lopez, 2016). The word-embeddings are scaled before being used for training the classification layer. The data is fed to the model in batches of 32 and over 5 epochs. The learning rate parameter was tuned by trying different values and comparing the evaluation metrics. It was found that a constant learning rate of 0.0001 performed best. The scikit-learn module (Pedregosa et al., 2011) is used to load this logistic regression algorithm in python.

Both the classification layer for the liberal and conservative and the communism and conservative datasets are trained the same way.

3.4 Classification Task

To assess political polarization, the word- embeddings for the classification dataset are classified by both the model trained on the communism and conservatism and liberalism and conservatism datasets. The outcome of the probabilistic models is a NumPy (Harris et al., 2020) array portraying the probabilities the comment belongs to either the left or right ideology. For the purpose of plotting and inferring from this plot, only the probabilities the comments belong to the right ideology are kept. Due to the complementary nature of probabilities and because there are only 2 classes in the data, these probabilities can be used to plot the ideological density histogram. Before plotting, the probabilities are multiplied by 100 to acquire the ideology score of a comment. This multiplication also provides the calculated variance to be better interpretable. In this plot, every comment that has an ideology score lower than 50 is assumed to belong to the left ideology group and every comment with an ideology score higher or equal to 50 is assumed to belong to the right ideology group.

3.5 Evaluation

The distilBERT model with its final layer is evaluated with a confusion matrix. Using this confusion matrix, the F1 score and the model accuracy is calculated. The F1-score offers an evaluation metric that is useful for imbalanced datasets. The accuracy score can offer additional context on model performance. However, the main model evaluation is assessed based on the resulting confusion matrix. Because the purpose of the model is ultimately reliant on probabilities and not on discrete class labels, the ROC-AUC score is used as an additional evaluation metric. This gives an idea how good the model is in distinguishing between the two labels, and not just how well it predicts the class labels.

The classification results are evaluated on the change in ideological diversion and ideological alignment. These two measures form the basis to assess if there has been political mass polarization visible in the US between the year 2016 and 2019 in the chosen social

media environment. The Ideological divergence value is defined as the change in the distance in the means of the left and right ideological groups between two periods in time.

$$Ideological\ divergence\ value\ = (\ \mu(r)_x\ -\ \mu(l)_x\ )\ -\ (\ \mu(r)_y\ -\ \mu(l)_y\ )$$

Where μ is the mean for either the right (r) or left (l) ideological group for moment in time x or y, where x is the latest time period and y is the earlier time period.

To assess if ideological diversion is significantly different between the years, a t-test is done between the right ideological group for both years and the left ideological group for both time periods.

The ideological alignment value is defined as the change in variance for the right or left ideological group between time periods.

$$Ideological\ alignment\ value\ = \sigma^2(z)_y\ -\ \sigma^2(z)_x$$

Where $\sigma^2$is the variance of z, which is either the left or the right ideological group for time period x or y, where x is the latest period in time and y is the earliest period in time.

Both the ideological diversion value and the ideological alignment value are expressed in a percentage of the change relative to the value of the earliest time period. These scores give context to the values in terms of the relative size of the change in the ideological values. The ideological diversion and ideological alignment score are calculated as follows:

$$Ideological\ diversion\ score = (Ideological\ divergence\ value \div (\mu(r)_y\ -\ \mu(l)_y)) * 100$$

$$Ideological\ alignment\ score\ =\ (\ Ideological\ alignment\ value\ \div \sigma^2(z)_y\ )\ *\ 100$$

The interpretation of the scores for these indicators are compared to the discourse around political polarization in the US. The goal for this comparison is to see to what extent the discourse and political sentiment towards political polarization in US society matches the results found in the social media environment.

4. Results

In this section, the model performance and political polarization indicators for both the model trained on the liberal and conservative and communism and conservative datasets are presented.

4.1 Liberal and Conservative

The logistic regression model performance is evaluated on the confusion matrix, the accuracy-, F1- and ROC AUC- scores. The confusion matrix below gives an overall view of the model performance on the test dataset. The model over-samples the minority class to 50% and under-samples the majority class to 80% of the resulting amount of minority class after oversampling. The learning rate is set to a constant value of 0.0001.

In table 3 the confusion matrix and in table 4 the values for the evaluation metrics are displayed. The confusion matrix shows that the model predicts wrong regularly. Especially the liberal class is predicted with less accuracy. The moderate model performance is also reflected by the evaluation metric values.

Table 3
Confusion matrix for the model trained on the liberal and conservative dataset.

Actual \ Predicted	Liberal (0)	Conservative(1)
Liberal (0)	**23.377**	**16.847**
Conservative (1)	**13.206**	**37.185**

Table 4
Values for the evaluation metrics for the model trained on the liberal and conservative dataset.

Evaluation metric	Value
Accuracy-Score	**0.668**
F1-Score	**0.712**
ROC AUC-Score	**0.726**

The overall visualization of the results of the model on the classification dataset for both 2016 and 2019 are displayed in figure 1. Some more detailed values used to calculate the political polarization indicators are presented in table 5 and table 6.

Figure 1

Histograms displaying the distribution of the ideology scores of the comments in the classification dataset with the model trained on the liberal/ conservatism dataset.

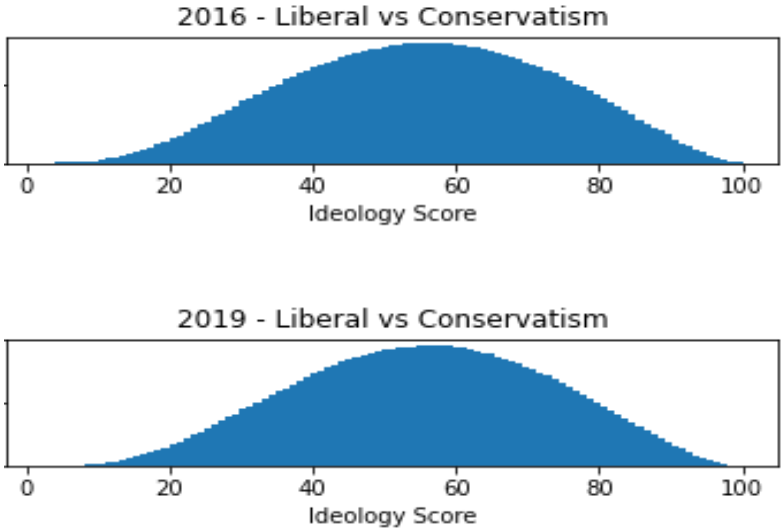

Table 5

The values calculated from the model output that are used to calculate the value for the ideological divergence indicator

Means	Value
$\mu(r)x$	66.600
$\mu(l)x$	37.118
$\mu(r)y$	67.032
$\mu(l)y$	36.733

Table 6

The values calculated from the model output that are used to calculate the value of for the ideological alignment indicator.

Variance (σ^2)	z = Liberal	z = Conservative
x	78.689	124.267
y	81.204	129.296

Table 7
The calculated values and scores for the political polarization indicators on the liberal and conservative dataset.

Indicator	*Value*	*Score*
Ideological divergence	- 0.817	**- 2.70**
Ideological alignment	2.515 (p = 0.00)	**3.10**
Ideological alignment	5.029 (p = 0.00)	**3.89**

In Table 7, the values and scores for the political polarization indicators are presented. The scores indicate only small changes relative to the initial values for the mean and variance. The scores indicate that the means of the two groups are getting closer, but that the variance of the two groups is less.

4.2 Communism and Conservative

The model has been optimized using different parameters for the SMOTE oversampling and under sampling. The proportions chosen were to over-sample the minority class to 50% of the majority class, and to under sample the majority class that the minority class is 80% of the majority class. The learning rate is a constant value of 0.0001. These values produced the best balanced performance according to the confusion matrix and the evaluation metrics.

As seen in table 8, the confusion matrix predicts the right labels reliably well. However, the model is better at predicting the conservative class as opposed to the communism class. The evaluation metric shown in table 9 shows that the model performs well on the test set, both in the discrete evaluation measures as the probabilistic measure.

Table 8
Confusion matrix for the model trained on the communism and conservative dataset.

Actual \ Predicted	Communism (0)	Conservative (1)
Communism (0)	**32.683**	**7.546**
Conservative (1)	**5.596**	**44.790**

Table 9
Values for the evaluation metrics for the model trained on the communism and conservative dataset.

Evaluation metric	Value
Accuracy-Score	**0.872**
F1-Score	**0.855**
ROC AUC-Score	**0.931**

In figure 2, the frequency of the probabilities predicted by the model on the classification dataset for both the year 2016 and 2019 are displayed. In table 10 and table 11, the values that are used to calculate the political polarization indicators are displayed.

Figure 2

Histograms displaying the distribution of the ideology scores of the comments in the classification dataset with the model trained on the communism and conservatism dataset.

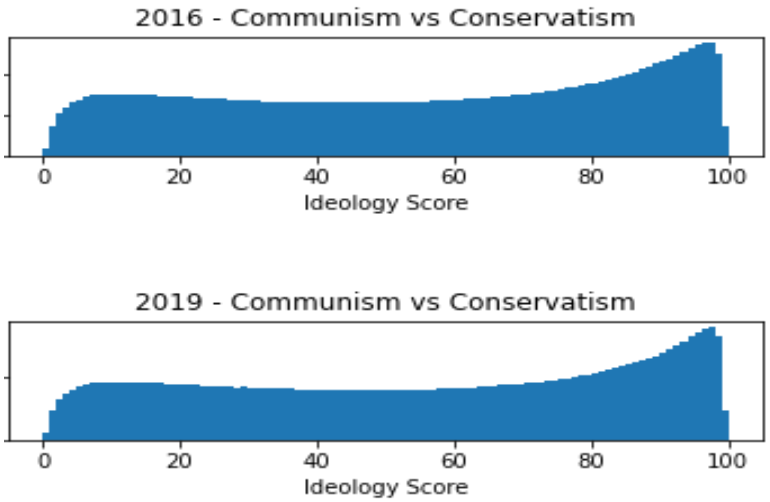

Table 10

The values calculated from the model output that are used to calculate the value for the ideological divergence indicator.

Means (μ)	Value
$\mu(r)x$	78.334
$\mu(l)x$	24.717
$\mu(r)y$	78.170
$\mu(l)y$	24.750

Table 11

The values calculated from the model output that are used to calculate the value of for the ideological alignment indicator.

Variance (σ^2)	z = Communism	z = Conservative
x	192.409	216.913
y	192.039	213.753

In Table 12, the values for the political polarization indicators are displayed. The scores for ideological divergence and alignment indicate only very marginal changes. Also the negative scores show a depolarizing trend for the political alignment factor.

Table 12
The calculated values and scores for the political polarization indicators on the communism and conservative dataset.

Indicator	*Value*	*Score*
Ideological divergence	0.198	**0.37**
Ideological alignment	- 0.370 (p = 0.02)	**- 0.19**
Ideological alignment	- 3.160 (p = 0.00)	**- 1.48**

5. Discussion

In this section the performance of the models trained on the liberal and conservative dataset and communism and conservative dataset is interpreted. Also, the results of both models are interpreted. Results for both models are compared, conclusions are drawn and some study limitations are described. Lastly, the possibilities and directions for future work are explained.

5.1 Liberal and Conservatism

The performance or the trained model on the liberal and conservative dataset seems to be moderate, as the values for the evaluation metrics are not too convincing. The confusion matrix shows that the model predicts conservative comments more accurately than it does liberal comments. This might be due to the fact that the model is trained on data that has, despite the over- and under sampling, more data labeled conservative than liberal, resulting in a bias towards the conservative ideology. Therefore, conclusions drawn from the classification results of this model should be with the model performance in mind.

The distribution of comment ideology scores shows a single peak, indicating no clear ideological groups in the data. The plot does seem to show the classification data in both 2016 and 2019 to be leaning towards the conservative ideology, but only slightly. Taking the bias of the model towards the conservative ideology into consideration, this perceived lean towards conservatism is contested.

The resulting values for ideological divergence and ideological alignment do not suggest a politically polarizing trend between the liberal and conservative ideologies between 2016 and 2019 in the classification dataset. The negative value for the ideological diversion score points to the left mean and the right mean to be coming closer to each other. In this distribution without two groups, this indicates a steeper peak for the distribution. The negative values for ideological alignment scores reinforce this steepening trend. However, the scores are small, meaning that this effect is minor.

5.2 Communism and Conservatism

Model performance for the communism and conservatism perspective is rather good, as indicated by the evaluation metrics and the confusion matrix. All three evaluation metrics suggest that the model predicts well. The high ROC AUC-score indicates that the model does well in distinguishing between comment from the communism subreddit and from the conservatism subreddit. However, the confusion matrix offers more detail and shows that the model predicts the conservative class more often and accurately. This can be seen as evidence that the model has a bias towards the Conservative class.

The overall distribution of the probability values show two distinct groups in the data plot, with the highest peak on the Conservative side of the distribution. This would suggest that the data is more heavily right leaning. However, due to the bias of the classification model, this is probably inflated. Additionally, these two peaks are on the extremes of the distribution, suggesting the presence of some polarity in the data.

The positive score for ideological diversion indicates the means of the two groups to be polarizing, however the negative values for the ideological alignment scores are suggesting that these two groups are getting more dispersed. However, both the ideological divergence score and the ideological alignment score are relatively small, meaning minimal change in the two factors that indicate political polarization. The values of the political polarization indicators thus do not show clear signs of political polarization in the classification dataset between 2016 and 2019.

5.3 Conclusion and Limitations

Comparing the results of both models, the model trained on the more contrasting ideologies seems to be better at quantifying and measuring political polarization in a social media environment. Mostly because the model trained on the less extreme ideologies does not produce convincing values for the evaluation metrics and the confusion matrix and the more extreme model does. The comparison of both plots show a single, central group for the less extreme ideologies and more indication of two distinct groups in the more extreme comparison. The less extreme model is found to be less certain in its prediction between the two classes compared to the model trained on the more contrasting ideologies. This big difference in the two histograms produced by the models is not expected, as both histograms are produced using the same data. The expected outcome would be that both plots would be similarly distributed, except on a different ideological scale.

To assess the accuracy of the findings, they should be compared to the existing narrative of political polarization in the same time frame. In the Introduction of this thesis, the popular discourse around political polarization was introduced. The narrative surrounding political polarization in the last few years is not mirrored by the data. Additionally, the polarization trend reported by the political polarization survey by Pew Research Center (2017) is not continued in the data. Therefore, the results of this thesis offer no evidence that political polarization can be measured in a social media environment.

Following this, the research question of this thesis cannot be answered concludingly. The results of this study do not offer evidence that political polarization can be directly measured in a social media environment. However, this study has found a model that predicts the difference between comments with a communism nature and with a conservative nature well.

This study was subject to some limitations. Time was a limitation and limited the extensiveness of the study. Also, the method of data collection limited the data that could be obtained. Additionally, this also limited the kind of data that could be acquired, both in terms of number of comments and in terms of time periods. Reddit has removed a lot of data that can be useful to train the model on far-right comments.

5.4 Future work

This study was an initial attempt to measure non-topic related political polarization in a social media environment from an ideological perspective. It attempted to create a more generalized measurement that is not limited by a specific topic or views political polarization

from a partisan perspective. This study provides initial results and insight into how such a measure could be approached, and what approach does not seem to work. Such initial insight could be valuable for future work in the attempt to understand political polarization from an ideological perspective and the role of social media in this. This understanding could ultimately play an important role in combating the negative consequences of political polarization.

Future work should consider using a classification model that is trained on clearly contrasting political ideologies, making the model better in distinguishing between different ideologies. The results of such a model make ideological groups more distinct and visible. Additionally, future attempts would benefit from more clearly distinct labeled comments to serve as ground truth to improve distinction further. Labeling all the comments by their subreddit name might not be the best way to do this. Anyone can comment on any subreddit, possibly resulting in mislabeled comments. When using more extremely contrasting ideologies, a challenge would be to obtain data that is far-right. Although this should be preferred in future work.

Additionally, a balanced dataset to reduce the bias as much as possible is desired. This improves the certainty and precision of the inferences from the results. Both the model bias and the degree to which the model can distinguish between the two labels seem to be important for the results to be reliable and interpretable. To answer the research question more conclusively, different social media environments should be examined with the same model and compared with the expected output. Testing if the trained model generalizes to more online environments is valuable for future endeavors towards ideological polarization measurements.

Although this thesis does not offer any evidence that political polarization can be measured in an online environment from an ideological perspective, it marks an initial attempt to create a general and direct measure for political polarization in social media. Future work can learn from the results of this study and use the insight into future attempts to measure political polarization in terms of ideologies.

References

- Abramowitz, A., & Saunders, K. (2005). Why can't we all just get along? The reality of a polarized America. In *The Forum (Vol. 3, No. 2, pp. 1-22). De Gruyter*.
- Azzimonti, M. (2013). The Political Polarization Index. *Working Paper Number NO. 13-41, Research Department, Federal Reserve Bank of Philadelphia*
- Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. F.& Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences, 115(37)*, 9216-9221.
- Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political analysis, 23(1), 76-91*.
- Brownstein, R. (2021, January 19). Trump leaves America at its most divided since the Civil War. *CNN.* Retrieved from
- Carothers, T., & O'Donohue, A. (Eds.). (2019). Democracies divided: The global challenge of political polarization. *Brookings Institution Press.*
- Cassese, E., Kalmoe, N. P., Mason, L. & Theodoridis, A. (2021, January 21). Pro-Trump Capitol riot violence underscores bipartisan danger of dehumanizing language. *NBC NEWS.* Retrieved from https://www.nbcnews.com/
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research, 16, 321-357.*
- Clark, T. S. (2009). Measuring ideological polarization on the united states supreme court. *Political Research Quarterly, 62(1), 146-157.*
- Demszky, D., Garg, N., Voigt, R., Zou, J., Gentzkow, M., Shapiro, J., & Jurafsky, D. (2019). Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. *arXiv preprint arXiv:1904.01596.*
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.
- DiMaggio, P., Evans, J., & Bryson, B. (1996). Have American's social attitudes become more polarized?. *American journal of Sociology, 102(3)*, 690-755.
- Farrell, J. (2016). Corporate funding and ideological polarization about climate change. *Proceedings of the National Academy of Sciences, 113(1), 92-97.*
- Fiorina, M. P., Abrams, S. J., & Pope, J. C. (2005). Culture war. The myth of a polarized America. *New York: Pearson Longman*.
- Garimella, K., De Francisci Morales, G., Gionis, A., & Mathioudakis, M. (2018, April). Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 World Wide Web Conference (pp. 913-922).*
- Gupta, S., Bolden, S., Kachhadia, J., Korsunska, A., & Stromer-Galley, J. (2020, October). PoliBERT: Classifying political social media messages with BERT. I*n Social, Cultural and Behavioral Modeling (SBP-BRIMS 2020) conference. Washington, DC.*

- Harris, C.R., Millman, K.J., van der Walt, S.J. et al. Array programming with NumPy. *Nature 585, 357–362 (2020).* DOI: 0.1038/s41586-020-2649-2.
- Hetherington, M. J. (2009). Putting polarization in perspective. *British Journal of Political Science, 413-448*.
- Heywood, A. (2017). Political ideologies: An introduction. *Macmillan International Higher Education p. 15-16.*
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation, 9(8), 1735-1780.*
- Hong, S., & Kim, S. H. (2016). Political polarization on twitter: Implications for the use of social media in digital governments. *Government Information Quarterly, 33(4), 777-782.*
- Howard, P. N., Ganesh, B., Liotsiou, D., Kelly, J., & François, C. (2019). The IRA, social media and political polarization in the United States, 2012-2018. *Oxford: University of Oxford.*
- Indra, S. T., Wikarsa, L., & Turang, R. (2016, October). Using logistic regression method to classify tweets into the selected topics. *In 2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS) (pp. 385-390). IEEE*
- Iyyer, M., Enns, P., Boyd-Graber, J., & Resnik, P. (2014, June). Political ideology detection using recursive neural networks. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1113-1122).*
- Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. Journal of documentation.
- Klinkner, P. A., & Hapanowicz, A. (2005, July). Red and blue Déja Vu: measuring political polarization in the 2004 election. In *The Forum (Vol. 3, No. 2). De Gruyter.*
- Körösényi, A. (2013). Political polarization and its consequences on democratic accountability. *Corvinus Journal of Sociology and Social Policy, 4(2*), 3-30.
- Lelkes, Y. (2016). Mass polarization: Manifestations and measurements. *Public Opinion Quarterly, 80(S1), 392-410.*
- Manogaran, G., & Lopez, D. (2018). Health data analytics using scalable logistic regression with stochastic gradient descent. *International Journal of Advanced Intelligence Paradigms, 10(1/2), 118. doi:10.1504/ijaip.2018.089494*
- Mehndiratta, P., & Soni, D. (2019). Identification of sarcasm using word embeddings and hyperparameters tuning. *Journal of Discrete Mathematical Sciences and Cryptography, 22(4), 465-489.*
- Melki, M., & Pickering, A. (2014). Ideological polarization and the media. *Economics Letters, 125(1), 36-39*.
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Morini, V., Pollacci, L., & Rossetti, G. (2020) Capturing Political Polarization of Reddit Submissions in the Trump Era. *University of Pisa, Italy.*
- Musolff, A. (2017). Metaphor, irony and sarcasm in public discourse. *Journal of Pragmatics, 109, 95-104.*

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research, 12, 2825-2830*.
- Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543)
- Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *NAACL*.
- Pew Research Center. (2017, oktober). The Partisan Divide on Political Values Grows Even Wider. https://www.pewresearch.org/politics/2017/10/05/the-partisan-divide-on-political-values-grows-even-wider/
- Pranckevičius, T., & Marcinkevičius, V. (2017). Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Baltic Journal of Modern Computing, 5(2), 221*.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog, 1(8)*, 9.
- Rao, A., & Spasojevic, N. (2016). Actionable and political text classification using word embeddings and lstm. *arXiv preprint arXiv:1607.02501*.
- Roitero, K., Bozzato, C., Della Mea, V., Mizzaro, S., & Serra, G. (2020, April). Twitter goes to the Doctor: Detecting Medical Tweets using Machine Learning and BERT. *In Proceedings of the International Workshop on Semantic Indexing and Information Retrieval for Health from Heterogeneous Content Types and Languages.*
- Saad, L. (2021, January 11). Americans' Political Ideology Held Steady in 2020. *GALLUP.* Retrieved from https://news.gallup.com/poll/328367/americans-political-ideology-held-steady-2020.aspx
- Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108.*
- Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification. *Augmented Human Research, 5(1). doi:10.1007/s41133-020-00032-0*
- Smith, S. (2021, January 10). Inequality, racism and polarisation ravaged US democracy. Then came Trump. *The Guardian.* Retrieved from https://www.theguardian.com/
- Sunstein, C. R. (2018). # Republic: Divided democracy in the age of social media. *Princeton University Press*.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems, pages 6000–6010*.
- Yu, B., Kaufmann, S., & Diermeier, D. (2008). Classifying party affiliation from political speech. *Journal of Information Technology & Politics, 5(1), 33-48.*
- Zhang, H. (2005). Exploring conditions for the optimality of naive Bayes. *International Journal of Pattern Recognition and Artificial Intelligence, 19(02), 183-198.*