Us vs. Them: A Dataset of Populist Attitudes, News Bias and Emotions

Huguet Cabot, P.-L.; Abadi, D.; Fischer, A.; Shutova, E.

Publication date
2021

Document Version
Final published version

Published in
The 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume

License
CC BY

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

Download date:01 Aug 2021
Us vs. Them: A Dataset of Populist Attitudes, News Bias and Emotions

Pere-Lluís Huguet Cabot\textsuperscript{1,3}, David Abadi\textsuperscript{2}, Agneta Fischer\textsuperscript{2}, Ekaterina Shutova\textsuperscript{1}

\textsuperscript{1} Institute for Logic, Language and Computation, University of Amsterdam
\textsuperscript{2} Department of Psychology, University of Amsterdam
\textsuperscript{3} Babelscape Srl, Sapienza University of Rome

perelluis1993@gmail.com
{d.r.abadi, A.H.Fischer, e.shutova}@uva.nl

Abstract

Computational modelling of political discourse tasks has become an increasingly important area of research in natural language processing. Populist rhetoric has risen across the political sphere in recent years; however, computational approaches to it have been scarce due to its complex nature. In this paper, we present the new Us vs. Them dataset, consisting of 6861 Reddit comments annotated for populist attitudes and the first large-scale computational models of this phenomenon. We investigate the relationship between populist mindsets and social groups, as well as a range of emotions typically associated with these. We set a baseline for two tasks related to populist attitudes and present a set of multi-task learning models that leverage and demonstrate the importance of emotion and group identification as auxiliary tasks.

1 Introduction

Political discourse is essential in shaping public opinion. The tasks related to modelling political rhetoric have thus been gaining interest in the natural language processing (NLP) community. Many of them focused on automatically placing a piece of text on the left-to-right political spectrum. For instance, much research has been devoted to detecting bias in news sources (Kiesel et al., 2019) and predicting the political affiliation of politicians (Iyyer et al., 2014) and social media users, more generally (Conover et al., 2011; Pennacchiotti and Popescu, 2011; Preotiuc-Pietro et al., 2017). Other works conducted a more fine-grained analysis, identifying the framing of political issues in news articles (Card et al., 2015; Ji and Smith, 2017). Recently, the field has also turned attention towards modelling the spread of political information in social media, such as detecting fake news or political perspectives (Li and Goldwasser, 2019; Chandra et al., 2020; Nguyen et al., 2020).

Populism has taken the spotlight in political communication in recent years. Various countries around the globe have experienced a surge of populist rhetoric (Inglehart and Norris, 2016) in both the public and political space. Despite this, approaches to computational modelling of populist discourse have so far been scarce. Due to the flexible nature of populism, annotating populist rhetoric in text is challenging, and the existing research in this area has relied on small-scale analysis by experts (Hawkins et al., 2019). In this paper, we present a new dataset\textsuperscript{1} of Reddit comments annotated for populist attitudes and the first large-scale computational models of this phenomenon. We rely on research in social- and behavioural sciences (e.g., political science and social psychology) to operationalise a definition of populism and an annotation procedure that allows us to capture and generalise the crucial aspects of populist rhetoric at scale.

In social sciences, populism is essentially described as a not fully developed political ideology and a series of background beliefs and techniques (Aslanidis, 2016), traditionally centred around the Us vs. Them dichotomy. In one of the first attempts to fully define populism (Mudde, 2004), it is described as a thin ideology around the distinction between ‘the people’, which includes the ‘Us’, and ‘the elites’ describing the ‘Them’, and with politics being a tool for ‘the people’ to achieve the common good or ‘the popular will’ (Kyle and Gultchin, 2018; Rodrik, 2019). Through different platforms, populism uses this rhetoric that revolves around social identity (Hogg, 2016; Abadi, 2017) and the Us vs. Them argumentation (Mudde, 2004). While right-wing populism tends to be characterised by fear, resentment, anger and hatred, left-wing populism is associated with shame and guilt (Otjes and

\textsuperscript{1}Available at https://github.com/LittlePea13/UsVsThem
Louwerse, 2015; Salmela and von Scheve, 2017a). Moreover, emotions have been shown to be crucial in shaping public opinion more generally (Marcus, 2002, 2003; Demertzis, 2006; Rico et al., 2020).

The design of our annotation scheme and the dataset are inspired by this research, particularly the link between populist rhetoric and both social identity and emotions. Our dataset consists of comments posted on Reddit that explicitly mention a social group. We collect the comments posted in response to news articles across the political spectrum. Through crowd-sourcing, we annotate supportive, critical and discriminatory attitudes towards the group, as well as a range of emotions typically associated with populist attitudes. At the same time, given the relevance of news in the spread of such mindsets, we investigate the relationship between news bias and the Us vs. Them rhetoric. Our data analysis reveals interesting interactions between populist attitudes, specific social groups and emotions.

We also present a series of computational models, automatically identifying populist attitudes, based on RoBERTa (Liu et al., 2019). We experiment in a multi-task learning framework, jointly modelling supportive vs. discriminatory attitudes towards a group, the identity of the group and emotions towards the group. We demonstrate that joint modelling of these phenomena leads to significant improvements in detection of populist attitudes.

2 Related work

2.1 Psychology research on populism

Populist rhetoric revolves around social identity (Hogg, 2016; Abadi, 2017; Marchlewksa et al., 2018; Bos et al., 2020) and the Us vs. Them argumentation (Mudde, 2004). Social identity explores the relations of individuals to social groups. Turner and Reynolds (2010) study the evolution of research into social identity and explain the Us vs. Them as an inter-group phenomenon, exposing its relation to social identity where the “self is hierarchically organised and that it is possible to shift from intra-group (‘we’) to inter-group (‘us’ versus ‘them’) and vice versa.”

Emotions constitute a part of the populist rhetoric and have been essential for information processing and the formation of (public) opinion among citizens (Marcus, 2002; Götz et al., 2005; Demertzis, 2006). While social identity and socio-economic factors have been considered primary indicators of populism’s growth (Rooduijn and Burgon, 2018), emotional factors have lately become a focus within empirical studies, particularly regarding the reactions and spread of populist views (Hameleers et al., 2017). Specific appraisal patterns have characterised emotions, i.e. an adverse event for which one blames the other is felt as anger - a pattern of appraisals is referred to as Core Relational Themes (Smith and Lazarus, 1993; Lazarus, 2001), which are the central (therefore core) harm or benefit that underlies each of the negative and positive emotions (Smith and Lazarus, 1993; Moors et al., 2013). Latest attempts to scrutinise populism from the communication science and social psychological perspective have described populist communication and language (Abadi et al., 2016; Rico et al., 2017) and demonstrated its operationalisation through experimental research (Wirz et al., 2018) as being successful in inducing emotions (Bakker et al., 2020). According to the concept of media populism (Krämer, 2014; Mazzoleni and Bracciale, 2018), media effects can further evoke hostility toward the perceived ‘elites’ and (ethnic/religious) minorities, as it contributes to the construction of social identities, such as in-groups and out-groups (i.e., Us vs. Them).

2.2 Modelling political discourse in NLP

Handcrafted features such as word-frequency (Laver et al., 2003) were initially the base of NLP approaches to model political data. Thomas et al. (2006) introduced the Convote dataset of US congressional speeches, and applied an support-vector machine (SVM) classifier leveraging discourse information to identify policy stances in it. One of the first uses of neural networks on political text was the work of Iyyer et al. (2014), who used a recurrent neural network (RNN) to identify the party affiliation on the Convote dataset. Li and Goldwasser (2019) detected the political perspective of news articles using a long short-term memory (LSTM) and a graph convolutional network (GCN) on user data from Twitter. Other research investigated the framing effect in news articles, which is a mechanism that promotes a particular perspective (Entman, 1993). Card et al. (2015) presented the Media Frame Corpus, which explores policy framing within news articles. Ji and Smith (2017) developed a discourse-level Tree-RNN model to identify the framing in each article, by using this corpus dataset. Huguet Cabot et al. (2020) addressed this
task by leveraging emotion and metaphor detection in an MTL setup. Other works have also explored sentence-level framing (Johnson et al., 2017; Hartmann et al., 2019).

**Hate speech detection** is not limited to the analysis of political discourses. However, it is related to exposing populist rhetoric in digital communication (Meret and Pajnik, 2017; Estelles and Castellví Mata, 2020). Several NLP approaches (Mishra et al., 2020), as well as recent shared tasks (Zampieri et al., 2019, 2020) have been proposed to tackle this widespread problem.

While political bias and framing have been widely explored, research on modelling populist rhetoric is still in its nascent stages. Previous work in this area focused on a general description of populism to determine whether a particular text, such as a party manifesto or a political speech contains what is understood as populist rhetoric or attitudes (Hawkins, 2009; Rooduijn and Pauwels, 2011; Manucci and Weber, 2017). Manual annotation was necessary to perform this analysis, often by experts, which also limited the scope and amount of data used, while the resulting datasets are not sufficiently large to train current machine learning models. Furthermore, the description of what constitutes populist rhetoric is still diffuse and covers many different aspects. Hawkins et al. (2019) used holistic grading to assess whether a text is populist or not, to later determine the degree of ‘populism’ of individual political leaders, thus creating the only existing dataset of populist rhetoric, the *Global Populist Database*.

3 Dataset creation

**Data collection.** By annotating Reddit comments that refer to a social group, we monitored how online discussions target them and whether the text showed a positive or negative attitude towards that social group, ranging from support to discrimination. While this process did not ensure capturing the complexity behind the *Us vs. Them* rhetoric, we detected comments directed at certain groups (out-groups) and the attitude towards them within an online community (in-group). We restricted this to six specific groups that populist rhetoric has targeted as an out-group, **Immigrants, Refugees, Muslims, Jews, Liberals and Conservatives**. Current research has shown these groups are common targets of populism in the US, UK and across Europe (Inglehart and Norris, 2016; Mudde and Kaltwasser, 2018). Note that to annotate sufficient comments per group we limit the current work to six groups, which is by no means a complete list of targeted groups. We encourage future research to broaden the scope of groups covered.

We chose to extract data from Reddit, (1) due to its availability through the *Pushshift* repository² (Baumgartner et al., 2020) and the *Google BigQuery* service, (2) its social identity dynamics (in-group vs. out-group) as close-knit communities created by sub-Reddits, (3) its nature as a social news aggregation platform, and (4) that it has been shown to encourage toxic communication between users and hate speech towards social groups (Massanari, 2017; Salminen et al., 2020; Munn, 2020). To filter the data for annotation, we followed several steps. (1) We identified submissions in Reddit which shared a news article from a news source listed at the *AllSides* website³, (2) we extracted comments which are direct replies to the submission where both the news article title and the comment match any of the keywords for our groups. Keywords were devised using online resources from the *Anti Defamation League*⁴ as well as by consulting social scientists. The full list of keywords can be found in Appendix A Table 4. (3) We selected comments with a minimum of 30 words and a maximum of 250 words, and sampled from specific periods during which each group was actively discussed on Reddit. See Appendix A Table 3 for details. (4) We removed comments that contained keywords from multiple social groups to make the annotation process more straightforward. (5) We randomly sampled 300 comments per group and news source bias according to *AllSides* (left, centre-left, centre, centre-right, right), resulting in a total of 9000 Reddit comments. Note that the bias is not directly related to individual comments, but rather to the news article the comment responded to.

**Annotation procedure.** To capture the *Us vs. Them* rhetoric, we asked: **What kind of language does this comment contain towards _group_.?**, where _group_ corresponds to the specific social group that comment refers to. Respondents had four options: **Discriminatory, Critical, Neutral or Supportive**. An extended description and an example as presented to annotators can be found in Appendix A.2, and Figure 5. We asked anno-

---

² https://pushshift.io/
³ https://www.allsides.com/unbiased-balanced-news
⁴ https://www.adl.org/
tators a second question to capture the emotions expressed towards the group in the same comment. We extended Ekman’s model of 6 Basic Emotions (Ekman, 1992) to a 12-emotions model, which includes a balanced set of positive and negative sentiments. Specifically, we included emotions previously shown to be associated with populist attitudes (Demertzis, 2006). We also provided the annotators with a brief description of each emotion, inspired by the concept of Core Relational Themes (Smith and Lazarus, 1990). The positive emotions are Gratitude, Happiness, Hope, Pride, Relief and Sympathy, and the negative emotions are Anger, Fear, Contempt, Sadness, Disgust and Guilt. Detailed descriptions can be found in Appendix A.2 along with an example 6. The annotation was conducted on Amazon Mechanical Turk (MTurk) and its framework can be accessed here.

Annotation reliability. Once the MTurk annotation was completed, we deployed the CrowdTruth 2.0 toolkit (Dumitrache et al., 2018) to assess the quality of annotations and to identify unreliable workers. CrowdTruth includes a set of metrics to analyse and obtain probabilistic scores from crowdsourced annotations. Worker Quality Score (WQS) is a metric that measures each worker’s performance, by leveraging their agreement with other annotators and the difficulty of annotation. The Media Unit Annotation Score (UAS) is given for each comment and possible answer, indicating the probability with which each option could be the gold label. Finally, Media Unit Quality score (UQS) describes the quality of annotation for each comment. We removed annotations by workers with a WQS lower than 0.1 and those with high disagreement after manually checking their responses, and recomputed the metrics. We removed comments left with only one annotator and comments with a UQS lower than 0.2. This resulted in 4278 comments with 5+ annotators, and 2564 comments with less, which constitute our final dataset.

Following the same procedure as Demszy et al. (2020), we computed inter-rater correlation (Deldago and Tibau Alberdi, 2019) by using Spearman correlation. We took the average of the correlation between each annotator’s answers and all other annotators’ average answers that labelled the same items. We obtain a range between 0.5 and 0.13 per emotion. The lowest agreement is for Relief (0.13), in line with Demszy et al. (2020), where the lowest value for correlation agreement being

Of course Dems are stealing the elections. They are playing by a different set of rules - being ruthless and violent. Dems take no prisoners and show no mercy to their enemies. The sooner GOP realizes it, the better. Because we have to up our game. If things go this way, we have only one way to save this country - Martial Law and kick every single liberal out!

<table>
<thead>
<tr>
<th>Label</th>
<th>UsVsThem</th>
<th>Group</th>
<th>Emotions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discriminatory</td>
<td>1</td>
<td>Liberals</td>
<td>Contempt, Disgust &amp; Fear</td>
</tr>
<tr>
<td>Supportive</td>
<td>0</td>
<td>Refugees</td>
<td>Sympathy</td>
</tr>
</tbody>
</table>

Figure 1: Two samples of our Us Vs. Them dataset.

0.16, and 0.17 for Relief. The full distribution can bee found in Appendix A.3 Figure 7.

4 Data analysis

4.1 UsVsThem scale

For the Us vs. Them question, we aggregated the answers into a continuous scale. To obtain a score for each comment, we computed the CrowdTruth UAS for each of the labels assigned to it and then take a weighted sum. We assigned the weight of 0 to the Supportive label, 1/3 to Neutral, 2/3 to Critical and 1 to Discriminatory. The frequency distribution of comments on this scale can be seen in Appendix A.4 Figure 8. From now on, we will refer to it as the UsVsThem scale.

The scale is skewed, with an overall mean of 0.551±0.265. Although our data selection was random across the selected news sources and groups, there are more comments with negative attitudes towards selected groups than positive or neutral ones due to its nature and our keyword selection.

We performed a two-way ANOVA (Analysis of Variance) test (Fujikoshi, 1993) on news bias and social groups as independent variables and the UsVsThem scale as the dependent variable to see whether the interactions between groups and news bias are significant. One-way tests show statistical significance. Interestingly, there was a statistically significant interaction between the effects of social groups and bias on the UsVsThem scale, F(1, 20) = 12.33, p < 0.05. Values can be found in Appendix A.4 Table 5. Therefore, we explored the interaction between them and the influence of
news bias on how each group is perceived. We performed a Tukey HSD test to check for significance between means in the UsVsThem scale.

**Social groups.** There were differences between groups in terms of the UsVsThem scale when looking at the comment frequency distributions in Figure 2. For Refugees, the distribution was relatively flat as they received a similar amount of positive and negative attitude comments. Immigrants showed a similar distribution with fewer comments in the higher end, i.e. the group received less discrimination than Refugees. Despite the two share many inherent similarities, these differences may be explained by negative media coverage of Refugees portrayed as a threat and being attributed to negative attitudes. Muslims received a higher amount of discriminatory comments than any other group. On the other hand, Conservatives showed a similar mean, due to a very high amount of critical comments. Liberals also received a relatively high amount of critical comments. Both share moderately low tails, as they received less support and discrimination. Finally, Jews showed lower critical and discrimination values, with most values around Neutral, having the lowest mean value of all social groups. These variations translate into a significant \((p < 0.05)\) difference between the means of each group, except for Conservatives and Muslims, and for Liberals and Refugees.

**News source bias.** In this case, the bias was not directly associated with the comment itself. However, differences in the distribution of comments and the out-group attitudes based on the original article’s bias can be observed, as shown in Figure 2. Moreover, means increased from the centre-left to the right bias. Interestingly, there was no symmetry at the centre bias, contrary to the Horseshoe Theory (Hanel et al., 2019), which argues both ends of the political spectrum closely resemble one another. In terms of significant differences, all biases were significantly different from the right bias \((p < 0.05)\), and there was a significant difference between centre-left and centre-right \((p < 0.05)\). The remaining groups showed no significant difference.

**Groups and news source bias.** In line with the above-mentioned bias effect, there was almost always a significant difference between right and centre-right bias and the rest for each group. Only right bias showed a distinct high value and a negative attitude towards Immigrants, which even exceeds those towards Refugees. With the exception of the attitude towards Conservatives, centre, centre-left and left showed lower degrees of negative attitude towards any of the groups. Full results can be seen in Appendix A.4 Table 6 and Figure 10.

### 4.2 Emotions

Instead of using CrowdTruth for emotions, we considered an emotion as being present in the comment provided that at least 1/4 of annotators selected it. In case more than half of annotators marked that comment as Neutral, it was labelled as Neutral. This way, a comment can contain more than one emotion, except for Neutral. Unless specified otherwise, in this subsection Neutral refers to emotionally neutral.

In Figure 3 we present the correlations between the values for each emotion dimension and the UsVsThem scale across all comments, in the same fashion as Demszky et al. (2020). We show the hierarchical relations at the top, demonstrating which emotions interact more strongly with each other. The frequency of emotions in our dataset is as follows: Anger 1724, Contempt 2538, Disgust 1843, Fear 1136, Gratitude 70, Guilt 170, Happiness 59, Hope 307, Pride 174, Relief 37, Sadness 122, Sympathy 1139, Neutral 2094.

**Emotions and the UsVsThem scale.** We were interested in the interaction between emotions and social identity by exploring how the UsVsThem scale is shaped for each emotion. Not surprisingly, comments with negative emotions showed a higher value on the UsVsThem scale and a high correlation
Figure 3: Correlation heat-map for different emotions.

with it, except for Guilt and Sadness. Contempt showed the strongest correlation, while Anger and Fear showed a higher proportion of Discriminatory comments. Guilt and Sadness, on the other hand, are characterised by a lower amount of comments in the Discriminatory range. In line with these results, the UsVsThem scale had a negative correlation with Sympathy and Neutral comments, and while Sympathy was the most frequent positive emotion, other emotions displayed a very similar relation to the UsVsThem scale. These results are visualised in Figure 3 including the UsVsThem scale and the distributions are summarised in Figure 9 in Appendix A.4.

Emotions and groups. We used a two-sided proportion z-test to check for significant differences since emotions are discrete variables. More than 25% of comments towards Muslims and Refugees showed Fear, with no significant difference between the two, followed by Immigrants at 21.5% and other groups at less than 10%. Another notable finding is that Contempt (47.7%) and Disgust (44.4%) were significantly higher towards Conservatives, particularly the latter, which for other groups never exceeded 30% of comments. Sympathy for Liberals (6.5%) and Conservatives (6.9%) was significantly lower when compared to other groups. Hope was present in a significantly higher number of comments for Liberals (7.9%). The values for all proportions can be found in Appendix Appendix A.4 Table 7.

Emotions and bias. Not many differences between biases were found. Most salient was the right bias showing a higher value in all negative emotions, significantly for Anger (31.5%), Contempt (43.4%) and Fear (21.9%). All proportion values can be found in Appendix A.4 Table 8.

5 Modelling populist rhetoric

5.1 Main tasks

Our models’ main focus was to assess to which degree a social group is viewed as an out-group and whether in a negative or discriminatory manner. Our annotation procedure provided a scale from Supportive to Discriminatory for each comment. While this scale is artificial and highly dependent on our task’s context, it provides a good indication of how strongly a social group is targeted in social media comments.

Regression UsVsThem. In our models, we explored two different main tasks. The first task was to predict the values on the UsVsThem scale in a regression model. This scale provides a score for each comment, which illustrates the attitude towards a social group mentioned in the comment ranging from Supportive (closer to 0) to Discriminatory (closer to 1). Values in between depict an intermediate attitude, Neutral lies at 1/3, and Critical at 2/3. By predicting the score, we modelled the out-group attitude of each comment. We used 33% of the data as the test set, and 13.4% as the validation set.

Classification UsVsThem. Our second task was to classify each comment in a binary fashion as whether the comment shows a negative attitude towards a group, i.e., Critical or Discriminatory, or not, i.e., Neutral or Supportive. This task resulted in a relatively balanced dataset, with 56% of Critical or Discriminatory comments. We used the same splits as before.

5.2 Auxiliary tasks

Emotion detection. Interactions between populist rhetoric and emotions have been explored in political psychology through surveys and behavioural experiments (Fischer and Roseman, 2007; Tausch et al., 2011; Salmela and von Scheve, 2017b; Redlawsk et al., 2018; Rollwage et al., 2019; Nguyen, 2019; Roseman et al., 2020). This is consistent with our findings in section 4 and further motivates modelling emotions in the context of populist rhetoric. For each comment, emotions were annotated as a Boolean vector. For our task, some
emotions were rarely annotated or only present alongside more frequent ones. They increased the difficulty of the task while not providing relevant information. To simplify the auxiliary task, we considered the 8 most common emotions, Anger, Contempt, Disgust, Fear, Hope, Pride, Sympathy and Neutral.

**Group identification.** In the work of Burnap and Williams (2016), types of hate speech were differentiated based on race, religion, etc., and models were trained specifically on those categories. In ElSherief et al. (2018), data-driven analysis of online hate speech explored in profundity the differences between directed and generalised hate speech, and Silva et al. (2016) analysed the different targets of hate online. In our case, the Us vs. Them rhetoric metric showed significant differences for each group as we have seen in the previous section. Therefore, we hypothesised that the information bias (Caruana, 1993) the group identification task provides will help understand the Us vs. Them rhetoric aimed at the different social groups, which motivated its role as an auxiliary task.

5.3 Model architecture

We used the Robustly Optimized BERT Pretraining Approach (RoBERTa) (Liu et al., 2019) in its BASE variant as provided by Wolf et al. (2019).

**Multi-task learning.** In all setups, tasks shared the first eleven transformer layers of RoBERTa. The final 12th layer was task-specific, followed by a classification layer that used the hidden representation of the <s> token, to output a prediction. We used scheduled learning, where the losses of each task are weighted and changed during training. We also experimented with a three-task MTL model where the two auxiliary tasks are learned simultaneously.

We assigned three different loss weights associated with each task, \( \lambda_m \) for the main task, either regression or binary classification; \( \lambda_e \) for emotion detection; \( \lambda_g \) for group identification. For MTL with one auxiliary task, \( \lambda_m + \lambda_e = \lambda_m + \lambda_g = 2 \), while for the three-task MTL: \( \lambda_m + \lambda_e + \lambda_g = 3 \).

**Regression UsVsThem.** We used Mean Squared Error loss with a sigmoid activation function for the main task. For emotion identification as the auxiliary task, we used Binary Cross-Entropy loss, and for the group identification, we used Cross-Entropy loss, both with sigmoid activation. For all MTL models, there was a warm-up period of \( \omega \) epochs, after which the weight is changed to \( \lambda_g = 10^{-2} \) and \( \lambda_e = 10^{-5} \), and \( \lambda_e = \lambda_g = 10^{-5} \) for the three-task setting.

**Classification UsVsThem.** We used Cross-Entropy loss with a sigmoid activation function for the main task. The remaining tasks were kept the same as with the Regression case above. For all MTL models, there was a warm-up period of \( \omega \) epochs, after which the weight was changed to \( \lambda_g = 10^{-2} \) and \( \lambda_e = 10^{-2} \), and \( \lambda_e = \lambda_g = 10^{-5} \) for the three-task setting.

5.4 Experimental setup

**Regression UsVsThem.** We report model performance in terms of Pearson correlation coefficient (R). We found the optimal STL hyperparameters using the validation set: a learning rate of \( 3e-05 \), a lineal warm-up period of 2 epochs and dropout of 0.15. The batch size used was 128. These hyperparameters were kept constant across our experiments for the regression UsVsThem task. For the emotion detection MTL setup, \( \lambda_e = 0.15 \) and \( \omega = 8 \). For the groups MTL, \( \lambda_g = 0.15 \) and \( \omega = 5 \). For the three-task MTL model we obtained optimal validation performance by setting \( \omega = 8 \) and both \( \lambda_g = \lambda_e = 0.073 \), which was the equivalent of \( \lambda_g = \lambda_e = 0.05 \) for the two-task MTL.

**Classification UsVsThem.** Similarly, we ran a grid-search to find the best hyperparameters for the classification setup. For the STL model, we obtained a learning rate of \( 5e-05 \), a warm-up of 2 epochs and an extra dropout of 0.2. For emotions-MTL, \( \lambda_e = 0.2 \) and \( \omega = 8 \). For the groups-related MTL, \( \lambda_g = 0.25 \) and \( \omega = 5 \). For the three-task MTL, \( \lambda_e = 0.95, \lambda_g = 0.25, \omega = 8 \). For both regression and classification, we report performance averaged over 10 different seeds.

5.5 Results

Results are presented in Table 1. We find that MTL outperforms STL in both versions of our task.

**Regression UsVsThem.** The STL baseline showed a 0.545 Pearson R to the gold score. When emotion identification was used as an auxiliary task, the performance increased by almost one point, to 0.553. The groups MTL setup showed a higher increase, up to 0.557. Both improvements
were significant compared to the STL model, using the Williams test (Williams, 1959). Perhaps all the more interesting is that the three-task MTL model achieved the highest performance, even without its hyperparameters being specifically tuned as with the other setups. It resulted in a Pearson R of 0.570, i.e. over 2 points performance increase over STL, a statistically significant improvement over both STL and the remaining two MTL approaches.

**Classification Us Vs Them.** Although not shown in the table, the accuracy baseline for a majority class classifier would be 0.550. All models highly surpassed that, with the real baseline set by the STL setup achieving a 0.705 accuracy. Results for the MTL approaches were similar to what we observed in the regression task. Emotion-MTL increased performance by half a point, to 0.710, as did group-MTL, with 0.711. The best performing model was again the three-task MTL, at 0.717, yielding a statistically significant improvement over STL, using the permutation test.

**5.6 Analysis**

**Qualitative and error analysis.** We selected comments with higher values on the scale where MTL improved the STL baseline predictions for the regression task. Comments with high emotion valence were better predicted by models that included emotion identification. Comments that had group-specific rhetoric with references to (derogatory) terms such as ‘illegal aliens’ were better predicted by models that incorporated group identification (see the first example in Table 2).

The standard deviation of the difference between the STL and the three-task predictions was just 0.055. This means that MTL helped capture nuanced information that improved prediction; however, comments with high squared error for STL still showed similar behaviour for MTL models. This aspect is shown in Appendix A.5 Figure 11. All models’ squared error showed a pair-wise Pearson correlation higher than 0.92. This observation prompted us to investigate comments with a high squared error. We identified three different sources. (1) **Comments with emotionally charged language**, slurs, or insults, which may often be associated with a more negative attitude towards a group, were mispredicted due to not being negative towards such group or being used ironically or satirically. (see second example in Table 2). (2) **Reference to multiple groups**: we removed comments that included keywords from similar groups, however it was impossible to account for all the terms that may refer to other groups. Hence, there are comments for which the prediction seems to be about a target different than the one at annotation time (see the third example in Table 2). More examples can be found in Appendix A.5 Table 9. (3) **Annotation error** is expected in any crowd-sourced annotation. While these were not as frequent as to pose a problem during training, they did occur as incorrect model predictions that can mistakenly decrease performance.

**Analysis of model representations.** Our qualitative analysis showed that auxiliary tasks had a positive effect on predictions. Still, it cannot explain how the model changes its underlying structure and Reddit comments’ encoding. We explored how auxiliary tasks affected how the network encodes information through its layers to get a better understanding. We used **t-Distributed Stochastic Neighbour Embedding** (t-SNE) (van der Maaten and Hinton, 2008), a stochastic technique for dimension-
1 Does anyone else think it’s absurd that we have Muslims serving in our congress? Literal foreign agents acting as if they hold the interests of the country and its people at heart? They never talk about the will of the people. It’s always some bullshit about how white men (who founded and built this country) are evil and we need to let in more 3rd worlders who want to bomb and kill us. This is literal clown world tier nonsense.

2 You proud of yourselves, making 3 year olds represent themselves in immigration court? You fucking proud of that insanity? All for the sake of keeping out a gang that has already been in America for a long time, meanwhile regular home grown white kids are murdering dozens of their own classmates but goddam, at least they we’re legal, amirite

3 Conservatives have every right to revolt. If we don’t get our way we will destroy the country. I hope the left keeps pushing us to provoke a civil war. Or maybe Commiefornia should secede. Maybe that’s the best thing that can happen, a complete break up. That way we can have our ethnostate, and the left can have their degenerate cesspool without us paying taxes for it. The US is dead anyway. It’s time to burn this diverse shithole to the ground. It will be the ultimate proof that diversity doesn’t work.

<table>
<thead>
<tr>
<th>Label</th>
<th>MTL, E. &amp; G.</th>
<th>MTL, Emo.</th>
<th>MTL, Groups</th>
<th>STL</th>
<th>Group</th>
<th>Emotions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.000</td>
<td>0.872</td>
<td>0.870</td>
<td>0.759</td>
<td>Muslims</td>
<td>Anger, Contempt, Disgust &amp; Fear</td>
</tr>
<tr>
<td>2</td>
<td>0.02</td>
<td>0.774</td>
<td>0.874</td>
<td>0.740</td>
<td>Immigrants</td>
<td>Sympathy</td>
</tr>
<tr>
<td>3</td>
<td>0.071</td>
<td>0.729</td>
<td>0.773</td>
<td>0.747</td>
<td>Conservatives</td>
<td>Hope &amp; Pride</td>
</tr>
</tbody>
</table>

Table 2: Examples of predictions for comments. Predictions are averages over 10 seeds for each model.

We used it to visualise the hidden representations of the test set comments in-between transformer layers across the network. We present the results for both STL and the three-task MTL in Appendix A.5 Figures 12 and 13, where for both the first layers showed some structure not related to the tasks at hand. As we were using pre-trained weights from RoBERTa, this could be explained by the first layers modelling lower-level language characteristics as shown empirically in Tenney et al. (2019), where probing mechanisms indicate early layers were more relevant for tasks such as POS tagging. For STL, the last layers showed the UsVs-Them scale continuously in the y axis. Once we introduced the auxiliary tasks of group identification and emotion classification differences in the last layers were exacerbated. For three-task MTL the last layers showed clusters for each social group, and related groups were closer together, such as Refugees and Immigrants, or Liberals and Conservatives. We also observe a radial distribution with highly emotional comments further away from the centre. Comments with very distant values on the scale (Discriminatory and Supportive) were closer together than with those in the mid-range (Neutral and Critical) as seen in Figure 4 and Appendix A.5 Figure 15. While paradoxical, our interpretation is that the model leverages the valence of emotion, where Discriminatory and Supportive comments are more loaded with emotion. This leads to a better performance of MTL compared to STL. This idea is supported by the distribution of emotions on the last layer, where emotionally neutral comments are closer to the centre of the plot, while more emotionally charged comments radially increase, visualised in Appendix A.5 Figure 14. In Appendix A.5 Figures 16 and 17 we present the emotion distribution for the group and emotion-specific layers, respectively.

6 Conclusions

We presented a new, large-scale dataset of populist rhetoric and the first series of computational models on this phenomenon. We have shown that joint modelling of emotion and populist attitudes towards social groups enhances performance over the single-task model, further corroborating previous research findings in various social sciences. Future work may deploy social information (e.g., Twitter) or explore the interactions of populist attitudes and the political bias of news articles as provided in our Us Vs. Them dataset.

Funding statement. This research was funded by the European Union’s H2020 project Democratic Efficacy and the Varieties of Populism in Europe (DEMOS) under H2020-EU.3.6.1.1. and H2020-EU.3.6.1.2. (grant agreement ID: 822590) and supported by the European Union’s H2020 Marie Skłodowska-Curie project Knowledge Graphs at Scale (KnowGraphs) under H2020-EU.1.3.1. (grant agreement ID: 860801).
References


Shantanu Chandra, Pushkar Mishra, Helen Yankoukdakis, Madhav Ninmishakavi, Marzieh Saeidi, and Ekaterina Shutova. 2020. *Graph-based modeling of online communities for fake news detection*.


Marta Estelles and Jordi Castellvi Mata. 2020. *The educational implications of populism, emotions and digital hate speech: A dialogue with scholars from canada, chile, spain, the uk, and the us*. *Sustainability*, 12.


Christoph G Nguyen. 2019. Emotions and populist support.


Dani Rodrik. 2019. What’s driving populism?


Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. CoRR, abs/cs/0607062.


A Supplemental material

A.1 Data collection

<table>
<thead>
<tr>
<th>Time ranges</th>
<th>Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conservatives</td>
<td>2016/09/15 - 2016/12/15 Election periods</td>
</tr>
<tr>
<td></td>
<td>2018/09/15 - 2018/12/15</td>
</tr>
<tr>
<td>Liberals</td>
<td>2016/09/15 - 2016/12/15 Election periods</td>
</tr>
<tr>
<td></td>
<td>2018/09/15 - 2018/12/15</td>
</tr>
<tr>
<td>Muslims</td>
<td>2016/11/01 - 2017/11/30 Trump Muslim ban,</td>
</tr>
<tr>
<td></td>
<td>2018/04/01 - 2018/05/01 Mosque attacks,</td>
</tr>
<tr>
<td></td>
<td>2019/03/01 - 2019/09/06/01</td>
</tr>
<tr>
<td>Immigrants</td>
<td>2016/11/01 - 2017/11/30 Migrant caravans,</td>
</tr>
<tr>
<td></td>
<td>2017/01/15 - 2017/03/15 Children at the US</td>
</tr>
<tr>
<td></td>
<td>2018/06/17 - 2018/07/01 border</td>
</tr>
<tr>
<td></td>
<td>2018/10/01 - 2019/02/01</td>
</tr>
</tbody>
</table>

Table 3: Events and periods used for each group. If comments were not sufficient, they were sampled randomly from other time ranges. Refugees did not have enough overall comments to be filtered by time range.

<table>
<thead>
<tr>
<th>News Title</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Refugees</td>
<td>refugee, asylum seeker</td>
</tr>
<tr>
<td></td>
<td>refugee, asylum seeker, undocumented, colonization</td>
</tr>
<tr>
<td>Immigration</td>
<td>-migra-, undocumented, colonization</td>
</tr>
<tr>
<td></td>
<td>-migra-, undocumented, colonization</td>
</tr>
<tr>
<td>Muslims</td>
<td>muslim, arab, muhammad, muhammed, islam, hijab, sharia</td>
</tr>
<tr>
<td></td>
<td>muslim, arab, muhammad, muhammed, islam, hijab, sharia</td>
</tr>
<tr>
<td>Jews</td>
<td>-jew(i/s)-, heeb-, sikey-, -zionis-, -semit-</td>
</tr>
<tr>
<td></td>
<td>-jew(i/s)-, heeb-, sikey-, -zionis-, -semit-</td>
</tr>
<tr>
<td>Liberals</td>
<td>antifa, libtard, communist, socialist, leftist, liberal, democrat</td>
</tr>
<tr>
<td></td>
<td>antifa, libtard, communist, socialist, leftist, liberal, democrat</td>
</tr>
<tr>
<td>Conservatives</td>
<td>alright, alt-right, cuckervative, trumpster, conservative, republican</td>
</tr>
<tr>
<td></td>
<td>alright, alt-right, cuckervative, trumpster, conservative, republican</td>
</tr>
</tbody>
</table>

Table 4: Keywords used in our data filtering process. The use of more emotionally laden terms is justified by their low occurrence compared to more common terms just to ensure a more diverse dataset.

A.2 Description of the annotation options

Discriminatory or Alienating. Annotators were asked to mark this in case the comment was either, (A) alienating or portraying a social group as negative, (B) a threat, danger or peril to society, (C) trying to ridicule it and attack that group as lesser or worthless.

Critical but not Discriminatory. In case the comment was critical, but not to the extent of the first option, annotators were asked to mark this option.

Supportive or Favorable. This answer refers to comments expressing support towards that group, by defending it or praising it.

Neutral. This option was offered in case none of the above applied, either because the group was only mentioned but the comment was not addressed at them, or there was no opinion whatsoever expressed towards the group, such as expressing purely factual information.

Annotators were first asked to select whether the comment showed a ‘Positive’, ‘Negative’ or ‘Neutral’ sentiment towards the specified group. With this approach, we intended to simplify the task and guide annotators, which then were offered to choose from 6 positive or 6 negative emotions according to sentiment they initially chose. In case annotators selected Neutral no further options were provided. The descriptions for each emotion were:

Positive emotions: Gratitude Someone is doing/causing something good or lovely. Happiness/Joy ⁵ Something good is happening. Something amusing or funny is happening. Hope Something good/better might happen (sooner or later). Pride Someone is taking credit for a good achievement. Sympathy Someone shows support or devotion.

Negative emotions: Anger. Someone is causing harm or a negative/undeserved outcome, while this could have been avoided. Someone is acting in an unjustified manner towards people. Someone is blocking the goals of people. Anxiety/Fear Something negative might/could happen (sooner or later), which threatens the well-being of people. Contempt Someone is inferior (for example, immoral, lazy or greedy). Someone is incompetent (for example, weak or stupid).

Neutral ⁶ Someone has experienced a loss (for example, death or loss of possessions). Moral Disgust ⁷ Someone behaves in an offensive way (for example, corrupt, dishonest, ruthless, or unscrupulous behavior).

Guilt/Shame ⁷ Someone sees him-/herself as responsible for causing a harmful/immoral/shameful/embarrassing outcome to people.

Referred to as Happiness for simplicity
Referred to as Disgust for simplicity
Referred to as Guilt for simplicity
Figure 5: Example of the first question as reference presented to MTurk annotators.

Figure 6: Example of the second question as reference presented to MTurk annotators.
A.3 Reliability

We also applied the algorithm called *Leave-One-Rater-Out* PPCA Cowen et al. (2019), using Bonferroni correction on p-values. Principal Preserved Component Analysis (PPCA) finds principal components which instead of preserving variance within a single dataset as conducted in PCA, preserve the cross-covariance between two different datasets, in our case being a comparison between annotations by one rater and a random set of other raters. In this manner, we can assess the degree of agreement and whether all component dimensions are significant, indicating significant emotion dimensions to be preserved. In our setup, the largest p-value for a dimension was $1.2e-03$, with all other dimensions showing much smaller values. This supports the idea that our emotion dimensions are significant, in order to be kept.

A.4 Data analysis

Figure 8: Distribution for the *UsVsThem* scale. Values closer to 0 are more supportive towards the target group, while higher values indicate a higher degree of criticism or eventually discrimination.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Sum of Squares</th>
<th>.df</th>
<th>Mean Square</th>
<th>F</th>
<th>p</th>
<th>partial (\eta^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>2582.47</td>
<td>1</td>
<td>2582.47</td>
<td>46 x 10^3</td>
<td>0.000</td>
<td>0.004</td>
</tr>
<tr>
<td>Groups</td>
<td>22.05</td>
<td>5</td>
<td>4.41</td>
<td>78.73</td>
<td>0.000</td>
<td>0.04</td>
</tr>
<tr>
<td>Bias</td>
<td>4.82</td>
<td>4</td>
<td>1.21</td>
<td>21.52</td>
<td>0.000</td>
<td>0.01</td>
</tr>
<tr>
<td>Groups x Bias</td>
<td>13.82</td>
<td>20</td>
<td>0.69</td>
<td>12.33</td>
<td>0.000</td>
<td>0.03</td>
</tr>
<tr>
<td>Error</td>
<td>492.63</td>
<td>8794</td>
<td>0.06</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Two-way ANOVA test.
Table 6: Mean UsVsThem Regression scale for each group and bias. Statistical significance is shown as super-indexes, in case the mean is statistically different with other biases for that group. $^l$ left, $^c$ centre-left, $^c$ centre, $^r$ centre-right, $^r$ right. Tested using Tukey HSD test.

<table>
<thead>
<tr>
<th>Group</th>
<th>Conservatives</th>
<th>Liberals</th>
<th>Immigrants</th>
<th>Refugees</th>
<th>Jews</th>
<th>Muslims</th>
</tr>
</thead>
<tbody>
<tr>
<td>left</td>
<td>0.668$r,c,r$</td>
<td>0.540$r,c,r$</td>
<td>0.471$r$</td>
<td>0.524$r$</td>
<td>0.433</td>
<td>0.554$r,c,r$</td>
</tr>
<tr>
<td>centre-left</td>
<td>0.669$r,c,r$</td>
<td>0.513$r,c,r$</td>
<td>0.447$r$</td>
<td>0.507$r$</td>
<td>0.422</td>
<td>0.556$r,c,r$</td>
</tr>
<tr>
<td>centre</td>
<td>0.646$r,c,r,l$</td>
<td>0.487$r,c,r$</td>
<td>0.516$r$</td>
<td>0.541$l$</td>
<td>0.452</td>
<td>0.573$r,c,r$</td>
</tr>
<tr>
<td>centre-right</td>
<td>0.555$c,l,c,l$</td>
<td>0.602$c,l,c,l$</td>
<td>0.497$r$</td>
<td>0.557</td>
<td>0.433</td>
<td>0.682$c,l,c,l$</td>
</tr>
<tr>
<td>right</td>
<td>0.543$c,l,c,l$</td>
<td>0.638$c,l,c,l$</td>
<td>0.646$r,c,c,l,l$</td>
<td>0.625$c,l,c,l$</td>
<td>0.467</td>
<td>0.696$c,l,c,l$</td>
</tr>
</tbody>
</table>

Table 7: Percentages of comments within each social group per emotion.

<table>
<thead>
<tr>
<th>Group</th>
<th>Anger</th>
<th>Contempt</th>
<th>Disgust</th>
<th>Fear</th>
<th>Gratitude</th>
<th>Guilt</th>
<th>Happiness</th>
<th>Hope</th>
<th>Pride</th>
<th>Relief</th>
<th>Sadness</th>
<th>Sympathy</th>
<th>Neutral</th>
</tr>
</thead>
<tbody>
<tr>
<td>left</td>
<td>22.5%</td>
<td>35.8%</td>
<td>25.8%</td>
<td>15.1%</td>
<td>0.7%</td>
<td>2.4%</td>
<td>0.6%</td>
<td>4.9%</td>
<td>2.5%</td>
<td>0.5%</td>
<td>2.0%</td>
<td>18.6%</td>
<td>31.5%</td>
</tr>
<tr>
<td>centre-left</td>
<td>21.1%</td>
<td>34.1%</td>
<td>27.0%</td>
<td>13.5%</td>
<td>1.2%</td>
<td>2.6%</td>
<td>0.7%</td>
<td>5.0%</td>
<td>2.1%</td>
<td>0.4%</td>
<td>1.8%</td>
<td>18.1%</td>
<td>32.6%</td>
</tr>
<tr>
<td>centre</td>
<td>24.4%</td>
<td>35.2%</td>
<td>26.1%</td>
<td>15.0%</td>
<td>1.5%</td>
<td>2.1%</td>
<td>1.1%</td>
<td>5.0%</td>
<td>2.7%</td>
<td>0.7%</td>
<td>1.8%</td>
<td>15.1%</td>
<td>32.8%</td>
</tr>
<tr>
<td>centre-right</td>
<td>25.9%</td>
<td>36.3%</td>
<td>26.5%</td>
<td>17.0%</td>
<td>1.2%</td>
<td>2.4%</td>
<td>0.9%</td>
<td>3.9%</td>
<td>3.2%</td>
<td>0.7%</td>
<td>1.2%</td>
<td>17.8%</td>
<td>29.4%</td>
</tr>
<tr>
<td>right</td>
<td>31.5%</td>
<td>43.4%</td>
<td>29.0%</td>
<td>21.9%</td>
<td>0.6%</td>
<td>2.9%</td>
<td>0.9%</td>
<td>3.6%</td>
<td>2.2%</td>
<td>0.4%</td>
<td>2.1%</td>
<td>13.5%</td>
<td>26.4%</td>
</tr>
</tbody>
</table>

Table 8: Percentages of comments within bias in the news source per emotion.

Figure 9: UsVsThem scale for each emotion.
Figure 10: Distribution for the *UsVsThem* scale per social group and bias. The mean for the scale is shown at the x axis.
A.5 Analysis

Figure 11: Squared error for STL and three-task MTL

Table 9: 1 and 2 are examples of predictions for comments with high values on the UsVsThem scale where MTL models showed an improvement over STL, 3 and 4 are examples of ambiguous and challenging comments and 5 is an example with mentions to more than one group with high error predictions. Predictions are averages of all 10 seeds predictions for each model.
Figure 12: Hidden representations at each layer of the Transformer model for the single task model. Red represents a value closer to 1 in the *UsVsThem* scale and blue closer to 0.
Figure 13: Hidden representations at each layer of the Transformer model for the three-task MTL. The last plots show the task specific Transformer layer output. Red represents a value closer to 1 in the UsVsThem scale and blue closer to 0.
Figure 14: Hidden representations for the three-task MTL main task specific layer. Emotions are represented by colours. Comments with more than one emotion show an average of the colours.
Figure 15: Hidden representations for the three-task MTL main task specific layer. Red represents a value closer to 1 in the *UsVsThem* scale and blue closer to 0.
Figure 16: Hidden representations for the three-way MTL group identification specific layer. Emotions are represented by colours. Comments with more than one emotion show an average of the colours.
Figure 17: Hidden representations for the three-way MTL emotion specific layer. Emotions are represented by colours. Comments with more than one emotion show an average of the colours.