UNIVERSITY OF GOTHENBURG

DEPARTMENT OF PHILOSOPHY, LINGUISTICS AND THEORY OF SCIENCE

GO BACK TO /R/CONSPIRACY: AN EXPLORATION OF METHODS FOR THE AUTOMATIC DETECTION OF AFFECTIVE POLARIZATION ON REDDIT

Klara Båstedt

Master's Thesis:	30 credits
Programme:	Master's Programme in Language Technology
Level:	Advanced level
Semester and year:	Spring, 2022
Supervisor	Asad Sayeed
Examiner	Rickard Johansson
Keywords	affective polarization, covid-19, social media, Reddit, NLP

Abstract

Affective polarization – the tendency to hold negative attitudes towards an out-group and biased, positive attitudes towards an in-group – is a hot topic in research and public debate. There are concerns that news media's tendency to focus on political conflict rather than issues is causing polarization to increase, but researchers lack methods to automatically asses levels of polarization in online debates and correlate them with news articles. This study examines the appropriateness of using Reddit Karma, word embeddings and existing NLP tools for automatic detection of affective polarization in discussions on Reddit. To achieve this, we collect and manually annotate Reddit discussions for expressions of affective polarization and fit multiple logistic regression models on the discussion features and metadata. We find a strong correlation between the probability to encounter expressions of affective polarization in the data and both word embeddings and the confidence scores of toxicity detection. We also find that patterns in the comment votes are good predictors of disagreement in the discussions. Moreover, we present a data set of Reddit-discussions about topics related to the covid-19 pandemic which can be used in further attempts to automatically detect affective polarization in interactive discourse on social media.

Preface

First and foremost, I would like to thank my supervisor Asad Sayeed for his support and encouragement, inspiring theoretical discussions and expertise about statistical modelling as well as online forums. In exploratory work of this kind it has been especially meaningful with the guidance of someone who always knows how to proceed.

I also want to thank Gregor Rettenegger for sharing important insights from the field of communication science as well as crucial advice on the annotation process. Moreover, I want to thank the GRIPES project for providing me with such a challenging and interesting thesis topic as well as supporting me with the annotation.

Last but not least, I would like to thank my teachers and classmates from the MLT programme for making the past two years memorable and enlightening.

Contents

1 Introduction

During the last years, affective polarization has been a hot topic in public debate as well as research. Affective polarization is commonly defined as the tendency to hold negative attitudes towards members of an out-group (e.g., supporters of an opposing political party) and positive, biased attitudes towards members of one's in-group (Iyengar et al., 2019). Typical for affective polarization is to dislike and distrust members of the out-group and ascribe them undesirable traits such as selfishness and hypocrisy. Studies indicate that affective polarization is on the rise in western democracies (Reiljan, 2019) and there are concerns that the increasing hostility towards people with opposing views constitutes a threat to democratic systems and political trust (Druckman & Levendusky, 2019). Researchers suggest a range of theories on what causes polarization to rise and social media as well as reporting in news media are frequently accused of driving this development (Nordbrandt, 2021; Kim & Zhou, 2020). However, the hypotheses have in common the difficulty to establish a relationship between possible causes and increased levels of affective polarization.

The recent progress in NLP enables automatized methods to analyze vast amounts of text which could facilitate research on what causes affective polarization to increase. Automated detection of expressions of affective polarization in online discussions would assist quantification of polarization levels. The detected polarization levels can in turn be correlated with reporting in news media or other societal phenomena and events to help establish a causal relationship. Multiple tools are available for classification of social media comments and posts as toxic, uncivil or hateful (e.g., Jigsaw, 2017; Davidson et al., 2020). While expressions of affective polarization on social media are often perceived as uncivil or hateful, this is not a sufficient condition. To the best of our knowledge, there are no existing methods that automatically detect expressions of affective polarization in online discussions.

The development of automated detection of expressions of affective polarization in online discussion comes with a number of challenges. Firstly, affective polarization is a more complex concept than, for instance, incivility since it is dependent on political and social context. It is also dependent on the group belonging of the author, as perceived by themselves and by others, as well as the target of any attacks and accusations. Secondly, it is typically not enough to analyze comments or posts in isolation since expressions of affective polarization may take different form depending on the presence or absence of the out-group in the discussion. Polarized attitudes include both negative attitudes towards out-groups and positive attitudes towards one's in-group, and the uttering of such attitudes are heavily dependent on the conversational context. With this in mind, it is often necessary to study the interaction between the participants of a discussion and establish their group belonging in order to understand to what degree the discussion is polarized.

The main challenge of designing methods to use for research purposes lies in performing classification that is true to theory about affective polarization. At the same time, the classifier must be able to handle large amounts of text and the immense variation that is typical of interactive discourse. This challenge cannot be overcome within the scope of this study. Instead, we conduct an exploration of existing methods and tools as a step towards automatic detection of affective polarization in discussions online. With this study, we hope to gain valuable insights on how affective polarization in online discussions can be detected and quantified. To further define our research aims, we formulate the research questions below.

(i) Can existing NLP tools predict affective polarization in discussions on Reddit?

(ii) Can comment votes predict affective polarization?

(iii) Can word embeddings predict affective polarization?

(iv) Do the results differ depending on the group belonging of the participants of a discussion?

In order to answer the research questions above, we collect and manually annotate Reddit discussions about the covid-19 pandemic for concepts related to affective polarization. We perform analysis of the discussions with a range of existing NLP tools and methods and fit logistic regression models to examine the correlation between the results of the applied NLP tools and the annotation. In the statistical analysis, we also include information about the votes obtained by the comments in the conversations to examine whether such meta data can improve polarization detection. According to theory on affective polarization, we attempt to establish presence of out-groups members in the conversations and investigate if it is reflected in the results of the automated analysis.

2 Background and Related Work

In this chapter, we account for theory and research on polarization that is of importance to our study. The first section outlines different aspects of polarization and possible explanations for its increase. The following sections explain affective polarization in greater detail, how it is traditionally measured and how it relates to theories about intergroup conflict and identity performance. We proceed with a summary of research on polarization in social media and account for the characteristics of Reddit in this context. Moreover, we describe previous attempts to automatically detect and quantify polarization in online discussions.

Research on polarization is of interest in multiple scientific fields such as sociology and psychology as well as political and communication science. As a result, the term polarization embodies a multitude of concepts which include political polarization, social polarization and affective polarization. Political polarization refers to the tendency of political views to move away from each other and towards the extreme (Fiorina & Abrams, 2008), social polarization is commonly defined as income inequality and social division (Hamnett, 2011), and affective polarization is used to describe negative feelings and attitudes towards out-groups and biased positive emotions towards the in-group (Iyengar et al., 2019). The term polarization in these senses is used to refer both to the state and the process of polarization. Throughout this report, we will focus on and refer to affective polarization when using terms such as ”polarized”.

Several studies suggest that polarization is increasing in the U.S. and other democracies (e.g., Reiljan, 2019; Iyengar et al., 2012). This development has negative consequences for political trust and interpersonal relations (Druckman & Levendusky, 2019) which can be considered a threat to democracy and public debate. An example of this is that, in the context of the U.S., people typically dislike, distrust and avoid interaction with supporters of the other party (Iyengar et al., 2012). An important limitation of polarization research is that a majority concentrates on the U.S. and their two-party system due to the facility of dividing the subjects in supporters of either the Democratic Party or the Republican Party. The study of polarization in the multi-party systems of many European countries is more complicated, since there are theoretical challenges when it comes to dividing the subjects into in-groups and out-groups (Wagner, 2021).

As stated in the introduction, the concerns that social media usage is causing polarization to rise has gained much popular attention. The increased personalization of social media platforms has led to suspicions that the algorithms it relies on create information bubbles that expose users mainly to opinions of like-minded and to news articles that confirm these views – a setting that could facilitate the spread of fake news and the increase of polarization (Pariser, 2011). Nevertheless, other studies question the impact and even the existence of such bubbles (e.g., Dubois & Blank, 2018). Additional disputed explanations of the increasing polarization are confirmation bias and selective exposure which allegedly cause people to selectively expose themselves to information that matches their believes (e.g., Stroud, 2008; Dahlgren, 2020).

Another possible cause of the increasing levels of polarization is reporting in news media and political news sources. News media's tendency to focus on conflicts between political parties instead of the issues themselves affect the psychological processes that can lead to polarization (Kim & Zhou, 2020). This is an example of news framing which can be described as the narrative of a news article, such as the exclusion or emphasis on certain aspects of the story, which in turn can influence the reader (Lecheler, 2019). Framing connects and gives meaning to the circumstances described in an article and can indirectly imply e.g., the core of an issue or possible solutions (Gamson & Modigliani, 1994). Wojcieszak et al. (2021) highlight that while partisan news have an effect on the increasing polarization, we lack knowledge about how news media content influences attitudes.

In contrast to the concerns about increasing levels of polarization, Druckman & Levendusky (2019) argue that the effects of affective polarization in the U.S. are exaggerated since their study finds that a large majority is still comfortable being friends or neighbours with people from the other party. They also find that U.S.

citizen particularly dislike the political elite of the other party, as opposed to their supporters. Along these lines, Siegel et al. (2018) suggest that hate-speech did not increase on Twitter during or after Trump's election campaign and Boxell et al. (2020) find that affective polarization did not increase during the covid-19 pandemic, perhaps despite the common conception.

2.1 Affective Polarization, Intergroup Conflict and Identity Performance

A more detailed account of affective polarization and how it is related to theories about social identity and intergroup conflict is given in the section below, given that affective polarization is the type of polarization that the experiments conducted in this study revolve around. In addition, we summarize a few studies that investigate affective polarization in the context of the covid-19 pandemic.

As was mentioned in the previous section, the term affective polarization describes the tendency to hold positive feelings towards one's in-group and negative feelings towards an out-group (Iyengar et al., 2019). An example of this is the study carried out by Iyengar et al. (2012) which shows that U.S. citizen judge supporters of the opposing party to be less intelligent and more selfish than supporters of their own party and that this tendency is increasing. Besides, the authors suggest that this tendency is not explainable by ideological differences alone and that affective polarization is a more suitable term for this phenomenon than political or ideological polarization.

Levels of affective polarization in society are traditionally measured through surveys where the subjects are asked to rate their attitudes towards an out-group (Wojcieszak et al., 2021; Iyengar et al., 2012). Such surveys commonly include feeling thermometers, questions to measure social distance and trait ratings. Feeling thermometers aim to quantify the nature and intensity of attitudes towards the in- and out-group. Social distance can be estimated through questions about how comfortable the subject is to, for instance, marry or be friends with someone from the in- and out-group. Trait ratings let the subjects associate traits from a selection of positive as well as negative attributes with members of the in- and out-group. Most importantly, surveys of this kind include questions that serve to establish group belonging of the subjects.

Despite the advantages offered by traditional surveys in terms of detail and depth, polarized attitudes can also be revealed through behaviour. Iyengar et al. (2019) highlight the need to study behaviour as well in order to understand to what degree polarized attitudes influence behaviour. Behaviour that can disclose polarized attitudes include political choices, social patterns and verbal expression (Yarchi et al., 2020), the latter being the subject of study in this thesis. We use the terms *in-group praise* and *out-group derogation*, coined by Wojcieszak et al. (2021), to refer to the voicing of positive attitudes towards an in-group and the expression of negative attitudes towards an out-group. While out-group derogation is often the focus of the public discussion about polarization, the literature suggests that in-group praise is an equally important aspect of affective polarization.

It is difficult to explore the mechanisms behind affective polarization without looking into theory about intergroup behaviour and social identity. According to Tajfel & Turner (1979), the term intergroup behaviour can be applied to individuals that behave in a certain way towards others based on the recognition that they are members of different groups. In other words, in- and out-groups do not need to be based on political ideology or partisanship to trigger intergroup behaviour, although affective polarization is notably stronger in political settings than in nonpolitical settings (Rudolph & Hetherington, 2021). According to the reasoning of Tajfel & Turner (1979), it is sufficient that the individuals involved in an online discussion are considered members of a group by themselves and/or by others to allow intergroup behaviour and in-group favoritism to occur.

In this regard, the concepts of intergroup behaviour are also related to theories about social identity. Social identity performance is claimed to have two main functions which are to express group belonging and increase status within the group, as well as to influence observers (Klein et al., 2007). According to Wojcieszak et al. (2021), both in-group praise and out-group derogation serve to distance an individual from their opponents and can be considered strategies for maintaining a positive in-group identity. In other words, behaviour such as in-group praise and out-group derogation can be considered identity-driven. In fact, Iyengar et al. (2019) suggest that the reinforcement of social identities is a central component of affective polarization.

Having discussed important aspects of polarization, we can argue that the discussions surrounding the covid-19 pandemic offer an interesting opportunity for computational research on polarization. The pandemic was a dominant topic in the news and discussions online for an extended period of time, which facilitates the study of the effects of news media's reporting on attitudes. During the often heated discussions, we have seen a division of people into groups that are for or against policies related to the pandemic such as vaccines and lockdowns. Although these groups did not exist before the pandemic, they rely on previous intergroup conflict to some extent. The politicization of issues related to the pandemic is exemplified in the work undertaken by Fridman et al. (2021), which shows that attitudes towards vaccines in the U.S. have remained stable among Democrats, but declined among Republicans during the pandemic. Similarly, the intentions to wear a face mask are greater among women in the US than among men (Capraro & Barcelo, 2020). In summary, we can expect to find behaviour related to affective polarization in discussions about covid-19 on social media between social groups that are in favor or against corona measures.

2.2 Affective Polarization on Social Media

A more detailed account of how affective polarization can be analyzed on social media in general, and on Reddit in particular, is given in the following sections. While it is common to discuss polarization on social media as a unified phenomenon, the characteristics of the discussions and the way polarization can be observed in them differ across social media platforms (Yarchi et al., 2020). In the literature, the users that the platforms attract and the affordances that the platforms offer are described as major influences on the varying discussion dynamics. The purpose of the platforms differ and consequently attract users with different goals, interests and personalities (Nordbrandt, 2021) and as a result, discussion dynamics and norms of conduct are slightly different across platforms (Yarchi et al., 2020). Social media that users utilize with the main purpose of maintaining offline relationships with family and friends encourage behaviour that is different from, for instance, anonymous forums with discussion as their main purpose (Nordbrandt, 2021).

The term affordance has come to be used in this context to refer to the possibilities and limitations offered by social media platforms (Nagy & Neff, 2015) in terms of how users can interact with each other and the way in which the content is presented to the user. Features such as anonymity and how the user is allowed to share, post, comment and follow have consequences for the types of discussions that emerge (Yarchi et al., 2020). For example, politeness is lower on more anonymous social media platforms compared to platforms where users typically use their offline identity (Halpern & Gibbs, 2013). With this in mind, it is necessary to consider the cross-platform differences in terms of affordances and users when studying affective polarization on social media. However, research on polarization in social media and social media dynamics in general, tends to focus on Twitter data (Yarchi et al., 2020) which may constitute a significant limitation.

Returning briefly to the challenge of finding a causal relationship between social media usage and affective polarization, there seems to be support for the claim that emotionally intense posts gain visibility more easily than more nuanced content (e.g., Brady et al., 2017). While there are aspects of social media platforms that may facilitate certain behaviour, there is support for the hypothesis that the causal relationship is reversed and that it is the level of affective polarization that leads to an increased use of social media. The findings in Nordbrandt (2021) suggest that individuals who feel strongly about a matter are especially appealed by social media and the opportunities they offer to express oneself and convince others. In consideration of individual and cross-platform differences, it is important to keep in mind that discussions on social media may not reflect discussions offline and that the quantity of expressions of polarization observable in them does not necessarily correlate with the levels of polarized attitudes in society in general.

2.3 Discussions on Reddit

Having stated that cross-platform differences affect online discussions and the expressions of affective polarization that can be detected in them, this section describes the affordances of Reddit and why we believe it to be a suitable social media platform for our research aims. Reddit is a popular discussion forum in several countries around the world and in January 2021 the platform had over 50 million active users each day, 100 thousand active communities and more than 13 billion posts and comments (Reddit, 2022a). Users on Reddit are anonymous and anyone can create posts to share text, links, images or videos. Reddit users can also comment on other users' posts and comments as well as upvote and downvote them.

In terms of affordances, the content on Reddit is organized in subreddits which function as communities that gather users with certain interests. For example, subreddits can gather users from certain geographical areas, with certain political ideologies, with interest in discussing certain topics or with the intention to share and watch funny cat videos. Reddit users can choose to follow subreddits of interest to them and get notified when new posts are made. However, anyone is free to post and comment regardless of the subreddit and, as a result, interaction across subreddits is common (Morales et al., 2021). Another important aspect of Reddit which may be of importance to our study is that anyone, even non-users, can access and read the discussion threads. Consequently, Reddit users are aware that people at a different time and place may read or reply to their contributions which reasonably affects how they express themselves.

Regarding the way in which the content is presented to Reddit users, the number of votes and the ratio between upvotes and downvotes on posts and comments are displayed on the site. The possibility to upvote and downvote is a way for users to engage in discussions without posting or commenting themselves and to see how other users react to the content. Besides, the voting feature affects the visibility of posts in the feed. The Reddit feed has a feature that allows visitors to sort posts by ”hot” or ”top” posts, but visitors can also browse posts by topic or with the search function. The visibility of individual comments in the discussion threads are also affected by their number of upvotes and downvotes. It is rarely possible to display all comments in a discussion because of its threaded structure which may consist of several thousand contributions.

Reddit discussions are chosen for this study since the main purpose of the platform is to encourage discussion. Because of the anonymity of the forum and the subreddit communities, we expect to find both out-group derogation an in-group praise in the discussions. Although, Morales et al. (2021) find that interaction between opposing groups is in fact common on Reddit. Regarding the discussion dynamics on Reddit, subreddits differ slightly in terms of characteristics and norms of conduct due to the broad range of topics and types of discussions that emerge on the forum. Subreddits can even have explicit rules about what content is allowed to be posted and edited and apply moderation. Reddit has been accused of being dominated by ”geek masculinity” that silences the voices of marginalized groups, despite their presence in the discussions (Massanari, 2015). Additionally, in our collected and annotated discussions, several users imply that the typical Reddit user is left-leaning and liberal, but this is not established in the literature.

2.4 Automatic Detection of Polarization on Social Media

The following section will describe previous computational studies on polarization in discussions on social media. As was pointed out in the introduction to this report, there are major challenges associated with computational detection and quantification of affective polarization in text. The main challenges consist of the difficulty to establish group belonging of the participants and the fact that the concept of affective polarization is highly complex and dependent on the societal and political context.

As previously stated, traditional research on polarization acquires information about individuals' group belonging through questionnaires, but computational research requires other strategies to establish group belonging of the participants of a discussion. An example of such strategies applied to Reddit discussions is the study of Chipidza (2021), who utilizes the explicit political ideology of certain subreddits to establish group belonging in his study on how toxicity levels in news articles affect their spread within and across the groups. Another example is Alsinet et al. (2021) who divide the discussion participants into two groups depending on whether their comments share the sentiment of the post title or not and proceed to measure negative sentiment in the interaction between members of the two groups. These two research methods are designed so that no manual annotation of the comments is needed which is a major practical advantage.

Computational research of affective polarization in online discussions has mostly been limited to automated quantification and comparison of toxicity levels. A drawback of this approach is that disagreement and disrespectful comments do not necessarily imply affective polarization. Besides, affective polarization consists of praising of the in-group as well as hostility towards out-groups and even out-group derogation does not occur exclusively in hostile confrontations. In order to perform automated and theory-sensitive detection of affective polarization in online discussions, large data sets with annotation grounded in theory are often necessary. Such data sets are not easily available as they are expensive to produce.

Although advanced language models facilitate automatic analysis of text data, humans have a major advantage over machines when it comes to understanding of theoretical concepts (Baden et al., 2020). To benefit from the assets of humans as well as machines, some researchers advocate for hybrid methods to detect abstract concepts in text (Baden et al., 2020; Yarchi et al., 2020). The study of Yarchi et al. (2020) examines interactional, positional and affective polarization on three social media platforms combining computational and manual analysis of political discussions. Despite the benefits of hybrid approaches, Baden et al. (2020) state that the challenge to choose algorithms that detect the patterns that human annotators base their judgements on remains and that there are yet no hybrid methods that ensure control over the classification process and at the same time offer a powerful solution that works on any constructs one might find in interactive discourse (Baden et al., 2020). On the other hand, human annotation is far from flawless and achieving sufficient reliability scores for complex concepts is a challenging task in itself.

This section has attempted to provide a brief summary of previous computational studies of polarization in online discussions. The chapter that follows moves on to consider the technical details of the tools and methods used in this study.

3 Technical Details

The chapter below describes the procedures and methods used in this investigation. We will describe the technical details of the data collection, the NLP tools applied to the data and the language model we utilize to obtain embeddings of the comments. Also, the statistical modelling used in the study will be accounted for.

3.1 Data Collection

With respect to the data used in this project, it consists of threaded discussions about the covid-19 pandemic collected from Reddit. In total, 2 186 posts with 1 229 490 comments are retrieved and the posts range in date with the oldest being from May 24th 2021 and the most recent from March 10th 2022. The discussions are collected using PRAW (Python Reddit API Wrapper) (Boe, 2016) which is a Python module that facilitates the collection of Reddit data using Reddit's API. PRAW allows the user to search Reddit posts and comments by e.g., subreddit, username or topic and provides metadata for the collected posts and comments. The data collected with PRAW is limited to text and the results do not contain images or videos.

To obtain discussions that are likely to concern the covid-19 pandemic, PRAW's keyword search is used with the search terms *vaccine*, *covid*, *corona*, *pandemic*, *covid-19*, *virus*, *wuhan*, *dosis* and *lockdown*. The keyword search returns posts that match each search term, sorted by relevance. The relevance sort takes into account the relative rarity of the words in the search query, the creation date of the posts and the numbers of votes and comments the posts have (Reddit, 2022b). There seems to be a limit to the number of posts returned by the keyword search since only 224–251 posts are returned with each call. The relevance search also seems to heavily prioritize more recent posts. The distribution of the collected posts by month is visualized in Figure 1 below.

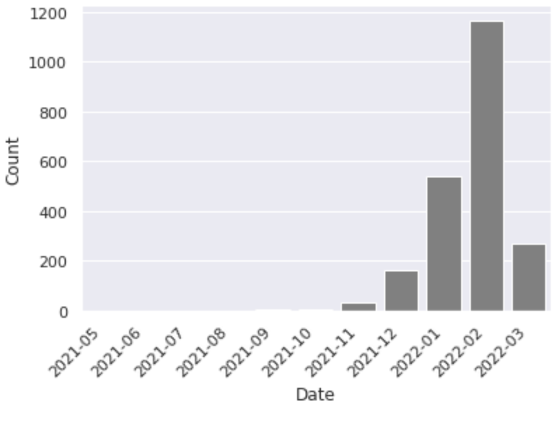

Figure 1: Distribution of collected posts by month.

The metadata collected together with the posts consists of unique post ID's and creation date. The metadata collected together with the comments contains the username of the author, vote score, creation date, comment ID and parent comment ID. The discussions are structured in such a way that each comment is a reply to exactly one parent comment and the parent comment ID represents this relation. That is, the threaded structure of the discussions can be maintained using the comment ID's and parent comment ID's. Regarding the usernames of the comment authors, they are not accessible if the account has been deleted after creation. In the case of the vote score, it represents the difference between the number of upvotes and downvotes for each comment at the time of collection. While it is possible to obtain the vote ratio for posts, which is the relashionship between upvotes and downvotes, this is not possible for comments using PRAW.

3.2 NLP tools

This subsection describes the three NLP tools that are used to analyze the collected comments and quantify their toxicity, anger and positive sentiment. The tools are Google Perspective's toxicity detection and two pre-trained and fine-tuned transformer models that predict the expressed emotion and sentiment of a text. The tools are chosen due to their potential to capture expressions of affective polarization in our data.

3.2.1 Toxicity Detection

The tool used to detect toxicity in our data is Google Perspective API. Perspective API is a freely available resource that utilizes machine learning to automatically detect toxicity in comments posted online (Jigsaw, 2017). Perspective's model classifies text on the comment level and returns a score between 0 and 1 that indicates how likely it is for that comment to be interpreted as toxic. The purpose of this resource is to assist human moderation of online forums and comment sections, offer immediate feedback to the commenters and to facilitate research on abusive language online. The service has been used previously in similar studies on polarization in social media discussions (e.g., Chipidza, 2021).

The toxicity attribute of Perspective API currently supports 17 languages [1]. The developer's definition of a toxic comment is "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion" (Jigsaw, 2017). For the languages English, French, German, Italian, Portuguese, Russian and Spanish, additional attributes are supported. These attributes include severe toxicity, insult, profanity, identity attack, threat and sexually explicit. While the attributes profanity, threat and sexually explicit are less relevant for our research aims, we hypothesize the attributes toxicity, insult and identity attack to be more likely to correlate with the presence of affective polarization in the collected data.

Perspective's model is trained on millions of comments from online forums which are annotated for different aspects of toxicity. For the attributes severe toxicity and toxicity, each comment is rated as either very toxic, toxic or not toxic. For the attributes insult, profanity, identity attack, threat and sexually explicit, the comments are coded according to whether or not they contain each of the concepts. The model is then trained to represent the annotation so that a comment that is considered toxic by 8 out of 10 annotators is assigned a toxicity score of 0.8. In other words, the score returned by the API represents probability and is not intended to specify the degree of toxicity in the input.

Perspective's toxicity detection is suitable for our study, as our data resembles the data on which the model is trained. In both cases, the data consists of comments from online forums. However, Perspective declares that their training data and consequently their model is biased. They state that the presence of certain words such as "muslim" or "feminist" may trigger a higher toxicity score, due to their over-representation in hateful comments (Jigsaw, 2017). Moreover, they acknowledge that their classifier makes mistakes and is unable to detect new patterns of toxicity. For these reasons, Perspective advise against the use of their service for fully automated moderation or to make judgements about the character of the comment author.

3.2.2 Emotion Detection

To detect expressions of emotions in our data, we utilize a pre-trained language model that has been fine-tuned for this purpose. We use DistilBERT (Sanh et al., 2019) which is a lighter version of BERT (Devlin et al., 2019) and allows faster building of task-specific models. The model we use has been fine-tuned for emotion detection by Tunstall et al. (2022) on the emotion data set created by Saravia et al. (2018).

[1] Supported languages are Arabic, Chinese, Czech, Dutch, English, French, German, Hindi, Hinglish, Indonesian, Italian, Japanese, Korean, Polish, Portuguese, Russian, and Spanish.

The emotion data set consists of tweets in English that are collected using hashtags. The hashtags have an additional function, which is to serve as ground truth during training of the model. For example, the hashtags *#depressed* and *#grief* indicate that sadness is expressed in the body of the tweet and the hashtags *#mad* and *#pissed* indicate expression of anger.

The fine-tuned emotion detection model assigns probabilities to the basic emotions sadness, joy, love, anger, fear and surprise given a text input. The model is designed so that the probability assigned to the six emotions sums to 1. According to Tunstall et al. (2022), the fine-tuned model obtains an accuracy of 92.7% which is calculated by checking whether the most probable emotion predicted by the model coincides with the ground truth. For our purposes however, we consider only the probability assigned to anger, even when other emotions are more probable. As with the toxicity detection, the score is intended as a probability score and not a representation of the level of anger in the input.

A potential issue with the emotion detection model is that it has been trained on Twitter data exclusively, which differs from Reddit data in several ways. Firstly, tweets have a limited length which is not the case for Reddit comments. Therefore, the emotion detection may not perform as well with longer input. Secondly, tweets and Reddit comments differ in style and purpose. While Twitter users often use their offline identity on the platform, Reddit is an anonymous discussion forum. Tweets are also intended to be read in isolation while comments on Reddit are created in an interactional context and typically require this context to be understood fully.

3.2.3 Sentiment Analysis

To detect positive sentiment in our data, we again use a pre-trained version of DistilBERT (Sanh et al., 2019) that has been fine-tuned for this purpose. The task-specific model is available through the Transformers library (Wolf et al., 2020) and has been trained on Stanford Sentiment Treebank (Socher et al., 2013). The treebank is a collection of sentences retrieved from movie reviews and their annotation. Each sentence has been labelled with either positive or negative sentiment according to what can be inferred from the review.

The sentiment model performs binary classification of a sentence and assigns it a positive or negative label along with a confidence score. This means that when a comment is predicted to be negative, no confidence score for positivity is provided. Consequently, the positivity scores range from 0.5 to 1. To avoid this and to obtain positivity scores between 0 and 1, like we do for toxicity and anger, we subtract the negativity score from 1 when a sentence is labelled as negative. Perhaps the most serious disadvantage of the sentiment analysis four our purposes is that the model is trained on movie reviews which are very different in character from comments in online discussions.

3.3 Word Embeddings

To represent the linguistic information in our data, we utilize embeddings from the pre-trained language model BERT (Devlin et al., 2019). BERT is trained on unlabelled data from books and Wikipedia containing over 3 million words. The training is unsupervised and the model learns language representations through two tasks: masked language models (MLM) and next sentence prediction (NSP). For the MLM objective, 15% of the tokens are hidden from the model at random and the task consists of predicting the masked tokens based on their context. NSP serves to learn relationships between sentences and the task consists of, given a sentence, predicting if a proposed sentence is the following one.

BERT has an attention mechanism which helps the model to maintain global connections between the input and the output. The attention layers allow the model to learn what parts of the context to attend to when making a prediction, which is particularly useful for long-range dependencies (Vaswani et al., 2017). In contrast to earlier transformer language models (e.g., OpenAI's GPT, Radford & Narasimhan, 2018), BERT employs bidirectional self-attention. That is to say that the model has access to the context before and after the current token while unidirectional models can only attend to tokens to the left.

Due to BERT's architecture and training process, the language model is capable of creating rich, contextual representations of words, phrases and sentences. The text given as input to the model is encoded in a multidimensional space so that, for instance, embeddings of related words have a relatively small cosine distance and embeddings of more unrelated words are located further apart. The word embeddings are context dependent so that a particular word will get slightly different embeddings depending on the linguistic context it occurs in. To represent the linguistic information in our data, we give the collected comments as input to BERT and retrieve the embeddings from the output.

To obtain this, the comments are tokenized before given as input to BERT. To summarize the semantic content of the input, we use the hidden state of the output of the model as recommended in the documentation (Huggingface, 2020). The hidden state consists of 12 layers and 768 dimensions for each token of the input. Since the 12 layers encode different kinds of information, the most suitable method differs across tasks (Alammar, 2018). We take the penultimate layer of the hidden state and average it across the tokens, to get a representation of the whole input. The motivation behind this approach is that the penultimate layer has been proven to work well as a contextualized embedding of words, at least for the task of named-entity recognition CoNLL-2003 (Tjong Kim Sang & De Meulder, 2003). With this approach, we get a vector of 768 numerical values that represent the input comment.

Next, the embeddings obtained from BERT are dimensionality reduced before they are given as individual features to the logistic regression model. The dimensionality reduction serves both to facilitate the interpretation of the logistic regression results and to eliminate unnecessary or redundant information encoded in the high-dimensional vectors (Rao, 2019). For this purpose, we utilize Scikit-learn's principal component analysis (PCA) (Pedregosa et al., 2011). This tool projects the vectors to a lower dimensional space using Singular Value Decomposition (SVD). The features of the input vector are centered, but not scaled, before the SVD is applied. With this method, the vectors of 768 dimensions that we retrieve from BERT are reduced to vectors of 10 dimensions. The linguistic information in the reduced vectors is consequently not as rich, but we hypothesize that important features of the input are maintained.

The purpose of including reduced word embeddings in our study is to investigate whether linguistic information of this kind can assist in the detection of affective polarization. Specifically, we are interested in comparing the suitability of word embeddings with the output of the NLP tools described above. Although the toxicity, anger and positivity scores build on the linguistic information of the input, that information is not explicitly available in them. However, a possible drawback with pre-trained BERT is that while it is used successfully in a range of NLP tasks, it is not trained on comments from online discussions which may affect its capability to model our data.

3.4 Logistic Regression

So far this chapter has focused on the tools and methods used to represent our collected data. The following section will explain the method we apply to statistically model these representations and the results of the manual annotation. The goal of the statistical analysis is to decide the probability that a conversation contains expressions of affective polarization given its vote scores, the confidence scores of the NLP tools applied to it and its dimensionality reduced BERT embeddings. Additionally, we want to evaluate how useful each of these features are for detecting expressions of affective polarization in our data.

To obtain this, we perform logistic regression analysis. A logistic regression model provides the probability of an outcome, given the statistical properties of a data set. The probability of an outcome is referred to as the dependent variable and the data consists of one or more independent variables. The independent variables may include either continuous or categorical values. Besides predicting the probability of an outcome, the regression analysis defines the relationship between the independent variables in the data and the dependent variable and that relationship can be expressed with the following function.

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m)}}$$

To calculate the probability that a conversation contains expressions of affective polarization, we need to estimate the coefficients (β) of the logistic regression function. The coefficients are estimated during fitting of the model with the maximum likelihood method. Maximum-likelihood estimation determines the values of the coefficients so that they maximize the probability of getting the outcome as observed in the data (Hosmer et al., 2013). In this project, we use Statsmodels API (Seabold & Perktold, 2010) to fit the logistic regression model to our data and estimate the coefficients of the independent variables.

After fitting the model, we want to assess the significance of each independent variable in our data. What we want to investigate is whether the presence of a variable improves the prediction of the outcome in comparison to a model that does not contain the variable. The p-value is typically used for this purpose and is calculated by comparing the log-likelihood of the two models with and without the variable. This is commonly calculated automatically and reported in statistical software (Hosmer et al., 2013). The Statsmodels API package calculates and provides the p-value associated with each independent variable. P-values below 0.05 traditionally indicate rejection of the null hypothesis and can be interpreted as a chance of less than 5% of obtaining the observed value, given that the null-hypothesis is true. The null hypothesis in this case is that the coefficient has no effect on the model's outcome.

Once the values of the coefficients are known, the log odds of a conversation being polarized can be calculated using the following linear prediction function.

$$f(i) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_m x_{m,i},$$

The first term β_0 is the intercept and is obtained by adding an independent variable with a consistent value of 1. The following terms β_{1-m} are the coefficients of the features of our data. As the prediction function above shows, the values of the independent variables are multiplied with their coefficients and summed to calculate the log odds of an outcome. To clarify, the relationship between the coefficients and the values of the independent variables is assumed to be linear. Moreover, the impact of the feature values is assumed to be independent which implicates that it is not possible to distinguish the impact of combinations of features such as high toxicity and high positivity from high toxicity and low positivity.

The coefficients of the prediction function represent the log odds of the independent variables so that one unit's increase results in the corresponding increase in log odds. Log odds range from -∞ to ∞ and can be converted to odds through exponentiation. The odds in turn range from 0 to ∞ and represent the ratio between the probability that an outcome occurs and the probability that an outcome does not occur. Because of this relationship, the odds can be used to calculate probability which is also what the logistic regression model returns. This probability can be used for classification so that conversations with an assigned probability above 0.5 are classified as polarized. Subsequently, the performance of the prediction function can be evaluated through prediction on unseen test data.

4 Experimental Setup

The following part of this report moves on to describe in greater detail the scientific method and the experimental setup of the study. First, we briefly discuss some methodological challenges of content analysis of social media discussions and motivate our design choices with respect to them. Second, we account for how the collected data is represented for the manual annotation as well as the statistical analysis. Third, we provide a detailed description of the annotation process and the annotation guidelines.

As was mentioned in the introduction to this report, our study aims to explore a set of NLP tools and methods in the context of automated detection of affective polarization in Reddit discussions. The study is primarily exploratory and we hope to gain insight in the usefulness of such methods in relation to this task. That is to say, we do not aim to develop high-performing classification due to the limitations of this study and the well-known challenges related to detection of abstract concepts in text. Nevertheless, our ambition is that the findings in this study facilitate future work in this domain.

Regarding content analysis of online discussions, Herring (2004) argue for a revision of traditional content analysis methodology. She claims that studies of online discussions tend to be anecdotal and lack sufficient empirical grounding, but that established methods of old media content analysis are typically too narrow to capture phenomena in online discussions. As an illustration, Herring (2010) highlight the necessity of selecting research questions that can be answered with the available data and to define salient features in that data to analyze. In contrast, traditional content analysis relies heavily on predetermined research questions and coding categories as well as calculation and evaluation of inter-annotator agreement.

The methodology of this study resembles the one proposed by Herring (2004). We formulate research questions and annotation guidelines according to important aspects of affective polarization theory and the patterns encountered through manual inspection of the collected data. However, inter-annotator agreement is calculated on the annotated data as a measure of reliability. Regarding the collection of Reddit discussions, we employ event-based sampling, as suggested by Herring (2010), since we hypothesize that online discussions about the covid-19 pandemic increases our chances of finding expressions of polarization. We believe that event-based sampling is justified by the fact that the aim of this study is not to determine the frequency of the phenomenon. In terms of sampling of the collected comments we want to maintain their relation, since random sampling is unsuitable for analysis of conversations (Skalski et al., 2017).

On the question of applying methods of analysis to the collected data, we select logistic regression as it is suitable for binary prediction and to represent the presence or absence of a characteristic (Vittinghoff, 2012). A major advantage of logistic regression is that its output is comprehensive and transparent in a way that more advanced machine learning models are not (Baden et al., 2020). For more advanced models, unknown biases may govern the classification and consequently decrease researcher control over the process (Skalski et al., 2017). An additional benefit of logistic regression is that it can operate on smaller data sets. This is not unimportant due to the limitation of this study and the fact that we collect and annotate our data. However, logistic regression models lack the ability to detect more complex relationships in the data that more advanced machine learning models can capture.

4.1 Representations of Conversations

To annotate and perform statistical analysis on the scraped Reddit discussions, we need a suitable way to represent them. The over two thousand posts collected from Reddit vary in size and many posts are several thousand comments long which makes them difficult to read. The threaded structure of the discussions complicates comprehension further. However, analysis of the comments in isolation is not a suitable approach as it may lead to erroneous assumptions (Baden et al., 2020) and methods that are sensitive to the conversational context of e.g., forum contributions are shown to improve accuracy (Walker et al., 2012).

The conversational context is especially important for our research aims, as we expect the expressions of affective polarization to manifest differently depending on the presence or absence of members of in- and out-groups.

To overcome this issue and to maintain the interactive nature of Reddit discussions, we select snippets of the threaded discussions as representations. There are several things to consider when selecting snippets to represent the discussions. Baden et al. (2020) highlight some of the difficulties related to the tendency of online conversation to lack a clear beginning and end and that users can enter and exit the conversation at any point. With this in mind, the snippets should be long enough to allow an exchange of views and opinions. They should also be long enough to enable multiple participants to engage in the discussion and reply to contributions made by others. However, the snippets should be short enough for the manual annotation. Besides, as we are giving average scores to the logistic regression, snippets that consist of too many comments risk that single contributions that show expressions of affective polarization are lost. In consideration of this, we retrieve conversation snippets consisting of seven comments.

Moreover, we need to consider which part of the discussion tree is the most appropriate to represent the discussion as a whole. While there may be several suitable approaches, we decide to select the first seven comments of the longest thread of comments in the discussion. We hypothesize that the longest thread would regard topics that are of interest to the readers of the post since many of them choose to interact with it and that it therefore constitutes an adequate representation. Moreover, we find it probable that the largest thread holds, for our purposes, interesting phenomena due to the engagement it generates among the participants.

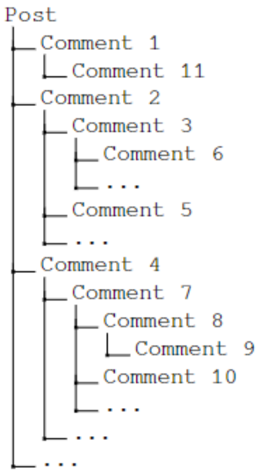

Figure 2: An abstraction of the threaded structure of a Reddit discussion.

The tree structure in Figure 2 to the left is a simplified example of a Reddit discussion. It consist of a post title followed by multiple comments. Comments 1, 2 and 4 are root comments since they are replies to the post title and consequently have the post as their parent. Comment 9 and 11 are leaf comments since they are not the parent to any other comment in the tree. In order to find the longest possible path, we count the number of comments from the leaves to the root. In this example, there are two possible paths. The first possible path is two comments long and consists of Comment 1 and Comment 11. The other possible path is four comments long and consists of Comment 4, Comment 7, Comment 8 and Comment 9.

The method described above is applied to every discussion in the collected data. Posts that do not contain paths that are seven comments or longer are discarded. Similarly, longer paths are truncated so that they contain the first seven comments only. This results in 1 234 conversation snippets that are exactly seven comments long. The conversation snippets are portrayed in two ways: First, in plain text for human annotation and second, with numerical values for the statistical analysis. This is illustrated with examples and explained in greater detail below.

936, svt4tm, Efficacy of Ivermectin on Disease Progression in Patients With COVID-19

936.0, hxi25t1, user_a
Good news.
All those people who suggested Ivermectin to others and to refuse all other treatments have blood on their hands.

936.1, hxi5k8a, user_b
can we get a consensus that advocating for this is reportable as misinformation?

936.2, hxicwpt, user_c
The eternal pro-censorship lefty Redditor.

936.3, hxjvlkg, user_d
Meanwhile right leaning people love a little censorship themselves.
Not one sided there bud.

936.4, hxjy3sf, user_c
I'm anti-censorship

936.5, hxjy8n9, user_d
Well hey, trying to bury misinformation isn't censorship, cause well, it's misinformation.

936.6, hxjyby5, user_c
It's still censorship.
It's so funny how many lefties say 'but, it's not censorship, because we're right and they're wrong!'
It's truly stunning.
/r/selfawarewolves material

The above example shows how conversations are displayed in text for the manual annotation. The post title is shown at the top in order to provide context to the comments that follow. The comments are numbered and the username of the author is displayed as well, in order to enhance comprehension further. Reddit provides a function to cite previous comments in a block quote which is used frequently in discussions. This is formatted accordingly with the Reddit interface and in our data, block quotes are marked with a '>' in the text and the quote is separated from the rest of the comment with a new line above and below. This is exemplified in comment 959.5 below. Such block quotes are excluded from the analysis with the NLP tools. However, "normal" quotes are not excluded and are displayed the way the author formatted them, as in comment 936.6 above.

959.5, hlj4yzi, user_a
These are the same companies that told us OxyContin was a miracle drug.

>It's exhausting having to re-explain this so many times.

Nobody is asking you to do anything.
If it's exhausting, don't do it.

4.2 Conversation Features

The conversations are represented with 15 features consisting of numerical values for the regression analysis. The features hold information about the vote scores obtained by the comments in the conversation snippet, the confidence scores of the NLP tools applied to them as well as their dimensionality reduced BERT embeddings. The values of the features of conversation 936 are exemplified in Table 1 below. The vote sum of a conversation is obtained simply by adding the votes of the seven comments and computing the logarithm of the value. Toxicity, anger and positivity scores are computed through application of the NLP tools on the conversation. Each comment is analyzed individually and the average of the scores is computed to represent the conversation snippet. The ten dimensions of the BERT embedding represent the linguistic information of the conversation snippet and are obtained in a similar way. To obtain them, the mean of the seven comment vectors is computed and the dimensions are subsequently reduced to ten (see 3.3 Word Embeddings).

4.2.1 Vote Swings

The vote swing feature is designed to capture swings in the pattern of comment votes. An observation of the vote scores of the comments in the conversation snippets reveals a tendency for vote scores to be higher for the first comment and decay towards to end. However, a considerable amount of conversations shows a different pattern of vote scores swinging up and down throughout the snippet. Manual inspection of the comments and their obtained number of votes leads us to believe that the swinging pattern often occurs when there is a disagreement and intense debate in the conversation. Since simple addition of the vote scores do not capture such swings, we design a measure to capture them.

Any vote score that is higher than the vote score of the previous comment is considered a vote swing. To calculate the magnitude of a swing, we use the ratio of the current comment's vote score and the previous comment's vote score. To calculate the value of vote swings in a conversation snippet, the vote scores of all comments need to be positive. To ensure this, the absolute value of the lowest score plus 1 is added to all vote scores when there are negative values in a snippet. When there are no swings in an array and every vote scores is lower than the previous, the vote swing value is 0. When there are several vote swings in a conversation, their values are multiplied. Subsequently, the logarithms of the vote swing values are computed.

Feature	**Value**
Vote sum	4.919980925828125
Vote swings	3.386808644210056
Toxicity	0.1052265785714285
Anger	0.4659729492185371
Positivity	0.3394889320646013
Dimension 1	-0.6663250923156738
Dimension 2	0.7394592761993408
Dimension 3	0.838313639163971
Dimension 4	-0.7391476631164551
Dimension 5	-0.16352471709251404
Dimension 6	-0.05322221666574478
Dimension 7	-0.199598029255867
Dimension 8	0.9864889979362488
Dimension 9	-0.4132081866264343
Dimension 10	0.4791337549686432

Table 1: Features of conversation snippet 936.

4.3 Annotation Process

This section describes the process of annotating the conversation snippets for features related to affective polarization. Based on our research questions, theory on affective polarization and a manual inspection of the items we identify concepts to code and suitable annotation guidelines. The annotation process consists of two test rounds and the annotators are researchers involved in the GRIPES project[2]. In the first round, three annotators annotate 100 conversations individually according to the guidelines and their inter-annotator agreement (IAA) is computed on the result. Tricky cases are discussed and the features are reconsidered and the annotation guidelines refined, in order to obtain higher reliability across the annotators. After this process, there is a second test round of annotation of 100 conversations across the same three annotators and IAA is computed again. This is followed by annotation of one of the annotators of the remaining conversations.

In order to decide what features should be coded, the conversation snippets are manually inspected. During this part of the process we aim to get an impression of what expressions of affective polarization as described in the literature can be found in them and how they can be identified. Most importantly, we consider what phenomena are necessary to capture to be able to answer our research questions. However, it is also necessary to explore what phenomena are possible to capture with limited resources in terms of number of annotators and time. In other words, we want the features and annotation guidelines to be complex enough to capture important aspects of polarization yet simple enough to ensure consistent coding across the annotators. It is also important that they are grounded in theory about affective polarization, identity performance and intergroup conflict.

During the first round of annotation, the conversation snippets are annotated for the following features: conversation in English, covid-related topic, in-group praise, out-group derogation, disagreement and discussion balance. Since both in-group praise and out-group derogation are examples of behaviour related to affective polarization, we choose to interpret them as indications of polarization. A further important concept according to theory is group belonging of the discussion participants. In the first round, we assume that a suitable way of representing that is to annotate whether or not the conversations contain disagreement.

[2] GRIPES – Gothenburg Research Initiative for Politically Emergent Systems

If a conversation contains disagreement, we also code the balance of the discussion. We do this by marking if the discussion balance is 1 vs. 1, 1 vs. many or many vs. many as theory suggest this has consequences for how the participants express themselves.

The first test round resulted in low IAA, especially for in-group praise and discussion balance. For this reason we decide to discard those features in the second test round. An important observation from the first test round is that in-group praise rarely occurs in our data in comparison to out-group derogation and is consequently not as common in our corpus as the literature suggests. For the remaining features, the annotators discuss examples with low IAA and clarify the guidelines accordingly. During the second round of annotation, the conversation snippets are consequently annotated for the features conversation in English, covid-related topic, out-group derogation and disagreement. The guidelines followed by the annotators during the second and final annotation rounds are described in the section below.

4.3.1 Annotation Guidelines

In this section, we summarize the annotation guidelines followed during the second and final round of annotation and on which the statistical analysis in this study are based on. The full codebook is available in Appendix.

The annotation involves evaluation of the four statements for each conversation snippet. If the annotator agrees with a statement, it is marked with 1. If the annotator disagrees with a statement, it is marked with 0. Annotation statements 1–2 concern the language and the topic of the conversation and a conversation is only relevant for further annotation if the annotator agrees with both statements. If statement 1 or 2 are marked with 0, the following statements are coded with 99 and excluded from the study. Annotation statement 3 measures the affective polarization in the conversation and annotation statement 4 estimates if there is any disagreement in the conversation. The statements are formulated in the following way:

1. The conversation is in English.
2. The conversation handles a topic related to the covid-19 pandemic.
3. The conversation contains derogation of the out-group.
4. There is disagreement in the conversation

The first annotation statement regards the language of the conversation. We decide to only consider conversations in English as our NLP tools are developed on English text and English is the mutual language of the annotators. There is a tolerance for words and phrases from other languages that are expected to be understood by monolingual speakers of English. A conversation is marked with 1 if the conversation is in English and with 0 if there are comments in other languages than English so that a monolingual English speaker cannot follow the conversation.

The second annotation statement regards the topic of the conversation and this study only consider discussions that relate to the covid-19 pandemic. The evaluation of this statement can be complicated since Reddit discussions tend to handle multiple topics that change along the course of the discussion, even when they are short like our selected snippets. Determining the topic of interactive communication is challenging compared to e.g., news articles (Baden et al., 2020). Additionally, there are many topics that are related to the covid-19 pandemic in ways that are difficult to anticipate. With this in mind, conversations that include at least two comments that touch upon a topic related to the covid-19 pandemic are annotated with 1. Our examples of covid-related topics include, but are not limited to, scientific or medical questions regarding covid-19 as a disease, the coronavirus and its spread, remedies against covid-19 such as restrictions and vaccines and protests against them. Economic and political consequences of the pandemic and related policies on a societal or individual level are also considered relevant as well as discussions or opinions about celebrities or politicians associated with strong attitudes towards the pandemic, vaccines and restrictions.

The third annotation statement evaluates if derogation of the out-group is expressed in the conversation. Derogation of the out-group involves expressing negative attitudes towards a group that the author does not identify or agree with (Wojcieszak et al., 2021). This involves expressing dislike or distrust for members of the out-group and assigning its members disadvantageous intentions and qualities such as dishonesty or hypocrisy (Nordbrandt, 2021; Iyengar et al., 2012). Denying members of the out-group positive traits can also be considered out-group derogation. Because of the difficulty to establish group belonging without context beyond the seven comments in the snippet, the conversation will be annotated with 1 if at least one participant in the conversation expresses negative attitudes towards another person or group, regardless of their presence in the conversation. It is not necessary that the derogation of a person or group is based on opposing opinions about the covid-19 pandemic. However, the attack must be directed towards a person or group of people which means that attacking e.g., a policy does not count as derogation of an out-group.

The fourth annotation statement measures if there is disagreement in the conversation snippet, which is meant as a proxy for the presence of out-group members. If there is any kind of disagreement expressed among the participants, it will be annotated with 1 on this task. This means that if at least one participant is questioning another participant's statement or opinion or expressing a different view on the present topic, the conversation will be considered to contain disagreement. The disagreement can be either explicit or implicit and the topic of disagreement does not need to be related to the covid-19 pandemic. However, the disagreement needs to occur between participants of the conversation and a participant expressing disagreement with a policy or politician not present in the conversation is not considered disagreement. Consequently, if all participants in a conversation agree with each other or express similar opinions, the conversation will be annotated with 0.

4.3.2 Inter-Annotator Agreement

To calculate inter-annotator agreement (IAA), we use Krippendorff's alpha which is a method emerging from content analysis. There are several softwares for computing Krippendorff's alpha and we utilize the Simpledorff package for Python (Perry, 2021). Krippendorff (2004) highlights that his alpha measures agreement across annotators and not reliability but that reliability of the annotation can be inferred from the agreement. A value of α=1 indicates perfect agreement, values above 0 indicate systematic agreement and values below 0 indicate systematic disagreement. It is calculated with the following formula, where D_o is the disagreement observed and D_e is the disagreement expected by chance.

$$\alpha = 1 - \frac{D_o}{D_e}$$

$$D_o = \frac{1}{n} \sum_{c \in R} \sum_{k \in R} \delta(c,k) \sum_{u \in U} m_u \frac{n_{cku}}{P(m_u, 2)}$$

$$D_e = \frac{1}{P(n,2)} \sum_{c \in R} \sum_{k \in R} \delta(c,k) P_{ck}$$

The fact that the agreement is compared with the agreement expected by chance has consequences when annotating imbalanced categories. When a category is very common or very rare, it has consequences for the value of alpha which will be lower. With this in mind, we also report IAA in terms of accuracy. The accuracy is calculated by dividing the number of items that were annotated equally by all three annotators with the total number of items. The tables below present the obtained IAA in terms of Krippendorff's Alpha and accuracy from annotation round 1 and 2 respectively.

Task	**α**	**Accuracy**
Conversation in English	1.0	100%
Covid-related topic	0.67	79.57%
In-group praise	0.25	64.52%
Out-group derogation	0.50	69.35%
Disagreement	0.61	77.42%
Discussion balance	0.16	18.92%

Table 2: Inter-annotator agreement test round 1.

Task	**α**	**Accuracy**
Conversation in English	1.0	100%
Covid-related topic	0.73	82%
Out-group derogation	0.35	57%
Disagreement	0.58	81%

Table 3: Inter-annotator agreement test round 2.

The values of alpha are generally low, except for the task *Conversation in English* which has perfect agreement in both annotation rounds. Krippendorff (2004) is hesitant to give thresholds that separate reliable annotation from unreliable annotation, but highlights the need for high agreement when the consequences of making erroneous conclusions are grave. When tentative conclusions are sufficient, he suggests that values above 0.667 can be accepted. However, only the tasks *Conversation in English* and *Covid-related topic* obtain alphas that meet this threshold.

Another thing worth mentioning is that IAA in terms of accuracy is higher in the second round for the tasks *Covid-related topic* and *Disagreement* but notably lower for *Out-group derogation*. The differences between Krippendorff's Alpha and accuracy as measures of agreement are well illustrated in Table 3 and the reported agreement for *Covid-related topic* and *Disagreement*. While all three annotators agree on 81% and 82% of the items respectively, alpha is notably lower for *Disagreement*. This is due to the class imbalance and that a large majority of the items are annotated to contain disagreement while covid-related topic is a more balanced category.

The low IAA in round two raises the question about the reliability of the data we use to fit the logistic regression model. While α=0.58 for *Disagreement* is close to meet the suggested threshold of 0.67, α=0.35 for *Out-group derogation* is critically low. It is easy to argue for the invalidity of the results due to the low IAA, but it may also be argued that it is problematic to blindly follow IAA thresholds on annotation categories with this degree of subjectivity. The annotators in Walker et al. (2012) who code stance classification in online discussions, which is another highly subjective task, obtain an accuracy of 77%. Davidson et al. (2020) also highlight the difficulties of achieving acceptable IAA when developing their incivility classifier. They perform several annotation rounds before finally obtaining a Fleiss's kappa of 0.663. According to the authors, sarcasm is often involved in the tricky cases that result in different judgements among the coders.

In this project, we solve the issue with inconsistent annotation by letting one of the annotators annotate the remaining conversations alone. The limited scope of this study does not allow further annotation rounds and it is our hope that analysis of the data may lead to important insight on affective polarization in online discussions, despite the low IAA. Sayeed (2013) identifies the difference in conservativity among annotators as a challenge in his criticism of IAA as a reliability measure in the context of opinion detection. Our IAA also suffers from the fact that the annotators disagree on the quantity of conversations that contain e.g. out-group derogation. A possible solution is the approach applied by Jigsaw (2017) in the development of their toxicity detection model (see 3.2.1 Toxicity detection). Since the judgement of annotators can be expected to vary on subjective tasks, they interpret the accuracy for each annotated comment as the probability that the comment is interpreted as toxic.

4.4 Regression on Conversation Features

This section describes the final step of the experimental setup which involves fitting a logistic regression model on the data. This is performed by treating the numerical values of the features extracted from the conversation snippets as independent variables and the binary annotation as the dependent variable. Only the conversation snippets in English that are annotated as related to the covid-19 pandemic are included in this step which results in 705 items. Since they are annotated with 1 or 0 for out-group derogation and with 1 or 0 for disagreement, the conversations can be divided into categories according to combinations of their annotation.

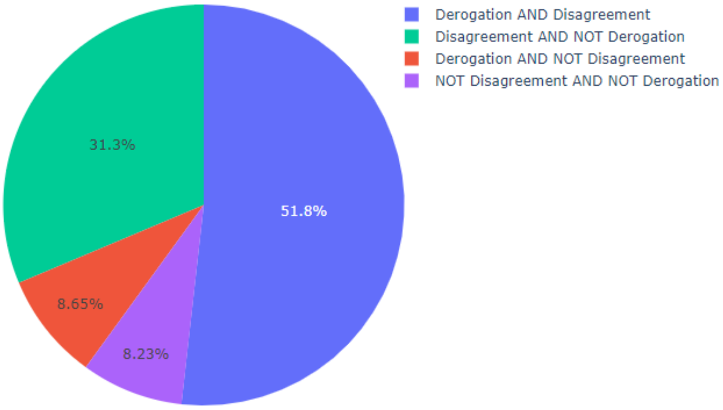

Figure 3: Distribution of annotated conversations.

Figure 3 to the right display the possible combinations of out-group derogation and disagreement. The pie chart illustrates that a majority of the conversations contain derogation and that a large majority contain disagreement of some kind. It also visualizes that while most conversations with disagreement also contain out-group derogation, a large share of conversations contain disagreement without out-group derogation. We are foremost interested in out-group derogation since it is a potential expression of affective polarization. Disagreement is of interest to test our hypothesis that the characteristics of the out-group derogation change depending on this. In regards to this, we will only consider two of the groups in the pie chart for the statistical modelling, namely derogation + disagreement and derogation - disagreement. We will also consider the groups of conversations that are annotated with 1 for derogation and 1 for disagreement for comparison. The table below displays the categories we will use in the regression analysis and what their annotation looks like.

Category	Out-group derogation	Disagreement
Out-group derogation	1	*
Disagreement	*	1
Out-group derogation AND Disagreement	1	1
Out-group derogation AND NOT Disagreement	1	0

Table 4: Categories of conversations. Asterisk indicates that the value can be either 0 or 1.

For each of the four categories in the table above, separate logistic regression models are fitted. For each model, the dependent variable is set to 1 if the annotation of a conversation matches the annotation as displayed in Table 4 for the corresponding category and to 0 otherwise. The separate models are fitted to correlate the conversation features with a) out-group derogation, b) disagreement, c) out-group derogation AND disagreement, d) out-group derogation AND NOT disagreement.

Subsequently, the fitted models are evaluated in two ways. The first evaluation approach considers the coefficients of the features and their significance. The purpose of this is to test the ability of the selected NLP tools to capture aspects of affective polarization in online conversations. The second approach involves testing the models on unseen data and compute evaluation metrics. To do this, we fit the model on 80% of the data and test it on 20%. We apply 5-fold cross-validation and use the split that results in the best performing model in terms of F1 for each task and for each set of features. To evaluate the predictive power of the models, we calculate and report accuracy, precision, recall and F_1 with respect to the positive class.

The four kinds of models that are fitted according to the conversation categories are divided into separate models with different sets of features. For each conversation category, we fit four separate models. The first model contains the features vote score, vote swings, toxicity, anger and positivity. The second model contains toxicity only. The third model contains 10 dimensions of dimensionality reduced BERT embeddings and the forth models contains all features. This results in a total of 16 fitted logistic regression models. Because of the very strong correlation between out-group derogation and toxicity, we test if toxicity alone can predict out-group derogation successfully. Similarly, we want to explore if linguistic information is more suitable for predicting affective polarization than the results of the selected NLP tools.

5 Experimental Results

In the following chapter, the results of the study will be presented. The first section examines the data and inspects the distribution of conversation in terms of annotation results and values of features. The second section evaluates and compares the predictive performance of the fitted models on test data. Moreover, the coefficients of the best performing models are inspected and their statistical significance examined.

5.1 Annotation Results

The annotation of the data results in 705 conversation snippets in English that handle a topic related to the covid-19 pandemic. As explained earlier, each snippet is annotated for out-group derogation and disagreement which is represented with 0 and 1. Each snippet is also represented with numerical values which correspond to vote sum, vote swings, toxicity, anger, positivity and dimensionality reduced BERT embeddings. To better understand the results of the logistic regression, we examine the distribution of conversations in terms of annotation and feature values and visualize the findings.

The first feature in the data is vote sum which is the logarithm of the sum of the vote scores obtained by the comments in the conversation snippets. Before computation of the logarithms, the values range from 3 to 31 196 with a median of 219 and a mean of 1060. In other words, a few conversation had extremely high values and taking the logarithms of the values serves to facilitate the modelling of this feature. Calculation of the logarithms results in feature values that range from 1.10 to 10.35, with a median of 5.38 and a mean of 5.46. The distribution of these values across the conversations is presented in Figure 4.

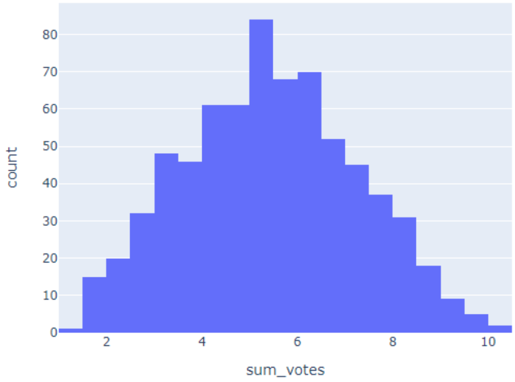

Figure 4: Distribution of conversations by vote sum.

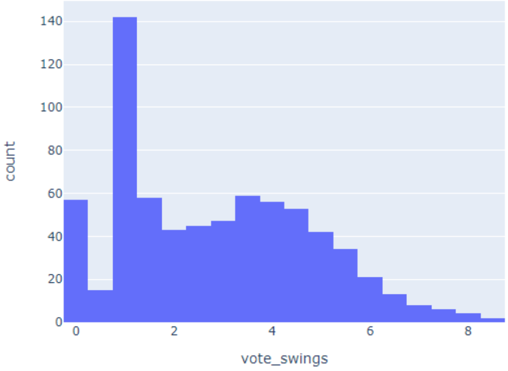

Figure 5: Distribution of conversations by vote swings.

The second feature is vote swings which approximates to what extent the vote scores of the comments swing up and down throughout the conversation snippets. Like the vote sum feature, the logarithms of the vote swing values are computed before the values are given to the regression model. Before this step, the lowest vote swing value among the conversations is 0, which indicates steadily decaying comment scores, and the highest value is 4393. The median value of vote swings is 12.5 and the mean is 114. After computation of the logarithms, the values range from 0.00 to 8.39 with a median of 2.60 and a mean of 2.83. The distribution of conversations by vote swing values is visualized in Table 5 above.

The third feature is toxicity which is the average of the confidence score of the toxicity detection applied to the comments in a conversation. The highest toxicity score of a conversation in our data is 0.81 and the lowest 0.04. The median toxicity score is 0.22 and the mean is 0.24. This suggest a more balanced distribution of the feature values, although only a smaller share of conversations has relatively high toxicity scores. Again, the distribution of the conversations by their average toxicity score is visualized in a histogram.

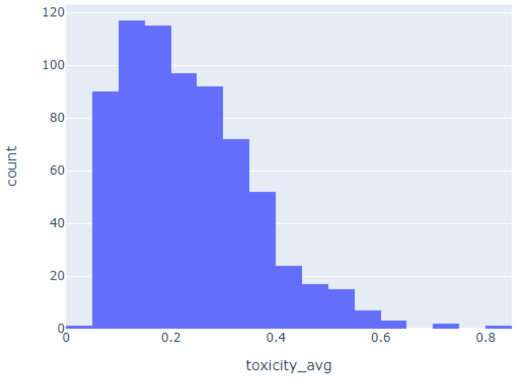

Figure 6: Distribution of conversations by toxicity.

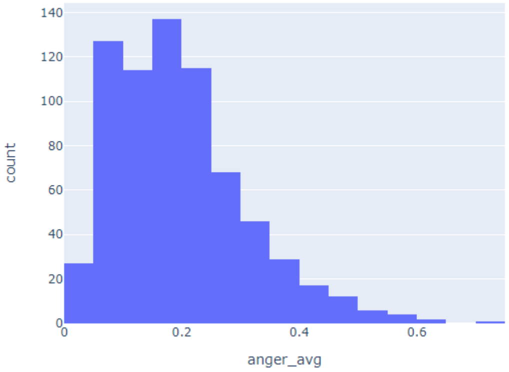

Figure 7: Distribution of conversations by anger.

Turning now to anger, which is the fourth feature in our data set. The anger features is the average of the probability of anger assigned to each comment in a conversation by the emotion detection model. The highest value assigned to a conversation snippet is 0.71 and the lowest is 0.02. The median is 0.18 and the mean is 0.20. The distribution of the values is similar to that of toxicity but with fewer high values. Figure 7 above visualized the distribution of average anger scores among the conversation snippets.

The last NLP tool applied to the collected conversations is sentiment analysis. The minimum value is 0.0008 and the maximum value encountered in the data is 0.86. The median value of positive sentiment is 0.17 and the mean is 0.21. The average of the conversations cluster around certain values as can be observed in Figure 8. This is due to the fact that most comments are labelled as negative and that the confidence score of the sentiment analysis is typically high. To clarify, when a comment is classified as positive, the confidence score is close to 1 and when a comment is labelled as negative, it is close to 0. Since the positivity scores are obtained by adding the confidence scores and dividing them by the number of comments, we obtain this distribution of feature values.

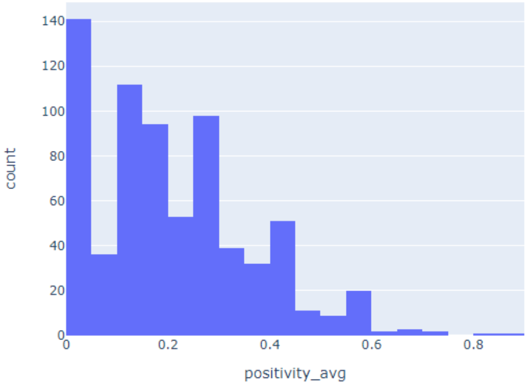

Figure 8: Distribution of conversations by positivity.

After an inspection of the distribution of the vote measures and the results of the NLP tools applied to the conversations, it is interesting to examine their distribution across the different categories of conversations. A majority (60.40%) of the annotated conversations are judged to contain out-group derogation according to the annotation guidelines. Disagreement, as defined in the codebook, is judged to be even more frequent in the annotated conversations and is present in 83.10% of them. While the logistic regression model is more suitable to find correlations between the values of the features and the annotation results, we inspect the median values of the features of these categories of conversation in the tables below.

Table 5 lists the median values of the features in conversations that contain out-group derogation and compare them to the conversations that do not contain out-group derogation. We can observe that the value of the vote sum feature is similar across the categories. The value of vote swings is slightly higher among conversations that contain out-group derogation compared to conversations that do not. The medians of the average confidence scores for toxicity and anger are also higher in conversations with out-group derogation. Regarding positive sentiment, the median of the average scores are exactly the same across the categories.

Feature	**Out-group derogation (60%)**	**NOT Out-group derogation (40%)**
Vote sum	5.35	**5.43**
Vote swings	**2.84**	2.15
Toxicity	**0.27**	0.15
Anger	**0.20**	0.15
Positivity	0.17	0.17

Table 5: Median feature values in conversations with out-group derogation compared to conversations without out-group derogation.

Feature	**Disagreement (83%)**	**NOT Disagreement (17%)**
Vote sum	5.22	**6.21**
Vote swings	**3.25**	0.88
Toxicity	0.21	**0.26**
Anger	**0.18**	0.17
Positivity	0.16	**0.26**

Table 6: Median feature values in conversations with disagreement compared to conversations without disagreement.

In Table 6, the median values of the features are compared between the group of conversations that contain disagreement and the conversations that do not contain disagreement. In this comparison, there is a larger difference between the median values of vote sum. The difference in vote swings is even greater which suggests that the feature is a good indication of disagreement. Quite surprisingly, the median of the toxicity feature is slightly higher in conversations without disagreement but the median values of anger are similar across the categories. Less surprising is that the median of the positivity feature is higher in conversations without any disagreement.

Feature	**Derogation AND Disagreement (52%)**	**Derogation AND NOT Disagreement (9%)**
Vote sum	5.18	**6.49**
Vote swings	**3.41**	0.83
Toxicity	0.26	**0.30**
Anger	0.20	**0.21**
Positivity	0.17	**0.28**

Table 7: Median feature values in conversations with out-group derogation AND disagreement compared to conversations with out-group derogation AND NOT disagreement.

Table 7 displays a comparison of median feature values between conversations that contain out-group derogation but not disagreement and conversations with both out-group derogation and disagreement. The results are similar to the ones presented in Table 6 above and the category of conversations that do not contain disagreement have a higher median toxicity score as well as a higher median value of positivity. As in Table 6, the median of the anger feature is similar across the categories. However, these data must be interpreted with caution because of the small size of the category of conversations with derogation and without disagreement. This category only contains 9% of the total number of conversations.

5.2 Logistic Regression Results

In the following pages, the results of the logistic regression models applied to the collected and annotated data will be presented. As was mentioned in the previous chapter, models with four different sets of independent variables are fitted to estimate the probabilities that a conversation belongs to one of the four categories: out-group derogation, disagreement, out-group derogation AND disagreement and out-group derogation AND NOT disagreement. This results in 16 fitted and evaluated logistic regression models. First, we evaluate the predictive power of the 16 models on unseen data in terms of accuracy, precision, recall and F_1 and the significance of the cross-model differences is reported. Second, we present and inspect the coefficients of the same 16 models and their statistical significance.

5.2.1 Model Evaluation

The table below presents the performance of the 16 models on the test data. To perform classification, the prediction function with its estimated intercept and coefficients is applied to the features of the conversations in the test set, resulting in probabilities between 0 and 1. To allow comparison between the models' predictions and the annotation, values equal to or greater than 0.5 are rounded to 1 and lower values to 0. Next, the performance of the fitted models is evaluated in terms of accuracy, precision, recall and F_1. The table allows us to compare the performance both across models and across tasks. By comparing the models' performance on the same task, we gain insight on how well the features they are fitted on represent the corresponding concept. On the other hand, through comparison of the models' performance across the conversation categories, we can evaluate if some categories are more challenging than others to model with the current experimental setup.

Model	Accuracy	Precision	Recall	F_1	Baseline F_1	Split
5 Features						
derogation	75.18%	72.92%	88.61%	80.00%	71.82%	2
disagreement	87.94%	90.77%	95.93%	93.28%	93.18%	5
derogation AND disagreement	72.34%	69.01%	74.24%	71.53%	69.44%	2
derogation AND NOT disagreement	92.91%	80.00%	30.77%	44.44%	95.17%	1
Feature Ablation						
derogation	74.47%	72.63%	87.34%	79.31%	71.82%	2
disagreement	87.23%	87.23%	100.00%	93.18%	93.18%	5
derogation AND disagreement	64.54%	64.86%	66.67%	65.75%	67.61%	5
derogation AND NOT disagreement	90.78%	0.00%	0.00%	0.00%	95.17%	1
BERT Embeddings						
out-group derogation	77.30%	82.29%	84.04%	83.16%	80.00%	3
disagreement	89.36%	90.30%	98.37%	94.16%	93.18%	5
derogation AND disagreement	75.18%	71.23%	78.79%	74.82%	69.44%	2
derogation AND NOT disagreement	92.2%	75.00%	23.08%	35.29%	95.17%	1
All Features						
out-group derogation	**82.98%**	80.90%	91.14%	**85.71%**	71.82%	2
disagreement	**92.91%**	93.80%	98.37%	**96.03%**	93.18%	5
derogation AND disagreement	**81.56%**	77.78%	84.85%	**81.16%**	69.44%	2
derogation AND NOT disagreement	**93.62%**	75.00%	46.15%	**57.14%**	95.17%	1

Table 8: Evaluation metrics of the logistic regression models on test data. For each task and model, the best split in terms of F_1 is presented. The highest accuracy and F_1 score for each task is marked in bold. An F_1 score below or equal to the majority group baseline calculated on the test set is marked with red.

In the table above, the four kinds of model are named 5 Features, Feature Ablation, BERT Embeddings and All Features. These models are identical in architecture but differ in terms of the set of features they are fitted on and have access to during prediction. The model named 5 Features is fitted on the features vote sum and vote swings as well as toxicity, anger and positivity score. The model called Feature Ablation has toxicity score as its only feature. The model labelled BERT Embeddings is fitted on the 10 dimensions of the reduced BERT embeddings only while the model named All Features has access to the features of the 5 feature model as well as the 10 dimensions of the reduced BERT embeddings.

The models evaluated in the table above are fitted on the train and test splits that result in the highest F_1 score. To determine the best split, we apply 5-fold cross-validation and the number of the split is reported in the rightmost column of the table. To allow meaningful evaluation of the F_1 scores obtained by the models, a baseline value is presented. The baseline scores are calculated on the test data of the corresponding splits and represent the values that would be obtained if the models always predict the most common outcome. Ideally, the F_1 scores of the models should notably exceed this baseline. However, on the tasks that are imbalanced (e.g., derogation AND NOT disagreement), results that slightly exceed the baseline are acceptable.

In the case of out-group derogation, all four models obtain F_1 scores that are higher than the baseline. The fourth model, which has access to all features in the data set, has the highest F_1 of 85.71%. This is notably higher than the baseline of 71.82%. The same model also has the best performance in terms of accuracy and obtains a score of 82.98%. This can be compared with the model with toxicity as its only feature which achieves the lowest accuracy of 74.47% as well as the lowest F_1 of 79.31%. The data reported here appear to support the assumption that word embeddings, despite the radical reduction of dimensions, provide information that benefits accurate prediction of out-group derogation in the conversations. However, the model that is fitted on BERT embeddings exclusively does not perform as well on this task as the all feature model. An additional interesting observation is that the performance of the model with toxicity as its only feature is not much lower than the 5-feature model or even the model fitted on BERT embeddings.

Let us now turn to the performance of the models on the task of predicting disagreement in the conversation snippets. Again, the best performing model in terms of accuracy and F_1 score is the model fitted on all features, but the performance on this task is more even across the models with F_1 scores ranging from 93.18% to 96.03%. However, the baseline F_1 is also high due to the imbalance of this conversation category. The model with access to toxicity predicts that all conversations in the test set contain disagreement and consequently does not exceed the baseline. Similarly, the five feature only slightly performs better than the baseline. In other words, the presence of word embeddings in the data set seems to marginally improve detection of disagreement as well.

Regarding the category of conversations that contain both out-group derogation and disagreement, similar patterns can be observed. The results suggest that the models with access to embeddings of the linguistic features perform better. By contrast, the model with access to all features obtains a substantially higher F_1 score of 81.16% than the model fitted on BERT embeddings exclusively. Moreover, the model trained on toxicity alone performs below the baseline and the other models, both in terms of accuracy and F_1. On the task of predicting out-group derogation and disagreement combined, it seems to be particularly beneficial to fit the models on all features in the data.

With respect to the last category of conversations which contain out-group derogation but not disagreement, no model performs better than the baseline. This category is highly imbalanced and baseline F_1 is high, just like the category of disagreement. Despite this, the best performing model obtains an F_1 score of only 57.14%. Again, the most successful model is the one with access to all features but all models struggle with low recall. It seems possible that the poor performance across the models is due to the class imbalance, but it may also be the case that this category is especially challenging to model with the available data.

To summarize the results presented this far, embeddings of the linguistic information of the conversations seem to improve prediction of out-group derogation and disagreement as well as their combinations. The models with access to such embeddings outperform the other models on all tasks and obtain F_1 scores higher than the baseline for three out of the four conversation categories. The results also suggest that while toxicity alone can predict out-group derogation, it is not sufficient for successful prediction on the other tasks as it performs worse or equal to the baseline. In general, all models perform well on the task of predicting out-group derogation, in relation to the baseline.

Having presented and compared the results of the models on test data, we will now move on to examine the significance of the cross-model differences. To achieve this, we apply 20-fold cross-validation to generate 20 F_1 scores for each of the 16 models. These results are compared pairwise according to the Wilcoxon signed-rank procedure. The test was first suggested by Wilcoxon (1945) to test the null hypothesis that two related samples come from the same distribution and is a suitable method for comparing the performance of machine learning models (Demšar, 2006). We use the SciPy (Virtanen et al., 2020) implementation of the test and present the results in Table 9 below.

Model 1	**Out-group Derogation**	**Disagreement**
Model 2	p	p
5 Features		
Feature Ablation	0.1769	0.6742
BERT Embeddings	0.0263*	0.0152*
All Features	0.0034**	0.0008***
Feature Ablation		
BERT Embeddings	0.0136*	0.0130*
All Features	0.0014**	0.0054**
BERT Embeddings		
All Features	0.7285	0.9702
Model 1	**Derogation AND Disagreement**	**Derogation AND NOT Disagreement**
Model 2	p	p
5 Features		
Feature Ablation	0.0362*	0.0061**
BERT Embeddings	0.0702	0.8651
All Features	0.0048**	0.1047
Feature Ablation		
BERT Embeddings	0.0034**	0.0115*
All Features	0.0005***	0.0011**
BERT Embeddings		
All Features	0.0826	0.0867

Table 9: Pairwise comparison of the differences across models in terms of F_1 applying 20-fold cross-validation and Wilcoxon signed-rank test. Values of p lower than 0.05 suggest rejection of the null hypothesis.

The results of the Wilcoxon signed-rank test discard some of the cross-model differences observed in Table 8 as non-significant. For instance, there is no statistically significant difference between the performance of the model fitted on word embeddings only and the model with access to all features. Likewise, no statistical difference is encountered between the 5-feature model and the model with toxicity as its only feature on the two conversation categories out-group derogation and disagreement. In contrast, there are significant differences between the models with and without word embeddings on these tasks. Furthermore, the test suggests significant differences between the toxicity only model and the word embedding models on all tasks. These findings give us an indication of which features in our data benefit successful prediction. To explore this more thoroughly, the following section proceeds to examine the significance on the coefficients of each model individually.

5.2.2 Feature Evaluation

So far this chapter has focused on the evaluation of the logistic regression models on test data. The following section will focus on the evaluation of the individual features. In Table 10 below, the individual features of the four kinds of model fitted on the data to predict out-group derogation and disagreement are presented. For each feature, its coefficient is reported along with standard error and p-value. The coefficients represent the change in log-odds that would occur when the value of the corresponding features increase with 1. The p-values indicates the significance of the coefficients and standard error indicates the uncertainty of the coefficients. This value is high when two features are highly correlated and it is hard to determine their individual contribution to the outcome.

The first feature we will examine is vote sum. This feature is present in the 5-feature model and the model with all features, but its coefficient shows no statistical significance in any of the models it occurs in. This is true for the task of predicting out-group derogation as well as disagreement. In all models the feature occurs in, its coefficients are estimated to very small numbers while the p-values are high. To clarify, the logarithm of the sum of votes in the conversation snippets has very little impact on the probability that the conversation contains out-group derogation or disagreement, predicted by the model.

The second feature to be inspected is vote swings which is present in the 5-feature model and the model fitted on all features. The correlation between the value of vote swings and the probability that a conversation contains out-group derogation or disagreement is statistically significant. Considering disagreement, the p-value associated with the vote swing coefficient is below 0.001 and consequently very likely to influence the outcome of the model. One unit's increase in vote swings results in an increase of 1.0261 in the log-odds that a conversation snippet contains disagreement when the other feature values remain the same. This suggests that the vote swing feature is a suitable predictor of disagreement in the conversation snippets.

The third feature is toxicity which represents the value of the average toxicity of the comments of a conversation. This feature is present in all models except the one with BERT embeddings only. The coefficients of the toxicity feature have a statistically significant correlation with out-group derogation in all models it occurs in, with p-values below 0.001. However, there is no statistically significant correlation between the toxicity feature and disagreement in any of the models. Despite the strong correlation between the coefficients of the toxicity variable and out-group derogation, the standard error of the coefficients are high in all models and range from 0.984 to 1.292. A possible cause is the high correlation with the anger feature.

The forth feature in our data is anger which represents the average probability that a conversation snippet expresses angry emotion. Due to the correlation between values of anger and toxicity in the conversations, the anger feature is assigned high values of standard error as well. This is true for all models in which the feature occurs and for both conversation categories. The coefficients of the anger feature are also associated with high values of p in the models that predict disagreement. However, the anger coefficient in the 5-feature model is statistically significant, but not in the model with access to all features.

The fifth feature to be evaluated is the positivity feature which defines the confidence score of positive sentiment averaged across the comments in the conversations. Like anger, the coefficient of the positivity feature is only statistically significant in one model, namely the 5-feature model that predicts disagreement. The fitted regression models show no statistical correlation between the coefficients of the positivity feature and out-group derogation. In fact, the highest p-value presented in the table is assigned to the positivity coefficient in the all feature model. Moreover, the standard error is relatively high for positivity as well, with values ranging from 0.600 to 0.996 which suggests confusion with other features.

	Out-group derogation			**Disagreement**		
Model	Coef	Std Err	p	Coef	Std Err	p
5 Features						
intercept	-2.2124***	0.463	0.000	0.5241	0.608	0.389
vote sum	-0.0070	0.054	0.897	-0.0554	0.070	0.425
vote swings	0.1230*	0.051	0.017	1.0261***	0.129	0.000
toxicity	8.0888***	1.027	0.000	-1.2820	1.049	0.222
anger	2.4054*	0.963	0.012	0.5966	1.282	0.642
positivity	0.6455	0.600	0.282	-1.6041*	0.777	0.039
Feature Ablation						
intercept	-1.3610***	0.221	0.000	1.7916***	0.235	0.000
toxicity	8.3546***	0.984	0.000	-1.1149	0.843	0.186
BERT Embeddings						
intercept	0.4938***	0.105	0.000	2.1263***	0.165	0.000
dim 1	-1.0484***	0.143	0.000	0.8699***	0.164	0.000
dim 2	1.0308***	0.154	0.000	0.7947***	0.196	0.000
dim 3	0.0130	0.160	0.935	0.2620	0.207	0.205
dim 4	-0.6544***	0.179	0.000	-1.5136***	0.239	0.000
dim 5	-0.7675***	0.202	0.000	0.1130	0.247	0.648
dim 6	-0.3188	0.208	0.126	-1.2736***	0.267	0.000
dim 7	-1.2305***	0.244	0.000	0.3309	0.307	0.281
dim 8	1.4366***	0.255	0.000	-0.0440	0.293	0.881
dim 9	0.1044	0.259	0.686	0.7779*	0.321	0.015
dim 10	0.8710**	0.271	0.001	0.1557	0.334	0.641
All Features						
intercept	-0.9971	0.572	0.081	-0.3697	0.841	0.660
vote sum	0.0436	0.062	0.481	0.0446	0.085	0.602
vote swings	0.0910	0.058	0.117	1.0063***	0.139	0.000
toxicity	5.9035***	1.292	0.000	1.4781	1.631	0.365
anger	-1.0970	1.184	0.354	-1.1928	1.652	0.470
positivity	-0.0306	0.712	0.966	0.6079	0.996	0.542
dim 1	-0.6999***	0.154	0.000	1.1281***	0.231	0.000
dim 2	0.6496***	0.165	0.000	0.5586*	0.240	0.020
dim 3	0.2222	0.178	0.211	0.1132	0.259	0.662
dim 4	-0.6982**	0.203	0.001	-1.2792***	0.283	0.000
dim 5	-0.6160**	0.203	0.002	0.0979	0.296	0.741
dim 6	-0.3688	0.220	0.094	-0.6788*	0.309	0.028
dim 7	-0.8960***	0.252	0.000	0.2200	0.356	0.537
dim 8	0.8073**	0.280	0.004	0.1529	0.405	0.706
dim 9	-0.0735	0.265	0.782	0.6519	0.373	0.081
dim 10	0.8849**	0.276	0.001	0.1231	0.398	0.757

Table 10: Feature evaluation for out-group derogation and disagreement on the best performing models for each task considering F_1. The asterisks next to the coefficients indicate their statistical significance. One asterisk means p<0.05, two mean p<0.01 and three mean p<0.001.

The last features in our data are the individual dimensions of the reduced word embeddings. In comparison to the features presented this far, they are not as transparent and explainable. Their contribution to the model predictions is therefore difficult to connect with knowledge about affective polarization in online debates. Nonetheless, the results in Table 10 indicate that all embedding dimensions except for 3 and 9 have a statistically significant association with the probabilities of out-group derogation. This is true both for the model fitted on word embeddings only and the model with access to all features. However, the coefficients of the reduced word embeddings do not correlate to the same extent with the prediction of disagreement. Only four of the dimensions are significantly associated with the outcome of the model trained on word embeddings only, and two dimensions considering the model with all features.

Let us now turn to the remaining conversation categories and their relation with the individual features. Table 11 below presents the four kinds of model fitted on the data to predict out-group derogation AND disagreement as well as out-group derogation AND NOT disagreement. The individual features are the same as in Table 10 and for each feature, its coefficient is reported along with standard error and p-value. The results are generally similar to the ones in Table 10 above, but fewer features are statistically correlated with the outcome when it comes to predicting the combinations of out-group derogation and disagreement.

Regarding the vote sum feature, its coefficients have no statistically significant correlation with the prediction of conversation categories in any of the models it occurs in. These results are in line with the ones presented in Table 10 and suggests that the feature consequently has no impact on the classification of the conversation snippets. On the contrary, the estimated coefficients of the vote swing feature are significantly associated with both of the tasks in Table 11 above. A likely explanation is that the values of the feature can indicate absence or presence of disagreement, which is what separates the two conversation categories from each other.

The coefficients of the toxicity feature are significantly associated with the predictions of the conversation category derogation AND disagreement in all models the feature occurs in. The feature also contributes to the prediction of derogation AND NOT disagreement in the 5-feature model and the toxicity model. Despite this, the coefficient of the toxicity feature likely has no impact of the probability of derogation AND NOT disagreement in the model that is fitted on all features. This inconsistency may be due to the difficulty of predicting this particular conversation category and the relatively few occurrences of it in our data set. In analogy with the results in Table 10, the values of standard error are high for the toxicity feature in all models.

The anger feature shows no significant association with the outcome of any of the models when fitted to predict the combinations of out-group derogation and disagreement in the conversations. The standard error values are high for the anger feature as well, reaching 2.160 in the all feature model that predicts the presence of out-group derogation and absence of disagreement. The results of the positivity coefficients are similar, but suggest statistically significant correlation with the model outcome on the task of predicting derogation AND NOT disagreement. A probable explanation is that the median positivity score among conversations with derogation and without disagreement is notably higher than in conversations with derogation AND disagreement, as stated in Table 6. However, these results need to be interpreted with caution due to the imbalance of the conversation category and the poor performance across the models.

In consideration of the individual dimensions of the reduced word embeddings, the results suggest that they are useful for predicting the combinations of out-group derogation and disagreement. In the models that predict derogation AND disagreement, all dimensions except 3 and 9 are significantly associated with the outcome. When it comes to predictions on the imbalanced class of derogation without disagreement, fewer dimensions have coefficients that are likely to impact the model outcome. Only the coefficients of dimensions 1 and 6 are correlated with the prediction of this category in a statistically significant manner in the model fitted on all features. The model fitted on word embeddings exclusively have three additional significant coefficients.

	Derogation AND Disagreement			**Derogation AND NOT Disagreement**		
Model	Coef	Std Err	p	Coef	Std Err	p
5 Features						
intercept	-1.9268***	0.444	0.000	-3.1693***	0.857	0.000
vote sum	-0.0788	0.052	0.130	0.0823	0.090	0.361
vote swings	0.3137***	0.052	0.000	-1.0980***	0.192	0.000
toxicity	5.5291***	0.891	0.000	5.0748***	1.360	0.000
anger	1.7313	0.900	0.054	1.1856	1.660	0.475
positivity	-0.1115	0.584	0.849	2.3987*	1.056	0.023
Feature Ablation						
intercept	-1.2059***	0.202	0.000	-3.5284***	0.351	0.000
toxicity	5.5726***	0.811	0.000	4.3698***	1.074	0.000
Bert Embeddings						
intercept	0.1183	0.096	0.216	-3.2201***	0.262	0.000
dim 1	-0.4844***	0.121	0.000	-1.1039***	0.241	0.000
dim 2	0.7076***	0.137	0.000	0.1327	0.275	0.630
dim 3	0.1820	0.153	0.234	-0.0714	0.284	0.802
dim 4	-0.7801***	0.170	0.000	1.0619***	0.289	0.000
dim 5	-0.5631**	0.185	0.002	-0.4326	0.343	0.207
dim 6	-0.8857***	0.200	0.000	1.5903***	0.361	0.000
dim 7	-0.8350***	0.229	0.000	-0.4041	0.432	0.350
dim 8	0.9152***	0.232	0.000	1.0633*	0.413	0.010
dim 9	0.2859	0.234	0.221	-1.0351*	0.438	0.018
dim 10	1.0131***	0.253	0.000	-0.3663	0.460	0.426
All Features						
intercept	-1.3729*	0.551	0.013	-1.9123	1.080	0.077
vote sum	-0.0141	0.059	0.811	-0.0150	0.105	0.886
vote swings	0.2906***	0.057	0.000	-1.1049***	0.221	0.000
toxicity	4.2795***	1.184	0.000	2.2691	1.955	0.246
anger	-1.0725	1.104	0.331	0.0945	2.160	0.965
positivity	-0.0640	0.685	0.926	0.3659	1.278	0.775
dim 1	-0.3178*	0.146	0.029	-1.0935***	0.310	0.000
dim 2	0.5593***	0.158	0.000	0.4327	0.341	0.204
dim 3	0.1973	0.169	0.242	0.3259	0.340	0.338
dim 4	-0.7846***	0.197	0.000	0.4492	0.323	0.165
dim 5	-0.5675**	0.194	0.003	-0.3978	0.389	0.307
dim 6	-0.6297**	0.215	0.003	1.3067**	0.404	0.001
dim 7	-0.7574**	0.242	0.002	-0.4841	0.502	0.335
dim 8	0.6101*	0.267	0.023	0.4970	0.540	0.357
dim 9	0.2274	0.251	0.364	-0.8204	0.497	0.099
dim 10	1.0423***	0.266	0.000	-0.0465	0.518	0.928

Table 11: Feature evaluation for derogation AND disagreement and derogation AND NOT disagreement on the best performing models for each task considering F_1. The asterisks next to the coefficients indicate their statistical significance. One asterisk means p<0.05, two mean p<0.01 and three mean p<0.001.

6 Discussion of Results

In the following pages, we will discuss the results obtained in this study in relation to the theory about affective polarization accounted for in Chapter 2. We will also evaluate the experimental setup and make suggestions for future computational research in the area of detecting affective polarization in online discussions. In the final section of the chapter, we will discuss challenges in the annotation process and propose solutions that would be adapted in a hypothetical additional annotation round.

To summarize the features we choose to numerically represent the conversation snippets, they are suitable to detect behaviour related to affective polarization to a varying degree. Considering the all feature model, which obtains the best performance when predicting conversation categories of unseen conversation snippets, toxicity, vote swings and word embeddings are among the more successful features. Toxicity plays a significant role in the prediction of out-group derogation, the vote swing feature is significantly associated with the prediction of disagreement and the word embeddings are related to both concepts as well as their combinations. The results suggest that especially dimensions 1 and 6 are related to the predicted probability of the conversation categories. Among the less suitable features are vote sum, anger and positive sentiment. Dimensions 3 and 9 of the reduced conversation embeddings are also among the features that rarely have statistical significance with the model outcomes.

The findings of this study suggest that even heavily reduced embeddings of the conversations improve the predictive power of the statistical models. One possible implication of this is that advanced language models are suitable to detect behaviour related to affective polarization, even on a smaller data set. Moreover, these results appear to support the assumption that the linguistic information of the conversations, even in the form of reduced embeddings, captures aspects of out-group derogation that an NLP tool such as toxicity detection does not. In fact, the non-significant differences between the models fitted on word embeddings exclusively and the models with embeddings as well as toxicity and vote swings may imply that the concepts of toxicity and disagreement are implicitly available in the reduced embeddings.

However, there are problems related to the employment of embeddings in a theory sensitive task like this. Most importantly, the use of embeddings interfere with the transparency and explainability of the results. The fact that dimension 1 and 6 are helpful for classification, but not dimension 3 and 9 does not lead to any insights about the characteristics of affective polarization in online discussions. In contrast, examining how levels of toxicity or distribution of votes correlate with expressions of affective polarization can increase our understanding of the behaviour. A further limitation of word embeddings in this context is that they are not context independent. The estimated prediction functions of the regression models fitted on word embeddings may not be applicable to conversations about other topics, which is not the case with the toxicity models. However, the toxicity models still inherit the documented biases from the toxicity classifier (see Toxicity Detection).

While embeddings improve performance of our models, the 5 feature model and the toxicity only model can be applied to perform decent predictions of out-group derogation, which is the most central conversation category for our research aims. Both models obtain F_1 scores notably above the baseline. The F_1 score of the model with toxicity as its only feature is 79.31%, compared to 85.71% of the model with access to all features. This suggests that the toxicity score alone is a good predictor of out-group derogation and that decent prediction with toxicity only is possible. This would result in a much simpler model that allows us to better interpret and explain its predictions. Another important benefit of using understandable features such as toxicity is that we get a topic-independent classifier. In other words, the prediction function can be applied to Reddit discussions of other topics without being triggered by e.g., the frequency or absence of certain terms that may be related to affective polarization in the context of covid-19.

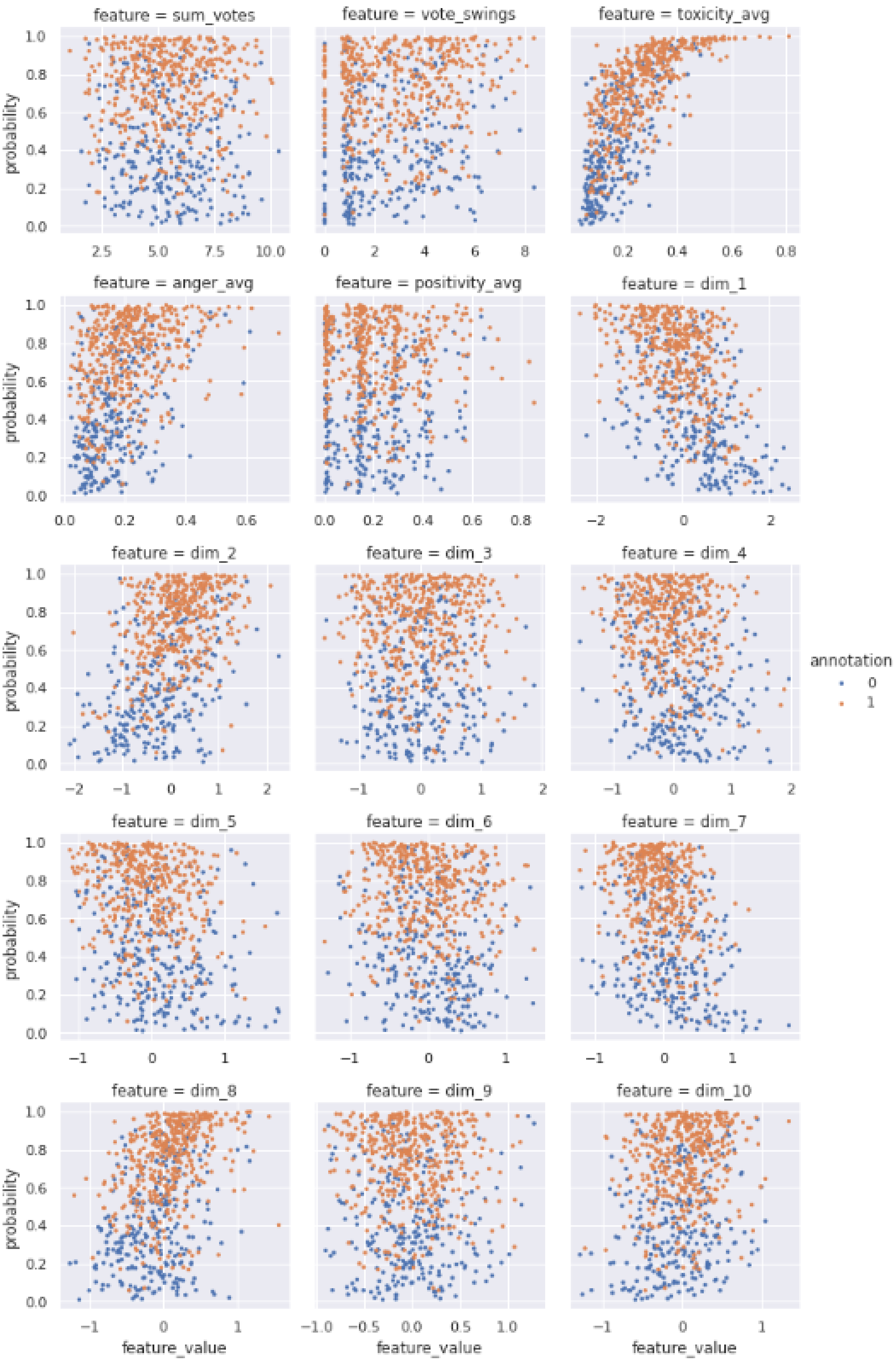

Figure 9: Predicted probabilities of out-group derogation on the complete data set, visualized by feature values. Orange dot indicates out-group derogation according to the annotation and probabilities above 0.5 are predicted by the model as out-group derogation.

On the question of the combinations of out-group derogation with disagreement, these categories seem particularly hard to model with our annotation and selection of features. For instance, all models perform better on the task of predicting out-group derogation than predicting the combined category of both derogation and disagreement. There are a few possible explanations for why these categories are harder to model statistically in our experimental setup. Most importantly, our hypothesis that out-group derogation would be expressed differently depending on the presence of disagreement in the conversation snippet could be wrong, or at least exaggerated. The opposite would however also be possible, namely that the assumption is correct but that the concepts are not represented to the same extent by our selection of features. A further possible explanation is that the relation between the features and the conversation categories are too complex for a logistic regression model to determine. As stated above, the product of features and their coefficients are added together to obtain the probability (See 3.4 Logistic Regression). To clarify, the regression model treats the values of the independent variables individually and is unable to distinguish between combinations of feature values.

Relating the results to the literature on affective polarization raises additional questions. As indicated previously, in-group praise is not expressed in our data to the extent suggested by previous work, nor is it as common as out-group derogation. A possible explanation for this is that in-group praise is expressed in more indirect manners than out-group derogation and that our annotation guidelines are unable do identify them. The more probable reason for the lack of in-group praise in our data is the fact that discussion participants online are aware of their audience and take it into account when posting their contributions (Klein et al., 2007). Especially in an open forum such as Reddit, it is fair to assume that users consider that others who might disagree with them can read and respond to their comments. In that sense, members of the out-group are always present which may explain the unexpected lack of in-group praise encountered in the collected data.

An important finding that confirms theory on affective polarization and identity performance is that out-group derogation is common even in conversations where everyone agrees. Among the 119 conversations that do not contain disagreement, 61 still contain out-group derogation. This is especially interesting as the annotation guidelines have a very low tolerance for disagreement. This confirms theories about how identity performance is crucial for explaining affective polarization (Iyengar et al., 2019), an important aspect that is often overlooked in attempts to automatically detect and quantify affective polarization in online discussions. While these results are not controversial in that sense, it is important to bear in mind that our data set is small when drawing conclusions of this kind. Below is an example of a conversation that contains out-group derogation, but not disagreement.

1025, sqyk89, "Green" from Indiana refused to live in fear and thought the virus was a joke, and he loved a good joke. He's now in the hospital with Covid pneumonia, still waiting for the punch line.

1025.0, hwohlqw, user_a

These people are so incredibly triggered by people wearing masks in their cars.

1025.1, hwoqv14, user_b

It's funny because, like, when I go out for errands I'm doing multiple stops at once.
And in between the stops where I'm around tons of people and touching tons of things, I keep my mask on because I don't think putting my grubby hands up around my face is a good idea.
And yet they think their clever.

1025.2, hwot2jp, user_a

Also this!
When I do the weekly shopping I have a three-store circuit.
I'm just going to leave the damn thing on rather than taking it on and off.
Also, who even cares?
It doesn't affect them at all but they just can't help being hysterical snowflakes.

1025.3, hwoxgp0, user_c

Right?!
It's as much business of theirs and affects them just as much as the kind of underwear other people wear.
Of course they think that's their business too because they are obsessed with other people's genitals

1025.4, hwp4xew, user_d

And other people's hair!

OMG a man with long hair!
A woman with short hair!
Anybody with pink/purple/blue/green hair!
THE WORLD IS ENDING, WHAT HAS HAPPENED TO AMERICA??

1025.5, hwp5ps4, user_c
Right?!
If only they could experience how much more pleasant life can be when you're not judging every person for every damn thing, especially things that are harmless and/or have nothing whatsoever to do with them!

1025.6, hwpbptz, user_d
If the world isn't catering to their every whim, they're BEING OPPRESSED!!!
And yes, having been raised by such, they ARE miserable assholes who aren't happy unless they're complaining about something or kicking someone else.
(but we're the snowflakes, right....)

Regarding the strong correlation between high toxicity score and out-group derogation, there is a concern that conversations with high toxicity may incorrectly be predicted to contain out-group derogation. When we inspect the predictions of the all feature model on the category out-group derogation, it is clear that many true positives have high toxicity scores which makes the reliability of the applied toxicity classifier important. Nevertheless, considering the models with multiple features, the embeddings also play an important role in the prediction and there are true positives with very low toxicity scores around and even below 0.1. The highest toxicity score among false positives is 0.47. That conversation is listed below and serves as an example of why the toxicity classifier is not always reliable and how that may have consequences for the performance of the model.

549, ss7rnp, first months of corona lockdown messed me up pretty bad. made this during that time. i'm sure many of you can relate

549.0, hwwerng, user_a
Better then most of the shit I've heard

549.1, hwwexiz, user_a
The pandemic has me messed up period...
The second year was harder, now this year I'm desparatly trying to gain stability.

549.2, hwwfy4h, user_b
so sorry to hear that mate..it's been really hard on me as well, my whole life fell apart... but i'm on my way up again.
wish you all the best, we got this!

549.3, hwwge68, user_a
By the way I do music to.... My goal to survive long enough to do hopefully one more album

549.4, hwwgsv1, user_b
in that case i hope it takes you 60+ years to make it!
are you getting help from anyone?
i put myself in a psych ward when things got unbareable for me last year.
hardest and best decision, truly saved my ass..

549.5, hwwhcse, user_a
Unfortunately my community wants me dead....

549.6, hwwl12k, user_b
don't give them the satisfaction! all the best to you, really.
things can go from good to shit real fast.
but they can also go from shit to good at the same speed!

For the reader, it is obvious that the comments in the conversation snippet above are not toxic. On the contrary, user_b expresses solidarity and support to user_a. It is plausible that the conversation still obtains a high average toxicity score due to the presence of certain words and the gravity of the discussion topic.

6.1 Suggestions for Improved IAA

In the following section, we will discuss the low inter-annotator agreement obtained during coding of the conversation snippets and a few suggestions for a hypothetical third annotation round. While low IAA is an issue for reliability of the results as well as reproducibility of the study, annotation that is not consistent enough will also make the data more difficult to model statistically. If the data is not annotated correctly and consistently, important patterns and correlations in the data may not be detected as desired. Due to the limited scope of this study, we solve this particular problem by only considering annotation from one annotator when fitting the logistic regression models. It is our hope that this strategy reduces unwanted variation in the judgement of the conversation snippets. However, we must consider the possibility that the low IAA suggests insufficiently elaborated annotation guidelines and if that is the case, there is a great risk that the annotation lacks consistency even though it is carried out by one individual.

Annotation of highly subjective tasks such as this is known to be challenging, partly due to the difficulty to foresee how concepts are referred to or expressed in interactive discourse, and this is especially true for sophisticated concepts such as toxicity (Baden et al., 2020). Despite this difficulty and the questioning of IAA thresholds on highly subjective tasks, we have a few suggestions for a hypothetical third annotation round. The examination and annotation of over 1000 conversation snippets and the evaluation of the IAA suggest that there are two main issues with the guidelines that would need to be addressed. First, the definitions of disagreement and out-group derogation need to be modified to better capture important aspects of affective polarization and group belonging. Second, the guidelines require further specificity and examples, especially for the trickier cases. Splitting up the concept of out-group derogation in smaller parts would be one way to approach this.

Disagreement in a conversation is not a necessary condition for it to contain expressions of affective polarization, but we annotate conversations for disagreement as a proxy for the presence of out-group members. However, the way disagreement is defined in the guidelines does not capture group belonging as intended. The idea is that when a participant expresses their views for or against a stance, we assume that they identify with other participants with similar opinions, but the current definition of disagreement is too generous and include any kind of disagreement in the conversation. This causes conversations with a general consensus on a topic related to the covid-19 pandemic to be annotated with disagreement, even when the disagreement regards a minor detail and not the stance. An updated codebook would therefore need a different definition of disagreement so that it more accurately represents the presence or absence of an out-group in the conversation snippets.

That is however not as easy as it may sound. The reason we opted for the more broad definition of disagreement is that stances on covid-related issues are related to other aspects, such as partisanship. To exemplify, we present a conversation snippet below in which the participants discuss the lasting impact on the sense of smell after a covid-19 infection. The participants agree on the gravity of the condition and that it needs to be taken seriously. An assumption is made that the people who do not take it seriously are conservatives and subsequently the conversation takes a turn to discuss negative traits of conservatives instead. One user comments on this and expresses their disagreement. This results in a conversation snippet that contains disagreement according to our current guidelines, but no disagreement on the issue related to the covid-19 pandemic.

292, sg908c, Like sewage and rotting flesh: Covid's lasting impact on taste and smell.

292.0, huv1q7e, user_a

Got Covid in June of 2020, to this day I still can't smell farts or poop (not necessarily a bad thing).
Also anything with a strong perfume smell like shampoo, soaps, men & women perfumes all smell like peanut oil.
So every time I take a shower all I smell is peanut oil...

292.1, huv786y, user_b

I haven't smelt anything since May 2020 and my taste is sensitive to salt and sugar.

Soda tastes like sewage.
I've lost 40 lbs because of it (good I suppose).
Last week I didn't realize I had my 1 year olds' shit on the underside of my sleeve.
Wife and kids and I took 30 minutes looking everywhere.
Because I was moving around the scent ended up in every room!
I sure do miss the pine scent when driving in the mountains, and also the smell of my wife.
Imagine if Covid made people go blind or deaf?

292.2, huvegfh, user_c

>Imagine if Covid made people go blind or deaf?

We'd still have people claiming it's fake, that all these people would just be pretending to be blind.

292.3, huvfr3v, user_b
My coworkers think I'm faking it.
Like when something is burning and they ask me to try and find the cause.
They roll their eyes.

292.4, huvxtlh, user_d
What do they think you are getting out of it?
I don't get conservatives that think the left is faking stuff.
Like, faking concern over global warming or faking being trans or faking school shootings.

292.5, huw7phy, user_e
Conservatives do it because they can imagine themselves doing it.
What they fail to understand is that no one would even consider it except other conservatives.
Being a conservative is a 24/7/365 life of projecting yourself and your shitty mind onto every around you.
Rant about pedophiles, rant about wearing mandates, rant about governments locking people up - all of it is them projecting what they want to do to others but feel that society prevents them.
That is the conservative way.

292.6, huxd7ui, user_f
I have tested positive for Covid three times in the past year.
Last time a month ago.
I am triple vaccinated.
I still have a lingering smell of burning wires that comes and goes.
I 100% know this is real and people die from it.
I lean conservative.
Please stop over-generalizing and putting all conservatives in the same bucket.

A further challenge of establishing group belonging in the conversation snippets is that there is not always a two-sided debate. According to our data, the participants of the conversations often reason about covid-related issues rather than debating them and many aspects of the issues are reviewed. The topics of the discussion also tend to change, even throughout these rather short fractions of conversation. Additionally, we observe a common behaviour of debating data and numbers among the participants which we believe can function as a way to express disagreement on a covid-related issue as well as distance themselves from one another. To clarify, facts and sources are frequently attacked or questioned by discussion participants instead of the conclusions that are drawn from them. These are examples of behaviour that would need to be accounted for in a third annotation round, given their frequency in the data as means of expressing stance and group belonging.

This leads us to the second main issue with the annotation guidelines which is the lack of specificity and examples. After completion of the annotation process, it is clear that many conversations are hard to evaluate and that detailed guidelines are necessary for the problematic cases. The difficulties are often related to the targets of the attacks and whether or not they can be considered an out-group. Quite frequently, governments, political parties, media and other authorities as well as their representatives are targets of the attacking that occurs in the conversations. In a hypothetical third annotation round, we would carefully evaluate if attacking of such organizations and individuals should qualify as out-group derogation. In an updated version of the codebook, only participants of the discussion and groups of people with a certain stance on a covid-related issue would be included. However, it is not always possible to decide whether e.g., a governmental organization is attacked because of their stance in relation to the covid-19 pandemic or because of their representation of an elite.

Lastly, updated annotation guidelines would need to establish clear boundaries between more and less direct attacks. For instance, there are clear differences in directness between ”that’s a stupid argument” and ”you are stupid”. Similarly, ”that’s not true” is different from ”you’re lying” and ”that’s misinformation” is different from ”you’re spreading misinformation”. A possible solution to this challenge is to only allow explicit attacks to qualify as out-group derogation. However, too ungenerous guidelines would risk to leave out the attacks that can be implied with comments such as “you obviously don’t understand statistics” or ”why do you keep deviating from the subject”. While it is possible to motivate both the inclusion and exclusion of such indirect attacks, it is important that the guidelines are clear to allow consistency in the annotation.

7 Conclusions

To sum up, we have explored a range of methods for detecting affective polarization in discussions on Reddit. We also propose a data set consisting of 705 conversation snippets that relate to the covid-19 pandemic. This data set might be of interest for future studies on affective polarization in social media discussions, and particularly in the context of the covid-19 pandemic. The collected discussions undoubtedly hold interesting phenomena related to affective polarization, intergroup conflict and identity performance, many of which were not possible to analyze within the scope of this study. Nevertheless, the experiments conducted on the data set can, at least to some extent, answer our proposed research questions.

(i) Can existing NLP tools predict affective polarization in discussions on Reddit?

According to our study, there is a strong correlation between the results of Perspective's toxicity detection model and expressions of affective polarization in the discussions. The regression analysis suggests that there is a statistically significant association between the average toxicity scores of the conversations and the predictions of the model, based on the toxicity feature only. That model obtains an accuracy of 74.47% with an F_1 score of 79.31%. Additionally, automatic detection of angry emotion in online conversations may also be suitable for detecting expressions of negative attitudes towards an out-group, since the anger feature of our regression model is also significantly related to the predictions of out-group derogation. Sentiment analysis however has no such potential, according to our findings.

(ii) Can comment votes predict affective polarization?

Our results do not suggest a significant correlation between the sum of votes of the conversation snippets and expressions of affective polarization. The pattern of the comment votes may however indicate disagreement in the conversations, due to the strong correlation between our vote swing measure and the annotation of disagreement. While disagreement is not in itself an expression of affective polarization, this insight might be useful for further studies on online discussions. For instance, researchers with an interest in collecting conversations that are likely to contain or exclude disagreement may use the information hidden in the vote scores to retrieve relevant conversations.

(iii) Can word embeddings predict affective polarization?

Our findings suggest that word embeddings are suitable for detecting expressions of affective polarization and related concepts such as disagreement. In fact, reduced word embeddings seem to capture aspects of out-group derogation that are not represented by the toxicity score. The best performing models in the regression analysis are fitted on reduced word embeddings and the difference in performance between them and the model with toxicity as its only feature is statistically significant. The results of the regression analysis also suggests that word embeddings are more suitable for predicting disagreement in the conversation than the vote swing feature.

(iv) Do the results differ depending on the group belonging of the participants of a discussion?

With the current experimental setup, we are not able to answer this question with certainty. The performance of the regression models suggest that the combined categories of out-group derogation and disagreement are harder to model than the concepts in isolation. Nevertheless, we cannot say with certainty if this is due to inconsistent annotation, imbalanced categories, the selection of features or the limitations of the statistical model. An additional plausible explanation is that disagreement is not a suitable indicator of group belonging of the participants. On the other hand, the results indicate that the values of toxicity as well as positive sentiment are higher in conversations with out-group derogation but without disagreement.

Furthermore, we can conclude that our study struggles with the known challenges of annotating theoretical constructs. When a classification task is this difficult even for humans, we cannot expect perfect performance from machine learning approaches. With this being said, humans are still superior when it comes to finding abstract concepts in interactive discourse and we need sophisticated models to detect patterns related to such concepts. Our findings suggest that embeddings from advanced language models are suitable to develop automatic detection of affective polarization that captures aspects beyond toxicity. However, due the lack of transparency of advanced models and the large amount of annotated training data required, theory sensitive, accurate and powerful detection of affective polarization remains a challenge.

Lastly, future work consists of careful elaboration and improvement of the annotation guidelines, expansion of the data set and an evaluation of the classifier on other data. The study of Boxell et al. (2020) would be a suitable benchmark as it correlates results of polarization surveys with events related to the covid-19 pandemic. It would be interesting to apply our prediction function to Reddit discussions along the same time line and evaluate the results in relation to the findings in Boxell et al. (2020). Despite the challenges related to automated detection and quantification of affective polarization in online discussion, it could be a way towards deeper understanding of the causes and effects of polarized attitudes. It is of interest to the society to be able to follow the development of behaviour related to intense negative attitudes towards out-groups in order to learn how it can be prevented.

References

Alammar, J. (2018). The illustrated bert, elmo, and co. (how nlp cracked transfer learning) [blog post]. https://jalammar.github.io/illustrated-bert/.

Alsinet, T., Argelich, J., Béjar, R., & Martínez, S. (2021). Measuring polarization in online debates. *Applied Sciences*, 11(24).

Baden, C., Kligler-Vilenchik, N., & Yarchi, M. (2020). Hybrid content analysis: Toward a strategy for the theory-driven, computer-assisted classification of large text corpora. *Communication Methods and Measures*, 14, 165 – 183.

Boe, B. (2016). Python reddit api wrapper (praw).

Boxell, L., Conway, J., Druckman, J. N., & Gentzkow, M. (2020). *Affective Polarization Did Not Increase During the Coronavirus Pandemic*. Working Paper 28036, National Bureau of Economic Research.

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Bavel, J. J. V. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28), 7313–7318.

Capraro, V. & Barcelo, H. (2020). The effect of messaging and gender on intentions to wear a face covering to slow down covid-19 transmission.

Chipidza, W. (2021). The effect of toxicity on covid-19 news network formation in political subcommunities on reddit: An affiliation network approach. *Int. J. Inf. Manag.*, 61(C).

Dahlgren, P. M. (2020). *Media echo chambers : selective exposure and confirmation bias in media use, and its consequences for political polarization*. Publications from the Department of Journalism, Media and Communication (JMG).

Davidson, S., Sun, Q., & Wojcieszak, M. (2020). Developing a new classifier for automated identification of incivility in social media. In *Proceedings of the Fourth Workshop on Online Abuse and Harms* (pp. 95–101). Online: Association for Computational Linguistics.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7, 1–30.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics.

Druckman, J. & Levendusky, M. (2019). What do we measure when we measure affective polarization? *Public Opinion Quarterly*, 83(1), 114–122.

Dubois, E. & Blank, G. (2018). The echo chamber is overstated: the moderating effect of political interest and diverse media. *Information, Communication & Society*, 21(5), 729–745.

Fiorina, M. P. & Abrams, S. J. (2008). Political polarization in the american public. *Annual Review of Political Science*, 11(1), 563–588.

Fridman, A., Gershon, R., & Gneezy, A. (2021). Covid-19 and vaccine hesitancy: A longitudinal study. *PLoS ONE*, 16.

Gamson, W. A. & Modigliani, A. (1994). The changing culture of affirmative action. *Equal employment opportunity: labor market discrimination and public policy*, 3, 373–394.

Halpern, D. & Gibbs, J. (2013). Social media as a catalyst for online deliberation? exploring the affordances of facebook and youtube for political expression. *Computers in Human Behavior*, 29, 1159–1168.

Hamnett, C. (2011). *Chapter 32: Urban Social Polarization*. Edward Elgar Publishing: Cheltenham, UK.

Herring, S. (2004). Computer-mediated discourse analysis: an approach to researching online communities. *Designing for Virtual Communities in the Service of Learning*, (pp. 316–338).

Herring, S. (2010). *Web Content Analysis: Expanding the Paradigm*, (pp. 233–249).

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. Wiley Series in Probability and Statistics. Chicester: Wiley, 3rd ed. edition.

Huggingface (2020). Huggingface.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The origins and consequences of affective polarization in the united states. *Annual Review of Political Science*, 22(1), 129–146.

Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, not ideology: A social identity perspective on polarization. *The Public Opinion Quarterly*, 76(3), 405–431.

Jigsaw (2017). Perspective api.

Kim, Y. & Zhou, S. (2020). The effects of political conflict news frame on political polarization: A social identity approach. (pp. 937–958).

Klein, O., Spears, R., & Reicher, S. (2007). Social identity performance: Extending the strategic side of side. *Personality and Social Psychology review*, 11(1), 28–45.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30, 411–433.

Lecheler, S. (2019). *News framing effects*.

Massanari, A. (2015). *Participatory Culture, Community, and Play: Learning from Reddit*.

Morales, G. D. F., Monti, C., & Starnini, M. (2021). No echo in the chambers of political interactions on reddit. *Scientific Reports*, 11(1).

Nagy, P. & Neff, G. (2015). Imagined affordance: Reconstructing a keyword for communication theory. *Social Media + Society*, 1(2), 2056305115603385.

Nordbrandt, M. (2021). Affective polarization in the digital age: Testing the direction of the relationship between social media and users' feelings for out-group parties. *New Media & Society*, 0(0), 14614448211044393.

Pariser, E. (2011). *The filter bubble : what the Internet is hiding from you*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Perry, T. (2021). LightTag: Text annotation platform. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 20–27). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Radford, A. & Narasimhan, K. (2018). Improving language understanding by generative pre-training.

Rao, D. (2019). *Natural language processing with PyTorch : build intelligent language applications using deep learning*. First edition. edition.

Reddit (2022a). Reddit. https://www.redditinc.com/press.

Reddit (2022b). Reddit. https://www.reddit.com/wiki/search/.

Reiljan, A. (2019). 'fear and loathing across party lines' (also) in europe: Affective polarisation in european party systems. *European Journal of Political Research*, 59.

Rudolph, T. J. & Hetherington, M. J. (2021). Affective Polarization in Political and Nonpolitical Settings. *International Journal of Public Opinion Research*, 33(3), 591–606.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Saravia, E., Liu, H.-C. T., Huang, Y.-H., Wu, J., & Chen, Y.-S. (2018). CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 3687–3697). Brussels, Belgium: Association for Computational Linguistics.

Sayeed, A. (2013). An opinion about opinions about opinions: subjectivity and the aggregate reader. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 691–696). Atlanta, Georgia: Association for Computational Linguistics.

Seabold, S. & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

Siegel, A. A., Nikitin, E., Barberá, P., Sterling, J., Pullen, B., Bonneau, R., Nagler, J., & Tucker, J. A. (2018). Measuring the prevalence of online hate speech with an application to the 2016 us election.

Skalski, P., Neuendorf, K., & Cajigas, J. (2017). *Content Analysis in the Interactive Media Age*, (pp. 201–242).

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1631–1642). Seattle, Washington, USA: Association for Computational Linguistics.

Stroud, N. J. (2008). Media use and political predispositions: Revisiting the concept of selective exposure. *Political behavior*, 30(3), 341–366.

Tajfel, H. & Turner, J. C. (1979). An integrative theory of intergroup conflict.

Tjong Kim Sang, E. F. & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003* (pp. 142–147).

Tunstall, L., von Werra, L., & Wolf, T. (2022). *Natural Language Processing with Transformers: Building Language Applications with Hugging Face*. O'Reilly Media, Incorporated.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., & SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17, 261–272.

Vittinghoff, E. (2012). *Regression Methods in Biostatistics Linear, Logistic, Survival, and Repeated Measures Models*. Statistics for Biology and Health. Boston, MA: Springer US, 2nd ed. 2012. edition.

Wagner, M. (2021). Affective polarization in multiparty systems. *Electoral Studies*, 69, 102199.

Walker, M. A., Anand, P., Abbott, R., Tree, J. E. F., Martell, C., & King, J. (2012). That is your evidence?: Classifying stance in online political debate. *Decision Support Systems*, 53(4), 719–729. 1) Computational Approaches to Subjectivity and Sentiment Analysis 2) Service Science in Information Systems Research : Special Issue on PACIS 2010.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6), 80–83.

Wojcieszak, M. E., Sobkowicz, P., Yu, X., & Bulat, B. (2021). What information drives political polarization? comparing the effects of in-group praise, out-group derogation, and evidence-based communications on polarization. *The International Journal of Press/Politics*, 27, 325 – 352.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., & Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). Online: Association for Computational Linguistics.

Yarchi, M., Baden, C., & Kligler-Vilenchik, N. (2020). Political polarization on the digital sphere: A cross-platform, over-time analysis of interactional, positional, and affective polarization on social media. *Political Communication*, 38, 98 – 139.

Codebook

Guidelines for annotating Reddit-conversations for polarization

This document contains guidelines and examples to ensure reliable and consistent annotation of the Reddit-conversations. The annotation tasks are listed below, together with descriptions and examples to clarify them.

General coding instructions

Each conversation consists of seven comments but should be treated and annotated as a unit. Although the individual comments contribute to the assessment, the annotation occurs on the conversation level and not on the comment level. Each conversation is indexed with a number and has a unique post id and a title. The purpose of revealing the title is to provide context to the comments that follow. However, the post title itself will not be considered part of the conversation. The first comment of a conversation is a reply to the post title, the second comment is a reply to the first comment and so on. Each comment has an index consisting of the post index and the comment number, a unique post id and (when possible) the username of the author. The purpose of revealing the usernames is to facilitate understanding of the conversation, but the usernames themselves should not be treated as part of the conversation.

The annotation tasks

1. The conversation is in English. 2. The conversation handles a topic related to the covid-19 pandemic. 3. The conversation contains derogation of the out-group. 4. There is disagreement in the conversation.

Variables:
1 – Yes 0 – No

The annotation involves evaluation of four statements. If the annotator agrees with a statement, it is marked with 1. If the annotator disagrees with a statement, it is marked with 0. When it is not possible to annotate one or several tasks, they will be coded with 99. This happens for statements 2–4 when the conversation is in a language other than English and for statements 3–4 when the topic of the conversation is not related to the covid-19 pandemic.

Annotation statements 1–2 concern the language and the topic of the conversation. The conversation is only relevant for further annotation if the annotator agrees with both statements. Annotation statement 3 measure one aspect of out-group derogation in the conversation that follows the title of the post. Annotation statement 4 estimates if there is any disagreement in the conversation.

Description of the annotation tasks

1. The conversation is in English.

Description:
A conversation is marked with 0 if there are comments in other languages than English so that a nonspeaker of that language cannot follow the conversation. This means that phrases in other languages than English can be accepted if a monolingual of English would be able to understand the conversation.

2. The conversation handles a topic related to the covid-19 pandemic.

Description:

Conversations that are considered related to the pandemic can discuss topics such as scientific or medical questions regarding covid-19 as a disease, the virus itself and its spread, remedies against covid-19 like restrictions and vaccines, opinions and protests against them or economic and political consequences of the pandemic and related policies on a societal or individual level. Topics related to the covid-19 pandemic also include discussions or opinions about celebrities and politicians associated with attitudes towards the pandemic, vaccines, and restrictions as well as conspiracy theories about the pandemic. Remember this is a non-conclusive list as we might not have thought of all cases.

If a covid-related topic such as the ones listed above is present in the conversation, it will be annotated with 1 on this task. A covid-related topic is considered to be present in a conversation if at least two comments explicitly or implicitly relate to covid-19 in the way described above. If this criterion is not met, the conversation will be annotated with 0 and not annotated further.

3. The conversation contains "derogation of the out-group"

Description:

Derogation of the out-group involves expressing negative attitudes towards a group that the author does not identify or agree with (Wojcieszak et al., 2021). This includes expressing dislike or distrust for the out-group and assigning its members disadvantageous intentions and qualities such as dishonesty or hypocrisy (Nordbrandt, 2021) (Iyengar et al., 2012). Derogation of the out-group can also involve denying its members positive traits.

If at least one commenter in the conversation expresses negative attitudes towards another person or group, regardless of them being present in the conversation, the conversation will be annotated with 1 on this task. The derogation of a person or group does not have to be based on opposing opinions about the covid-19 pandemic. However, the attack needs to be directed towards a person or group which means that attacking a policy does not count as derogating an out-group. In example a), multiple authors call the group that do not share their views on vaccines and restrictions unreasonable and illogical which is a clear example of derogation of the out-group. In example b), a participant is insulting Boris Johnson and his supporters and claims that they cannot be trusted. We can assume that the author does not consider themselves as part of that group which is why this is considered derogation of the out-group.

4. There is disagreement in the conversation.

Description:

If not all participants in the conversation agree with each other, it will be annotated with 1 on this task. This means that if at least one author is questioning another participant's statement or opinion or expressing a different view on the present topic, the conversation will be considered to contain disagreement. The disagreement can be either explicit or implicit and the topic of disagreement does not need to be related to the covid-19 pandemic. However, the disagreement needs to occur between participants of the conversation. This means that a participant expressing disagreement with a policy or politician not present in the conversation is not considered disagreement. If all participants in a conversation agree with each other or express similar opinions on present topic, the conversation will be annotated with 0.