UNIVERSITÁ DEGLI STUDI DI MILANO-BICOCCA

Dipartimento di Psicologia

Corso di laurea in Scienze Psicosociali della Comunicazione

Predicting Political Ideology: A Classification Approach through Moral Foundation Analysis

Numero di caratteri: 51608

Anno Accademico 2022/2023

Index

Preface

The topics, methodologies and form of this research all stem from the fight of interests (or maybe cognitive dissonance) that I experienced during my bachelor journey.

When I initially applied for the degree in Psychosocial Sciences Of Communication I was mainly interested in understanding social dynamics, from a psychological, social and communicative perspective. I wanted to understand why and how people interact the way they do, mainly from a theoretical and descriptive point of view.

I was initially drawn by courses like Social Psychology, Philosophy of Language and Linguistics, because they all contributed to unveil how beautifully complex we as humans are.

However, by proceeding with my studies, I also developed an interest in models of behavior. Courses such as Learning, Thinking and Decision Making introduced me to the notion that computational approaches could be employed to predict and model social and cognitive phenomena.

Eager to gain the knowledge to better understand how to perform such techniques, I embarked on a quest to add as much mathematical and statistical competences to my baggage as possible.

I wanted to be able to answer social and psychological questions using scalable and comprehensive approaches.

This thesis represents my attempt to connect the dots, merging the two main interests I developed during my undergraduate journey.

By observing the world and social trends, I came up with a research question that I consider both fundamental and important: Why do people disagree? And is it possible to ease discussions, fostering an intellectually honest and hate-less society?

To answer these questions, I conducted a research stage in which, followed by my advisor - who I'm very thankful to for her patience and guidance - I combined recent psychological results along with new cutting edge technologies in the field of statistics and data science.

Note that this is just an unpretentious attempt of mine to combine the two fields to answer pure social and psychological questions. If anything, I believe that this ride gave me the opportunity to learn a lot: I understood how to conduct research from start to finish, reading and understanding the data and communicating the results

while having the opportunity to hone my programming skills, mathematical and statistical knowledge.
To be fair, I also have to admit that I had a lot of fun during this journey.
In the future, I'm looking forward to having more of this, hence why I'm choosing to pursue a master's degree in the field of data science.
Technology advances and more opportunities are yet to come. I'm striving to be prepared to harness whatever the future will bring to the table, both to fulfill my curiosity and hopefully to make - even if little and insignificant - a positive impact to the world we live in.

1. Introduction

As political polarization grows around the world (Gidron, Adams & Horne, 2019; Pew Research Center, 2017), social media platforms are progressively becoming more like arenas where individuals defend their own ideas and attack others (Tucker et al., 2018).
In their comprehensive review, Tucker and colleagues (2018) discuss the interconnection between social media, political polarization and disinformation and their effects into the quality of policies. In their work they argue that growth in political division (Finkel et al., 2020) and cross-cutting information avoidance (Frimer, Skitka & Motyl, 2017; Barber, Jost, Nagler, Tucker & Bonneau, 2015) are related to an increase in spread of misinformation, worsening of civil respect during political debates and a reduced likelihood of compromise around policy actions between political parties (Tucker et al., 2018). The spread of misinformation through social media ought to be considered a serious matter as some researches have shown that false stories shared on those platforms might have played a role in the election results in 2016 (e.g. Gunther, Beck & Nisbet, 2019) and on COVID-19's pandemic public opinion (Caceres et al., 2022). Thus, polarization seems to carry real and potentially harmful consequences into the real world.
Furthermore, in their review, Kubin and Sikorski (2021) suggests that politicians might be intentionally sharing polarized content on social media, as it seems to be associated with an increase in readership (Hong & Kim, 2016).
Politicians seem to be leveraging specific communication tactics, the so-called *wedge* issues, rhetorical strategies, usually focused on social concerns, that are intentionally

constructed to divide party voters and polarize the public to gain political advantage (Wiant, 2002).

While it can be easy to observe the sharing of extreme political views on social medias, the relationship between such platforms and the general increase in polarization is still not so clear (Tucker et al., 2018).

Nonetheless, research (Ribeiro, Ottoni, West, 2020) has shown that social media's recommendation algorithms - such as the YouTube one - might play a role in political radicalization. The study showed how users tend to transition from moderately radical videos towards watching increasingly more politically extreme content, leading to the *radicalization* phenomena.

In the work I will describe in this dissertation we explored whether individuals who followed accounts characterized by polarized views would show differences in their expression of morality values.

For this, we studied social and political phenomena in the UK context through the theoretical lenses of Moral Foundation Theory (Haidt, 2012).

1.1 Moral Foundation Theory

The Moral foundation Theory (MFT; Haidt, 2012) is a theory in the field of social psychology developed to understand moral variations and communalities among cultures. Haidt (2012) argues that individuals' understanding of what is morally acceptable and what is not is based on "building blocks", that are called 'moral foundations'.

The theory emphasizes 5 main foundations:

- *Care/harm*: This foundation is related to the ability to feel (and dislike) the pain of others. It underlies virtues of kindness, gentleness, and nurturance.

- *Fairness/cheating*: This foundation is related to the evolutionary process of reciprocal altruism. It generates ideas of justice, rights, and autonomy.

- *Loyalty/betrayal*: This foundation is related to our tendency to form coalitions. It underlies virtues of patriotism and self-sacrifice for the group.

- *Authority/subversion*: This foundation underlies virtues of leadership and followership, including deference to legitimate authority and respect for traditions.

- *Sanctity/degradation*: This foundation is related to disgust and contamination and underlies religious notions of striving to live in an elevated, less carnal, more noble way.

Interestingly, and relevant for the present work, individuals with different political orientations seem to be characterized by their differential use of these moral foundations, which could help understand their disagreement on political topics (Graham, Haidt & Nosek, 2009).

Specifically, a seminal work in the field of moral psychology (Graham et al., 2009) effectively demonstrated the underlying moral differences between individuals with liberal and conservative orientations.

In the study, participants answered a self-report on their moral foundations (the Moral Foundation Questionnaire) and differences emerged in the way liberals and conservatives used moral foundations across different scenarios:

Liberals engaged more in what Haidt (2012) calls the "individualizing foundations" (*Care/Harm* and *Fairness/Cheating*), while conservatives seemed to rely equally on each of the five foundations (Graham et al., 2009). These results were confirmed by other studies resting on self-report questionnaires and are at the bulk of the Moral Foundation Hypothesis (MFH; Haidt, 2012).

Data gathered through other methods, such as linguistic and text-analysis, further found that liberals differentiated themselves from conservatives also for greater values in *Loyalty/Cheating* (e.g., Study 4, Graham et al., 2009; Frimer, 2019).

During the years, researchers have proposed several different methods to investigate individuals' moral foundations usage.

Tools like the Moral Foundation Questionnaire in various versions have been widely used in research on the topic (Frimer, 2019).

Other methods such as the Moral Foundation Dictionary have been proposed by researchers (Graham, Haidt): this is a comprehensive list of words which allow to score the proportions of morally violating (*vice)* words and morally virtuous words in a text (*virtue*) (see moralfoundations.org).

As language seems relevant for the expression of moral foundations, dictionaries can be valuable tools for capturing individuals' morality patterns (Frimer, 2019).

1.1.1 Moral Reframing

Beyond revealing the moral differences between individuals of different political stances (Graham et al., 2009), MFT has also been employed to enhance persuasion in communication (e.g. Wolsko, Ariceaga & Seiden, 2016; Feinberg, Willer, 2013). Moral reframing is the practice of framing a message in accordance with the target's moral values to improve messages' communicative effectiveness (Feinberg, Willer, 2013).

More specifically, the authors argue that when crafting a message, liberals and conservatives tend to frame it using their own moral values, reducing the persuasive efficacy when engaging with individuals of the opposing political stance (Feinberg, Willer, 2013; Feinberg, Willer, 2019).

Given the already discussed growth in political sectarianism (Finkel et al., 2020) and the instrumentalization of polarization among politicians, moral reframing might play a role in reducing polarization and filling the so-called "moral empathy gap" in political communication (Feinberg, Willer, 2019).

1.2 Research Outline and goals

In his comprehensive book *The Righteous Mind* (2012), Haidt discusses the importance of morality in shaping the way people think about topics such as politics and religion.

In this research, we explored the growing phenomenon of polarization on social media through the theoretical lenses of MFT, investigating how individuals of opposing ideologies employ moral foundations. More specifically, we tested whether previous results on liberals and conservatives (e.g. Graham et al., 2009) would replicate in a text-based analysis performed on social media data.We, therefore, studied differences in moral foundations usage among users who followed the Labor and Conservatives related Twitter accounts. We reasoned that individuals interested in following Conservative accounts would more probably also be characterized by a Conservative ideology (especially if they did not follow the Liberal accounts), which would emerge in the contents they posted, and vice-versa individuals who followed

the Liberal ideology would share a Liberal ideology. After identifying the target accounts (followers) we gathered their tweets data in order to analyze it upon their moral foundation content.

We next conducted a second study among users who followed accounts related to different types of environmental activism. This choice is motivated by the abundant research that ties political orientation and view about climate change issues. Specifically, we wanted to test for possible overlaps and correlations between political orientation and actions on climate change. Moreover, given the lack of research about climate change activism and MFT, we wanted to explore moral foundations patterns among users who are more slanted towards different types of climate action.

Both studies will consider accounts from the UK context. The rationale behind this choice is bound to our measurement tools, which allow us to analyze only english text.

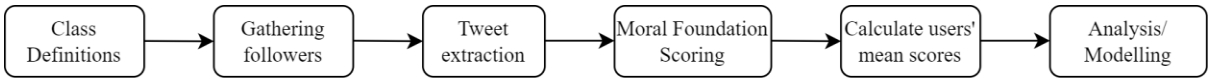

Figure 1: Flowchart of research's steps

Besides studying the differences in moral foundation usage from a purely psychological perspective, we trained a classification model for predicting the ideology of users using gathered moral foundation scores as predictors. Our aim is to test whether this methodological approach would allow the prediction of users preferences based upon their posted tweets.We believe that this goal in particular carries important implications. Being able to correctly predict the ideological preference of a user online can potentially open new paths for better communication in social and political scopes, leveraging techniques such as the already mentioned Moral Reframing. Moreover, our hope is that having such information could help reduce the overall polarization around topics, improving policies and fostering an overall better understanding of each other's worldviews without falling into excessive hyperpartisanship and unnecessary verbal animosity.

2. Materials and Methods

2.1 eMFD (Extended Moral Foundation Dictionary)

For the scoring phase, we adopted the Extended Moral Foundation Dictionary (eMFD; Hopp, Fisher, Cornell, Huskey & Weber, 2020), a thorough dictionary designed for extracting moral information from text corpuses. The eMFD offers several advantages over previous tools such as the Moral Foundation Dictionary (MFD, Graham et al., 2009; MFD 2.0, Frimer et al., 2017):

1. *Crowd-Approach:* Unlike previous approaches that relied on a limited number of expert researchers, eMFD employs a crowd-based approach for the dictionary creation. This method leverages the collective knowledge and perspectives of a diverse group of individuals, leading to a more accurate and unbiased representation of the moral language.

2. *Data-Driven Approach*: eMFD adopts a data-driven approach, focusing on identifying semantically similar words. By leveraging large amounts of data, the dictionary enhances the accuracy and comprehensiveness of moral information extraction.

3. *Continuous Scoring Method:* Unlike earlier dictionaries that employed discrete scoring methods, eMFD employs a continuous vector-based scoring method that assigns weights to each word of a corpus. This allows for a more nuanced understanding of moral dimensions by capturing subtle variations in moral expression, differently from previous approaches (e.g. MFD) where words mostly belonged to one and only one Moral Foundation.

4. *Consideration of the Syntactic Structure:* Differently from traditional "Bag of Words[1]" approaches, which treat a document as an unordered collection of words and focus only on the frequency of each word in the document, eMFD takes into account

[1] Bag Of Words (BOW): Model of text representation that does not rely on word order or syntax.

the syntactic structure of text. This enables a deeper analysis of the contextual meaning and moral implications present in the text.

Using eMFD to extract Moral Foundations from text has a wide array of benefits compared to previous dictionaries. Firstly, it has shown a much better accuracy in detecting morally relevant topics compared to other dictionaries; secondly, it more effectively detects distinctions between the moral language used by partisan news organizations (Hopp et al., 2021).
The eMFD has been used in a wide variety of researches in different fields to study social and economical phenomena, such as the impact of markets on morality (Harris, Colin and Myers, Andrew and Kaiser, Adam, 2022).

2.2 Frame Axis

Hopp and colleagues (2020) offer two scoring algorithms for the investigation of moral foundations through the eMFD. However these algorithms are unreliable when used with short texts - as in our case - since they rely on word count (Hopp et al., 2020). To overcome this issue, we used a scoring method stemming from the FrameAxis approach (Kwak, An, Jing, Ahn, 2020).

FrameAxis is an automated method to collect information about the framing (Kahneman, Tversky, 1979) of text. Manual annotations of the framings can be subject to individual biases and do not scale well on large amounts of text data.

Kwak et al. (2020) propose the FrameAxis, an unsupervised method for characterizing texts with respect to a variety of *microframes*, where each microframe is defined as a pair of antonyms words represented as vectors.

Mokhberian, Abeliuk, Cummings and Lerman (2020) implemented the FrameAxis approach to create a moral foundations scoring algorithm. The FrameAxis method for scoring moral foundations (Mokhberian et al., 2020) defines 5 semantic axes (microframes) – one for each moral foundation - in the latent space of word embeddings[2] and then calculates the relevance of any given text to those axes by

[2] Word Embeddings: Cluster of techniques to represent words into vector spaces

computing the cosine similarity between a corpus of text and the semantic axes. Each axis represents one moral dimension of the five proposed by the MFT; at the extremes of the axes are the semantically opposite poles of words related to the moral dimension. For example, for the *Care/Harm* dimension we find words such as “care”, “help*”* and “provide*”,* at the *virtue* pole and words such as “attack”, “violence” and “kill” at the *vice* one. Each document is evaluated upon the five axes on two main scores: *Bias* and *Intensity*.

Bias refers to the proximity of the document to a certain moral axis. It is obtained by calculating the weighted average of the cosine distance between each word of the document-vector and an axis. It is important to note that the absolute value of bias indicates how close a certain document is to the morality axis, while the sign reports the direction toward the pole of the axis. For instance, the word *help* would have a positive sign with respect to the *Harm/Care* dimension meaning that it’s oriented towards the virtue pole, while the word *attack* would have a negative sign.

Intensity is an index that captures how much a certain moral dimension is present in the document under consideration relative to the total distribution of documents. Intensity represents how much a document is relevant to a moral foundation. In other words, *Bias* indicates the proximity and the direction (positive or negative) of a document towards a microframe, while *Intensity* shows how much that microframe (moral foundation) is present over other possible options.

Research has shown that the FrameAxis approach ends up being more suitable for cases where the goal is to extract moral foundations scores from short text, such as tweets (Mokhberian et al., 2020).

2.3 Studies

In this research we conducted two studies, using 2 different comparison classes each. The expectancies with these studies are twofold and stem from two different epistemological goals. The first one, of a theoretical cut, is to assess whether results from past research on MFT and political views replicate with the tools and methodologies harnessed in our research.

Secondly, on a more practical side, we want to test whether classification models based on eMFD along with FrameAxis enable the prediction of the category a certain account falls into (i.e. whether an account follows liberal or conservative accounts) based solely on the text data posted on social media.
This second implication holds important consequences as knowledge about users' ideology can be used to suit communication styles to the target audience, empowering message's persuasiveness (Feinberg, Willer, 2013).
Moreover, through this work we want to explore and propose a general procedure to segment clusters of users on social media by leveraging moral foundations.

2.3.1 Study 1: Labor-Conservatives

In the first study, we chose two opposite political parties in the scope of the UK political scene, such as *Conservatives* vs *Labor*. This comparison is inspired and motivated by the abundant research on MFT and political engagement (Graham, Haidt, Nosek, 2009; Haidt, Graham, 2007; Stewart, Morris, 2021; Day, Fiske, Downing & Trail, 2014), showing that liberal individuals tend to rely more on individual moral foundations (*Harm*/*Care*, *Fairness*/*Cheating*), while conservatives show a higher employment of the binding foundations (e.g., Graham et al., 2009; Stewart, Morris, 2021).
However, past research on the topic heavily relies on questionnaires and controlled data collection scenarios while in our research we leverage a more ecological approach by analyzing posts on social media. Hence, potential different results can be yielded by different methodological approaches.
For example, Van Vliet (2021) employed a similar approach to our research, using the eMFD scoring algorithms (Hopp et al., 2021) to analyze moral language variations among UK politicians in regard to the Brexit context. The study shows that Labor politicians mostly relied on Authority and Care, while Conservatives scored significantly higher in Loyalty. The study also showed that Conservatives and Labor expressed the Authority foundation in different ways, using it to either question the current cabinet (Labor) or to support it (Conservatives).
It is worth noting that results in van Vliet's study (2021) do not replicate past results (e.g. Graham et al., 2009), showing incongruencies that might open up new questions on the methodologies and possibly on MFT itself.

2.3.2 Study 2: Insider-Outsider Activism

In our second test we sought to predict users' attitudes towards pro-environment communication. Research on moral foundation and environmentalism shows that self-identified political liberals tend to hold greater pro-environmental positions than conservatives. This might be related to the role of individualizing moral foundations (Milfont, Davies & Wilson, 2019; Fielding, Hornsey, 2016; Pew Research Center, 2019).

However, pro-environmental positions can have significant differences in their approach to address climate issues.

Here, we consider two different pro-environment approaches based upon Ozola's (2011) framework. We distinguish between Outsider and Insider activism.

Outsider activism is an activism approach that consists of social group strikes, such as confrontational demonstrations, sit-ins, blockades and direct confrontation (Ozola, 2011).

This type of activism is typically embraced by individuals lacking access to institutionalized platforms. It involves addressing social issues through collective action and participation in social movements organizations (SMOs) to bring about remedies and change (Briscoe, Gupta, 2016).

On the other hand, Insider activism acts by putting pressure on organizations and governments through petitions, policies, lobbying and financial investments to solve the proposed issue (Ozola, 2011)

Our first goal with this study is to investigate whether there are differences in moral foundations scores between users classified as Insider versus Outsider activism sympathizers, based on the accounts they follow. Research has shown that climate change over the years has become a polarizing issue across social media platforms (Falkenberg et al., 2022). Here, we sought to find whether polarization reflects onto users who seek different types of activism content on social media, exploring possible differences in terms of moral foundations usage between the two groups.

Second goal is to assess whether there are overlaps between political ideology and activism preference from a morality perspective. Our hypothesis is that users who follow liberal accounts might share similarities in terms of moral foundation usage with users who follow Outsider activists accounts.

Thirdly, as in Study 1, we sought to use moral foundations scores to train and assess a classification model to predict users among the two classes.
We believe that exploring differences between these types of users can enable benefits in terms of climate communication on social media, potentially reducing polarization around the topic of climate change.

2.4 Data Collection Pipeline

The objective of our research was to classify Twitter users into distinct social groups (classes) based on their moral foundations mean scores. Consequently, the initial step involves defining specific classes to label the users for the analysis.

2.4.1 Classes Definitions

The first step to conduct our studies was to give a formal class's definition.
First, we compiled two lists of Twitter accounts closely related with one class each (see Table 1). Hence, each class C_j is defined by a list of k_j representative Twitter accounts.
Membership of a user u_i to class C_j was established only if the user appeared to follow each of the k_j accounts in C_j definition.. Moreover, to ensure the robustness of the defined classes, we only considered the symmetric difference between the opposing groups. In other words, only users who followed Conservative accounts *and did not* follow Labor accounts were included.

For the class definitions in Study 1, we used a mix of politicians' and tabloids' Twitter accounts that are known to be oriented towards either a Conservative or Labor stance.

Similarly, class definitions in Study 2 are obtained by a list of accounts that are known to be related with either an Insider or Outsider approach to environmental issues.Average followers count (in millions) for the accounts defining the Conservative class is $\underline{x}_{\text{conservatives}} = 2.75$, $\underline{x}_{\text{labor}} = 3.62$ for the Labor class. As for Study 2, Insider activists accounts count on average $\underline{x}_{\text{insider}} = 54\text{k}$ and Outsider activist accounts have an average number of followers of $\underline{x}_{\text{outsider}} = 172\text{k}$.

Table 1. Class definitions for the two studies.

Class (C_j)	Accounts
Conservatives	@BorisJohnson, @DailyExpress, @Telegraph, @TheSun
Labor	@JeremyCorbyn, @novaramedia, @OwenJones84, @guardian
Insider Activist	@ClientEarth, @EnergySvgTrust, @thecarbontrust
Outsider Activist	@JustStop_Oil, @XrebellionUK, @GreenPeaceUK

2.4.2 Followers and Tweets collection

Here we describe the collected lists of users. For Study 1, we gathered N_{Labor}=67,974 *labor*-labeled users and $N_{\text{Conservatives}}$=77,976 *conservatives*-labeled users; for Test 2, we gathered N_{Insider}= 2,292 users and N_{Outsider}=9,698 users.

We then sampled from the users lists to extract class bounded tweets.

For the tweets extraction we didn't set any limit per user: we extracted every possible tweet up to the limit imposed by the APIs[3] themselves which is 3,200.

We gathered N=4,004,545 tweets for Labor class, N=4,003,950 for the Conservatives class, N=2,331,055 for Insider activism and N=4,003,245 for Outsider activism. Each tweet is associated with the corresponding user ID.

Additionally, a smaller dataset was created, comprising tweets solely containing the hashtag "#Brexit" to focus exclusively on this polarizing topic. The rationale behind this selection stems from our assumption that discourse surrounding polarizing issues, such as Brexit, would evoke a more lively moral language. This decision aims to provide a more comprehensive lens through which to explore and discern the predominant moral foundations employed by each class.

[3] Application Programming Interfaces (API): Sets of procedures that allow communication between two computers. In our case APIs are being used to retrieve data from Twitter's databases.

Tweets lists were then used to calculate and map scores for each of the 5 moral foundations to every user. Retweets were excluded from the analysis.

2.4.3 Moral Foundation Scores

We applied the eMFD dictionary with the Frame Axis scoring approach to obtain the moral foundation's scores. Due to the algorithm's nature and the dictionary employed, non-English tweets received a score of 0 in each category, resulting in their exclusion from the dataset.

The outcome of this process is a dataset comprising moral foundation scores for each tweet, measured on both the *Intensity* and *Bias* metrics described earlier. Next, to effectively classify random users by categories, we computed the mean scores of all the tweets of each user for each moral foundation.

Given possible imbalances in users tweet count, we added a *tweet_count* variable to the final dataset to keep it under control during the model validation phase.

The final datasets consist of a total of N=6,902 Conservatives users and N=4,410 Labor users for Test 1; N=2,065 for the Insider activism class and N=4,773 for the Outsider activism group.

Dataset filtered upon the "#Brexit" hashtag counts N=419 users for the Conservative class and N=934 users for the Labor class for Test 1; N=666 for the Insider activism class and N=887 for the Outsider one.

The resulting datasets from this procedure were then used to train classification models such as logistic regression. Our goal was to study which moral foundations would better explain the distinction between the classes and to assess the overall model's prediction power.

3. Results

3.1 Study 1: Labor-Conservative

3.1.1 Full Tweets Dataset Analysis

Firstly, we compared how Bias and Intensity scores are distributed among Conservatives and Labor labeled users on the entire tweets dataset. Figure 2 and 3 shows the score distributions for both metrics between the two classes.

Figure 2: Bias distributions comparison between Labor and Conservatives labeled users

The analyses on the bias dimension (see Boxplots in Figure 2), which reflects the polarity direction of a moral foundation of a given moral foundation is prevalent in the posts of the selected accounts, showed small but significant differences in the moral foundation orientations between Conservatives and Labor labeled users.

Surprisingly, a small difference emerged in the *Authority/Subversion* Bias metrics, "Cohen's $d = 0.19$, p-value < 0.0001", indicating that Labor users appealed to legitimate authority and respe ct for traditions slightly more than Conservative ones. Moreover, the Bias metric shows a significant yet small difference in the *Care/Harm* dimension in favor of the Labor accounts, "Cohen's $d = 0.15$, p-value < 0.0001". However, differences in moral intensities (see Boxplots in Figure 3) do not provide sufficient evidence to effectively replicate past results nor to draw any significant conclusions.

A glance at the Intensities values correlation matrix (see Heatmaps in Figure 4) provides insights into the relationship between the Intensity and Bias metrics.

In fact, there seems to be a generally low correlation between each pair of metrics for every moral foundation.

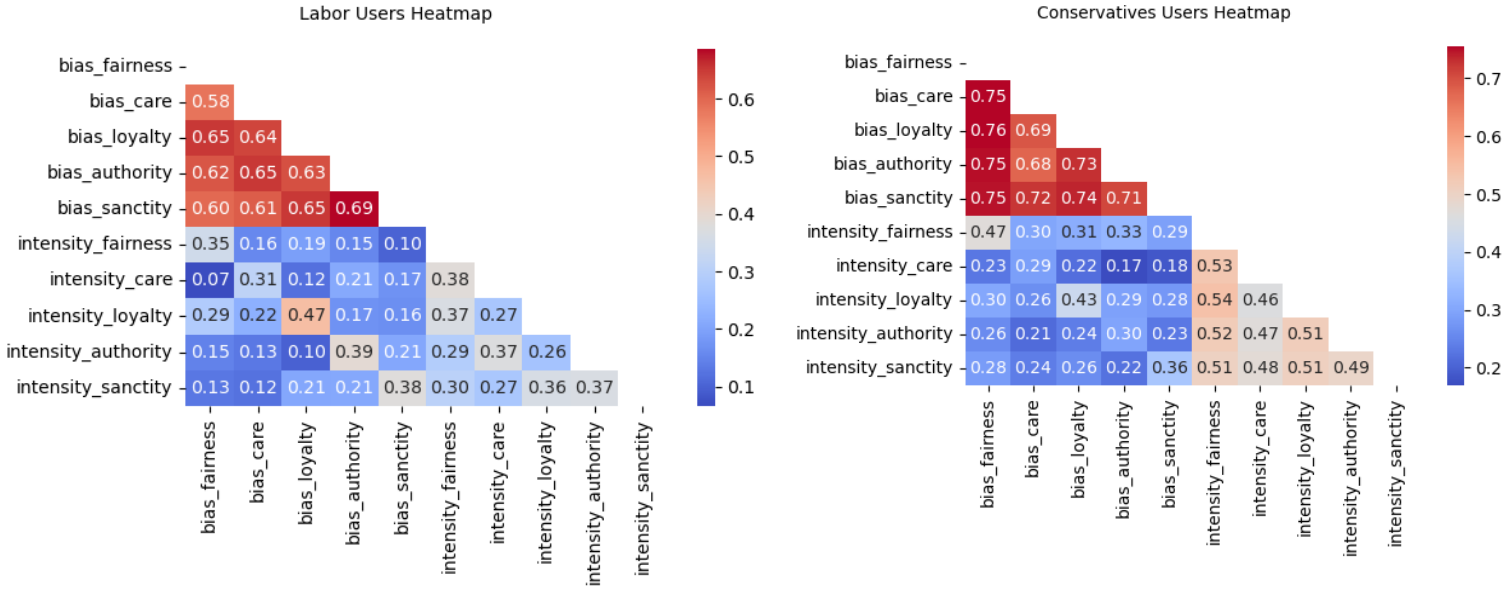

Figure 4: The heatmap on the left shows the correlation between the variables for the users who follow Labor accounts. Heatmap on the right reports the correlation between variables among users who follow Conservative accounts.

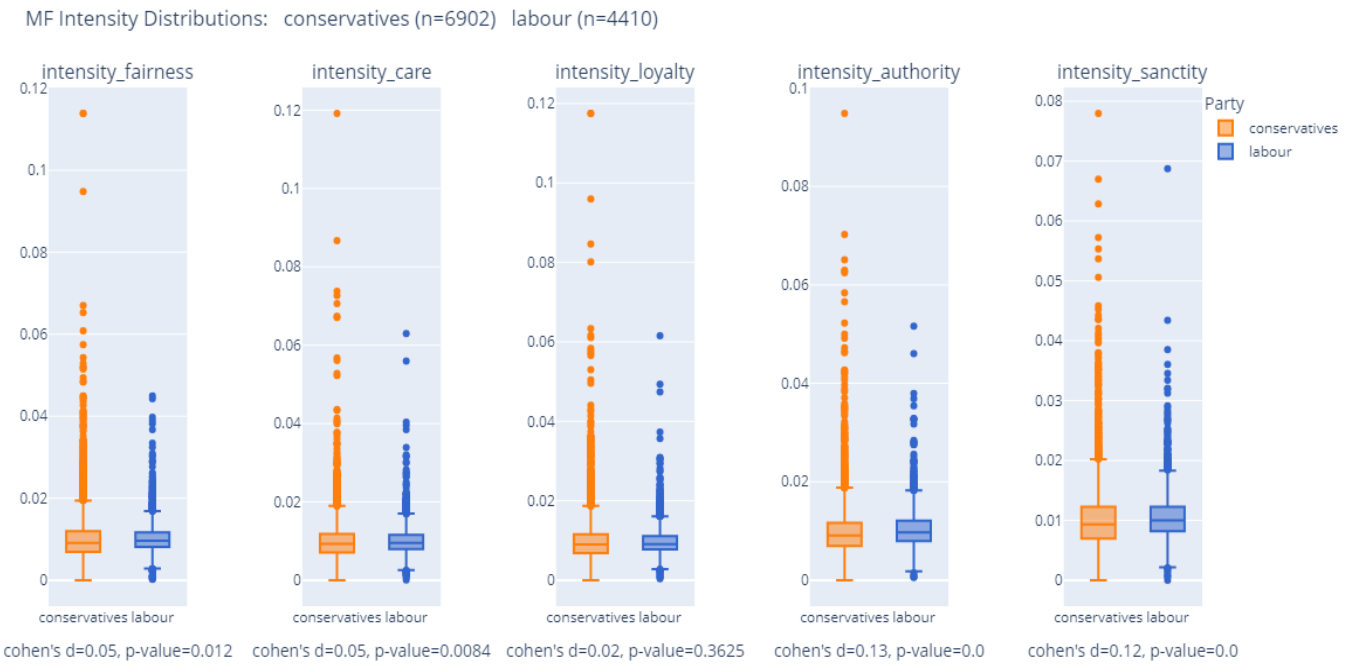

Figure 3: Intensity scores for Conservatives and Labor users

In general, we believe that these results pave the way for two possible interpretations. Firstly, it could be possible that no strong differences exist between the two classes. In other words Conservatives and Labor users would not show differences in moral foundations usage among text. This interpretation however seems to clash with most of the literature around the topic (e.g Study 4 in Graham et al., 2009; van Vliet, 2021). Another possible explanation for our results comes when noting that the analyzed data stems from day to day tweets. In fact, we considered every users' posted tweet, without applying any filtering around topics. This could've influenced our results by flattening the scores between the classes, not letting any difference emerge.
In support of this, we can remark on the fact that most research that investigates the differences between individuals holding different political views tend to leverage highly contextualized topics (e.g. Graham et al., 2009; Milesi, Alberici, 2016).
This leads us to the conclusion that moral differences between Conservatives and Labor users may not emerge in day to day tweets, necessitating a bounded context topic to emerge.
It can then be worth analyzing the same data after filtering it by topics that are morally relevant for the classes studied.

Note however that small effect sizes differences have been reported in some of the foundation's bias metrics. We will look forward to seeing whether these results replicate when considering only context related tweets.

3.1.2 Analysis of Tweets Filtered by Topic: #Brexit

As already previously mentioned, our assumption is that morality gets expressed more when the topic is morally relevant or polarizing.

To test this hypothesis, we performed another analysis of the same data, this time filtering tweets by hashtag.

Figure 5 and 6 shows users Bias and Intensity scores calculated by only taking into consideration tweets that contained #Brexit hashtag.

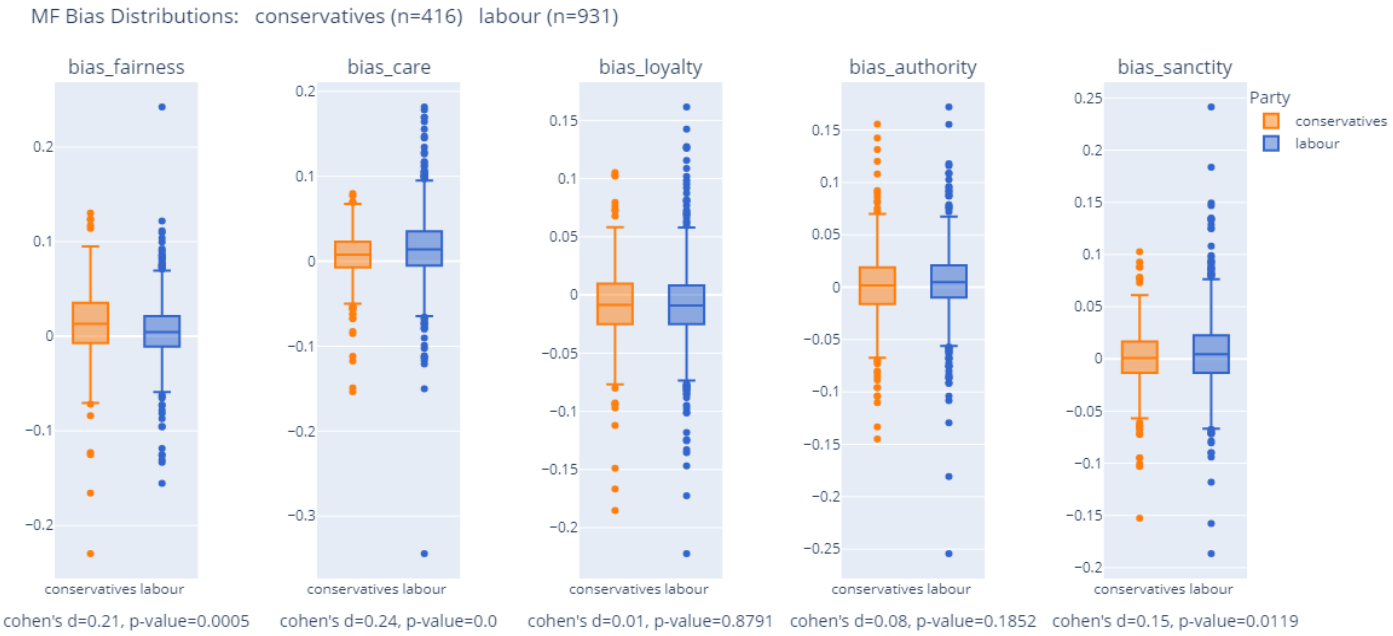

Figure 5: Conservatives and Labor labeled users Bias scores obtained from tweets containing #Brexit hashtag

Results confirm our hypothesis that selecting tweets around a polarizing topic would show more pronounced differences in the use of moral foundations between the two groups.

As for the Bias metric, differences seem to be thin (Figure 5), leading us to conclude that Conservatives and Labor labeled users share a common orientation (vice/virtue)

towards which moral foundations are being expressed. Is it possible to notice small differences in *Fairness/Cheating* foundation "Cohen's d=0.21, p-value<0.0006" and *Harm/Care* "Cohen's d=0.24, p-value<0.0001". Bias differences suggest that Conservatives hold a tendency to appeal to the *Fairness* words more than Labor, while the latter seem to be more prone to frame their language toward the *Care* foundation rather than the *Harm* one.

Intensity metric (Figure 6) shows much stronger differences in terms of effect size. As results in Graham et al. (2009) suggest, given the linguistic nature of our data we would expect to replicate results in their Study 4, observing higher scores in left wing users not only in *Harm/Care* and *Fairness/Cheating* foundations, but also in the *Loyalty/Cheating* one.

Conservative users seem to endorse the *Fairness/Cheating* "Cohen's d = 0.73, p-value<0.0001" foundation much more than Labor users, finding that clashes with classic literature on the topic (Graham et al. 2009).

On the other hand, Labor users tend to score much more than Conservative users on the *Care/Harm* dimension "Cohen's d = 0.78, p-value<0.0001" which goes along with the aforementioned literature.

Interestingly, Labor labeled accounts score much higher on the *Sanctity/Degradation* dimension, which is a moral foundation usually associated with the right and conservatives individuals.

Moreover, Conservative accounts tend to appeal to the *Authority/Subversion* foundation more than left wingers.

It is important to note that, even if significant results are reported, the moral foundations patterns we have found do not replicate older studies.

Hence, results suggest that differences in moral foundations usage do exist between Conservatives and Labor users and that these differences tend to emerge in context where a polarizing topic is discussed, such as Brexit.

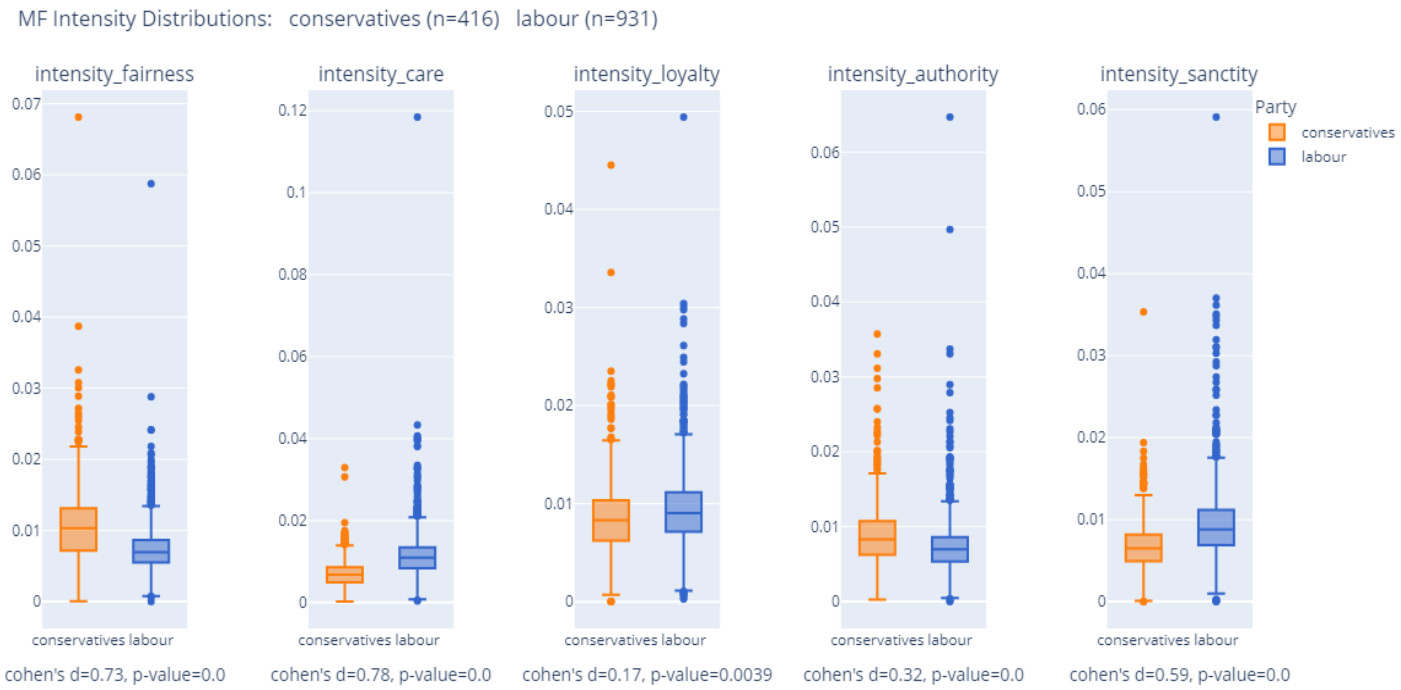

Figure 6: Conservatives and Labor labeled users Intensity scores obtained from tweets containing #Brexit hashtag

3.1.3 User Classification on Political Ideology

The second objective of our research was to classify users based on their tweets timeline.

We used t-Distributed Stochastic Neighbor Embedding (t-SNE; van der Maaten, Laurens & Hilton, 2008), a visualization technique widely employed for high dimensional data visualizations. Results of the t-SNE visualization suggest that Intensity scores (see Scatter Plot in Figure 7) might be good predictors for classifying users on political ideology while Bias scores (Scatter Plot, Figure 8) seem to not carry enough information for discriminating between the two classes.

Here, we use the filtered data discussed in 3.1.2 to train a logistic regression model to ask the following questions: which moral foundation do majorly impact the political ideology of a user ? How accurately can we predict the ideology of a user through moral foundation scores ?

Figure 7: t-SNE representation of Conservatives and Labor labeled users Intensity scores

First, we assigned each data observation to a *tweet range*, based upon the number of tweets retrieved for a user. We performed stratified sampling in the train/test split in order to get a more balanced sample and to avoid potential bias caused by tweet count. We then performed a 70-30 train/test split of the entire dataset.

Table 2 shows the logistic regression coefficients *p*-value and confidence intervals. Coefficient values show the importance of the Intensity metric in distinguishing between the two classes, while highlighting the small impact of the Bias.

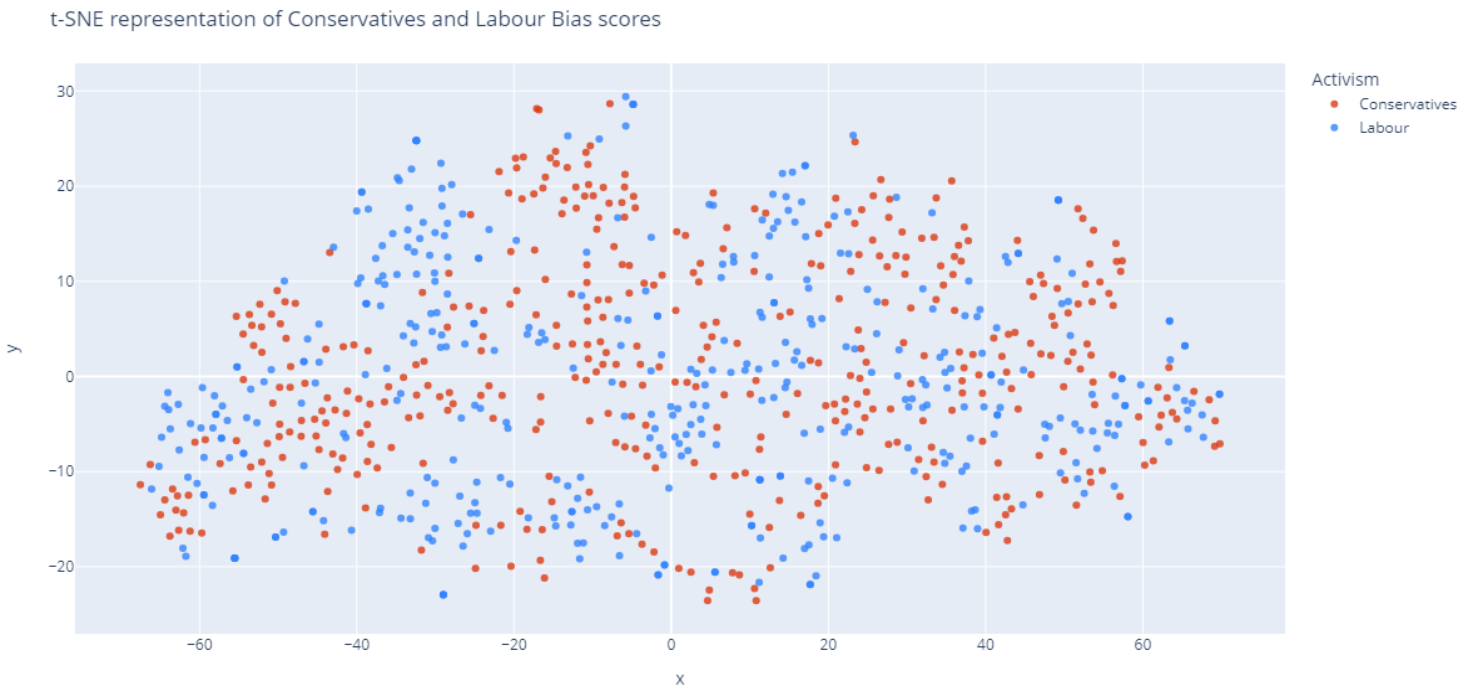

Figure 8: t-SNE visualization of Conservatives and Labor labeled users Bias scores

Moreover, Bias *p*-values suggested that Right and Left wing users expressed their moral foundation in the same way in terms of orientation. The important differences reside in which moral foundation they majorly relied on.

Tweet count was not a significant predictor.

Table 3 shows the classification report for the logistic regression model.

The model could correctly predict both Right and Left labeled users with good precision ($precision_{conservatives}$= 0.82, $precision_{labor}$= 0.88), while struggling to retrieve Right users ($recall_{conservatives}$ = 0.75).

The Area Under the Curve (AUC) metric reports a score of 0.93, indicating that the model we obtained is good at predicting between Left and Right users when they are discussing Brexit on social media.

Table 2. Logistic regression coefficients (Labor-Conservative classification)

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.1305	0.32	0.408	0.683	-0.496	0.757
bias_fairness	7.9596	5.59	1.424	0.155	-2.997	18.916
bias_care	-4.9793	5.82	-0.856	0.392	-16.387	6.428
bias_loyalty	-7.7491	5.826	-1.33	0.183	-19.167	3.669
bias_authority	-5.9078	5.827	-1.014	0.311	-17.328	5.513
bias_sanctity	1.6794	6.116	0.275	0.784	-10.309	13.667
int_fairness	440.4643	41.904	10.511	0.0	358.334	522.594
int_care	-425.6622	43.99	-9.676	0.0	-511.88	-339.444
int_loyalty	-47.2837	42.803	-1.105	0.269	-131.176	36.608
int_authority	202.8469	33.734	6.013	0.0	136.729	268.965
int_sanctity	-306.4789	41.623	-7.363	0.0	-388.059	-224.899
tweet_count	0.0089	0.005	1.838	0.066	-0.001	0.018

Table 3. Classification report for the logistic regression model

Class	Precision	Recall	F1-score	Support
Left	0.88	0.92	0.90	272
Right	0.82	0.75	0.78	133
Avg	0.85	0.84	0.84	405

3.2 Study 2: Insider-Outsider

3.2.1 Entire Tweets Dataset Analysis

In this section we discuss results obtained by the comparison of users who stick to an Insider environmental activism versus those who follow an Outsider one. Figures 9 and 10 respectively show the Bias and Intensity distributions for the two classes in the full tweets dataset analysis.

Insider activists seem to report higher Bias scores in every moral foundation, suggesting that they tend to communicate through a more virtuous language in their day to day life.

Study 2 reports also significantly higher effect sizes than Study 1 "lowest Cohen's d = 0.58, p-value<0.0001".

Conversely, Intensity metric doesn't show any significant difference, except for the *Authority/Subversion* moral foundation "Cohen's d = 0.19, p-value<0.0001".

As shown in Milfont, et al. (2019), individualizing moral foundations are a great predictor of pro-environmental behaviors. Therefore, we expected that Individualizing moral foundation Intensity scores between political classes and activists classes were significantly different.

Results coming from an independent sample t-test disprove our hypothesis, showing that Intensity means scores in *Fairness/Cheating* and *Harm/Care* foundations are not significantly different between the political and activism classes (Figure 9).

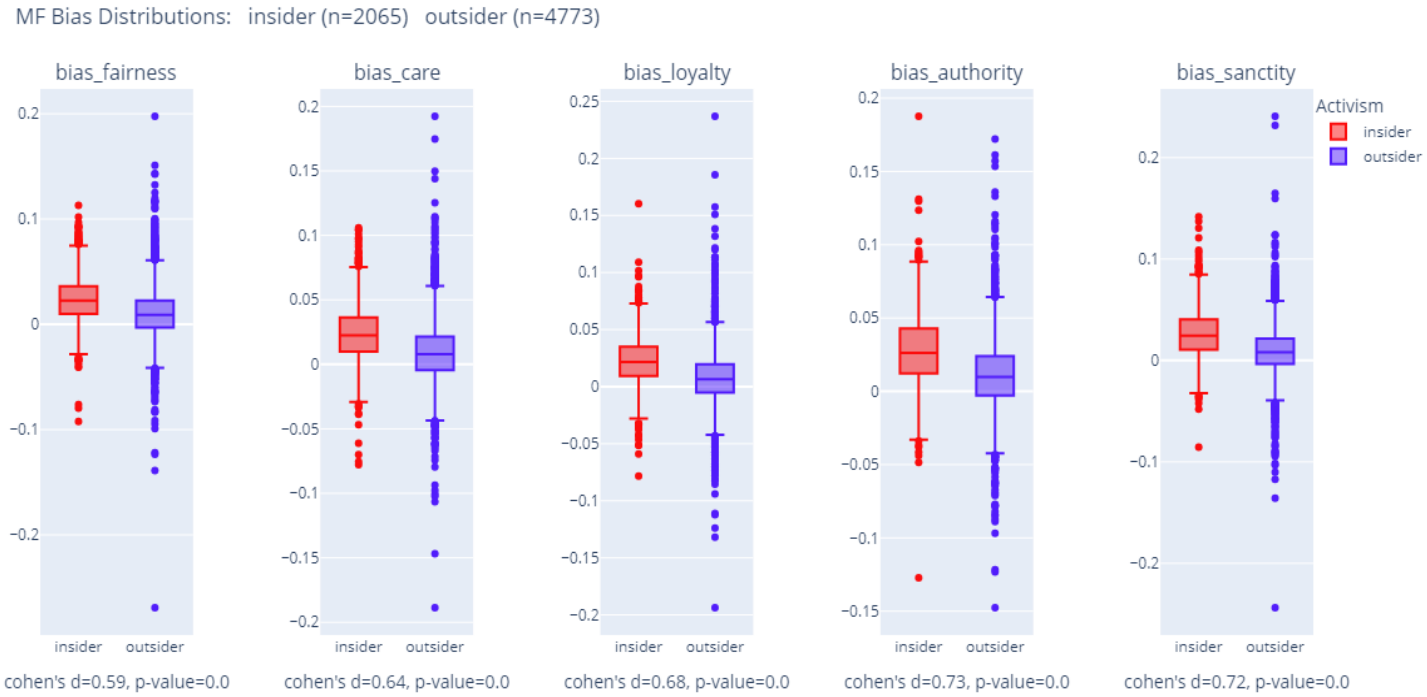

Figure 9: Bias distributions comparison between Insider and Outsider activist labeled users

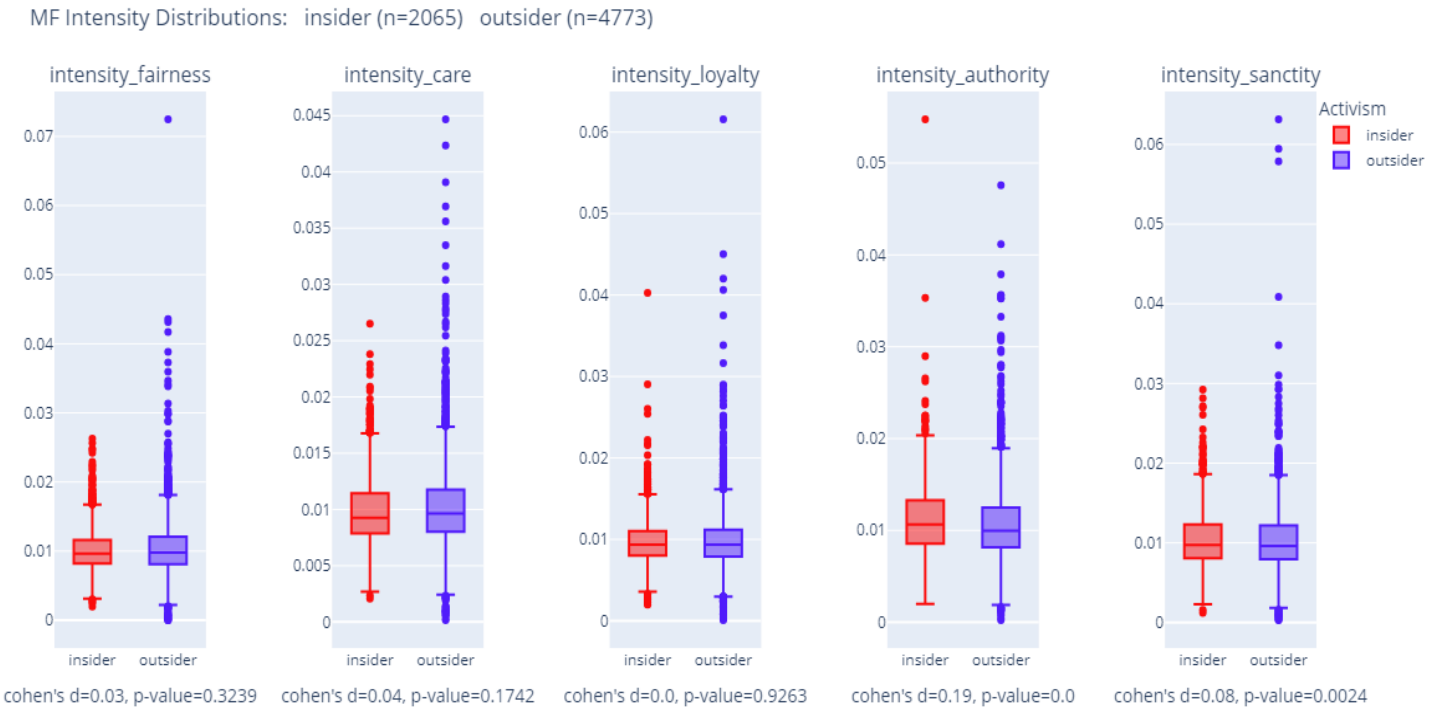

Figure 10: Intensity distributions comparison between Insider and Outsider activist labeled users.

3.2.1 Analysis of Tweets Filtered by Topic: #ClimateChange

Here, we repeat the same process employed in section 3.1.2, this time using the #climatechange hashtag to filter tweets.

Consistently with results from section 3.2.1, Insider class reports higher or equal to Bias scores in every moral foundation except, this time, for *Authority/Subversion*, which Outsider activists followers tend to address slightly more "Cohens'd=0.24, p-value<0.0001" (see Figure 9).

Huge "Cohen's d=1.2, p-value<0.0001" differences can be seen in the *Harm/Care* moral foundation usage (see Boxplots in Figure 9).

Lower overall Bias scores for Outsider activism accounts followers may suggest the employment of a more confrontational language when discussing climate issues.

Intensity scores (see Boxplots Figure 10) also show great differences in comparison with the results obtained without filtering tweets by hashtag (see Boxplots Figure 9).

Users who tend to follow Insider activists accounts tend to engage more in *Harm/Care* and *Loyalty/Cheating.* Users with Outsider activism preferences tend to score higher in both *Authority/Subversion* and *Sanctity/Degradation*.

In general ,we again notice different patterns of moral foundation usage when selecting tweets around a specific topic.

Moreover, large effect sizes may suggest some problems with the data, especially with the class definition we came up with.

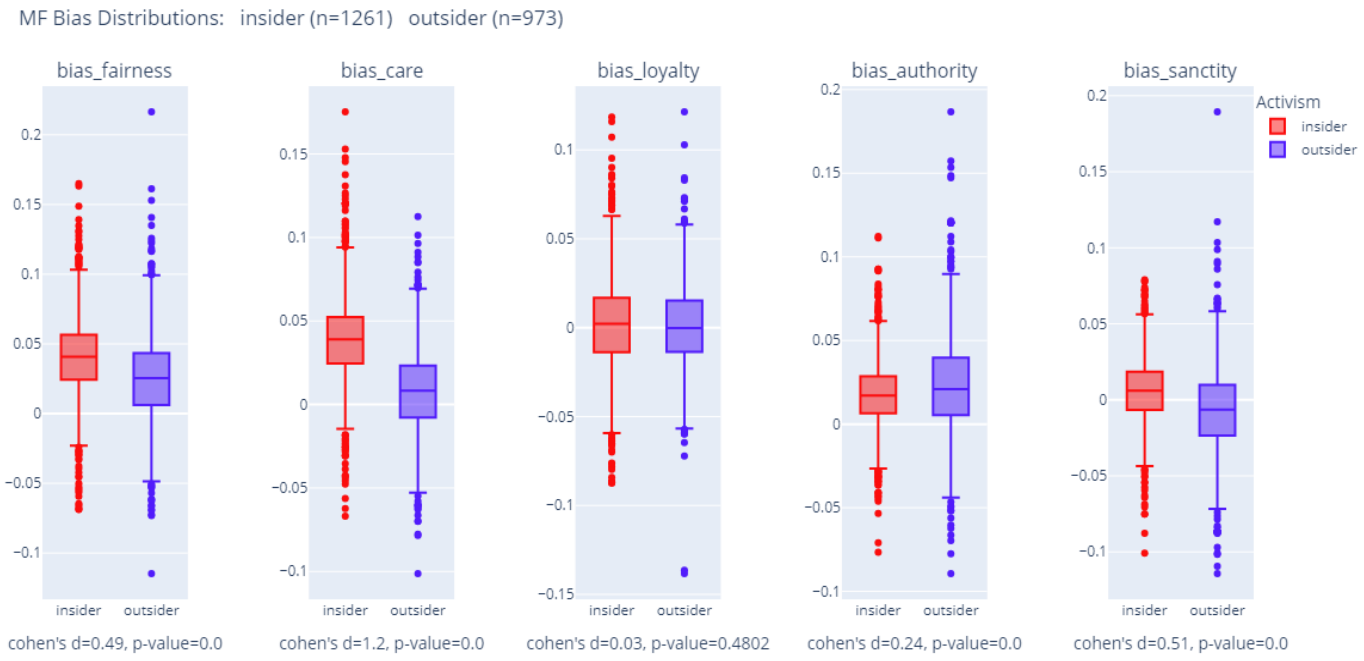

Figure 11: Bias distributions comparison between Insider and Outsider activist labeled users. Only tweets containing #climatechange hashtag are considered.

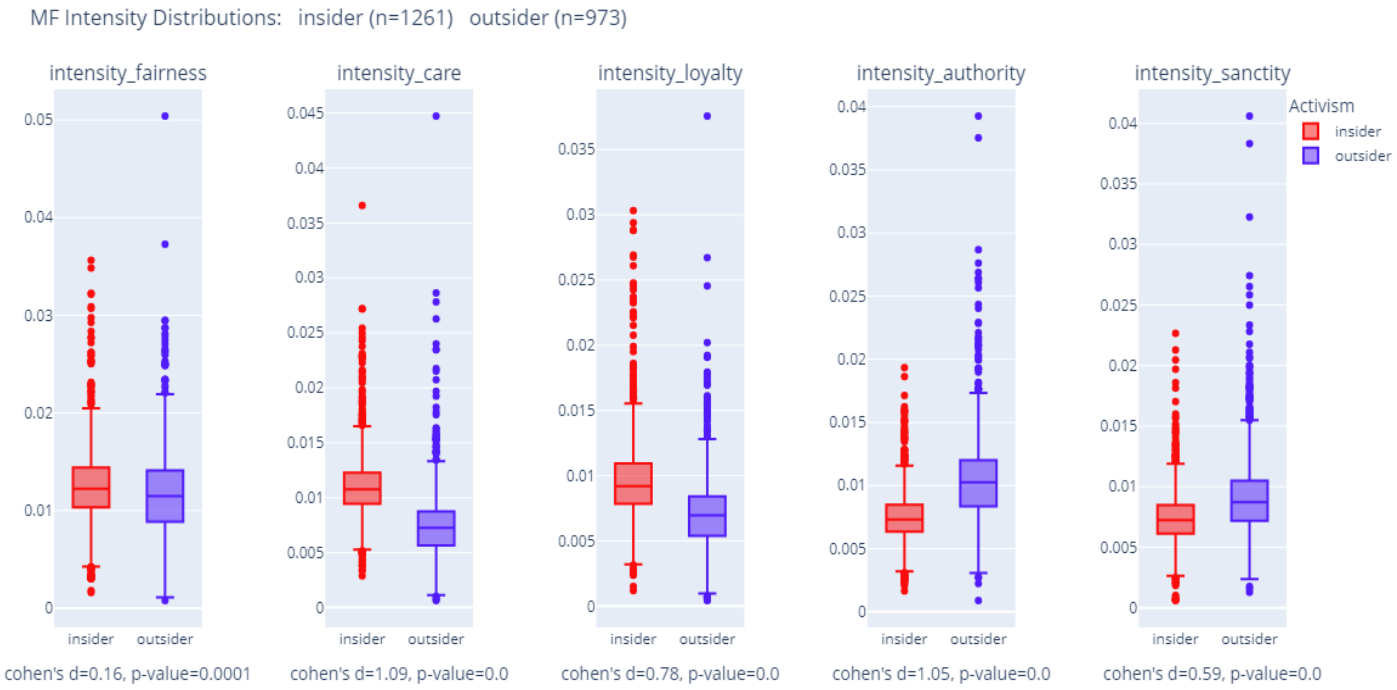

Figure 12: Intensity distributions comparison between Insider and Outsider activist labeled users. Only tweets containing #climatechange hashtag are considered.

3.2.2 User Classification on Activism Preference

As already shown in section 3.2.1, followers of either Insider or Outsider activism accounts tend to report high levels of feature separation. Thus, here we expected our classification model to perform well on the data. We trained a logistic regression model using the filtered data described in section 3.2.1 to see how moral foundations metric scores interacted and how accurately we could predict users on the classes.

A quick glance at the t-SNE scatter plot for the Intensity scores suggests that intensity features seem to discern between the two classes (see Scatter plot in Figure 13).

Table 7 shows logistic regression coefficients, *p*-value and confidence interval. Bias metrics tend to be less impactful in distinguishing between the two classes, while Intensity scores report the highest absolute values. As already shown in section 3.2.1, features related to *Authority/Subversion*, *Loyalty/Cheating* and *Harm/Care* seem to be the most important ones in the logistic regression model.

Table 6 shows the classification report revealing very high classification metrics; Area Under the Curve (AUC) metric reports a surprising score of 0,99.

Such high scores raised our warnings towards the model, convincing us to perform a *k*-fold cross validation for our model for a better understanding and assessment of the model.

Figure 14 shows metrics evaluation across 5 folds. Precision, accuracy and f1-score remain high across each fold, leading us to conclude that the model is good.

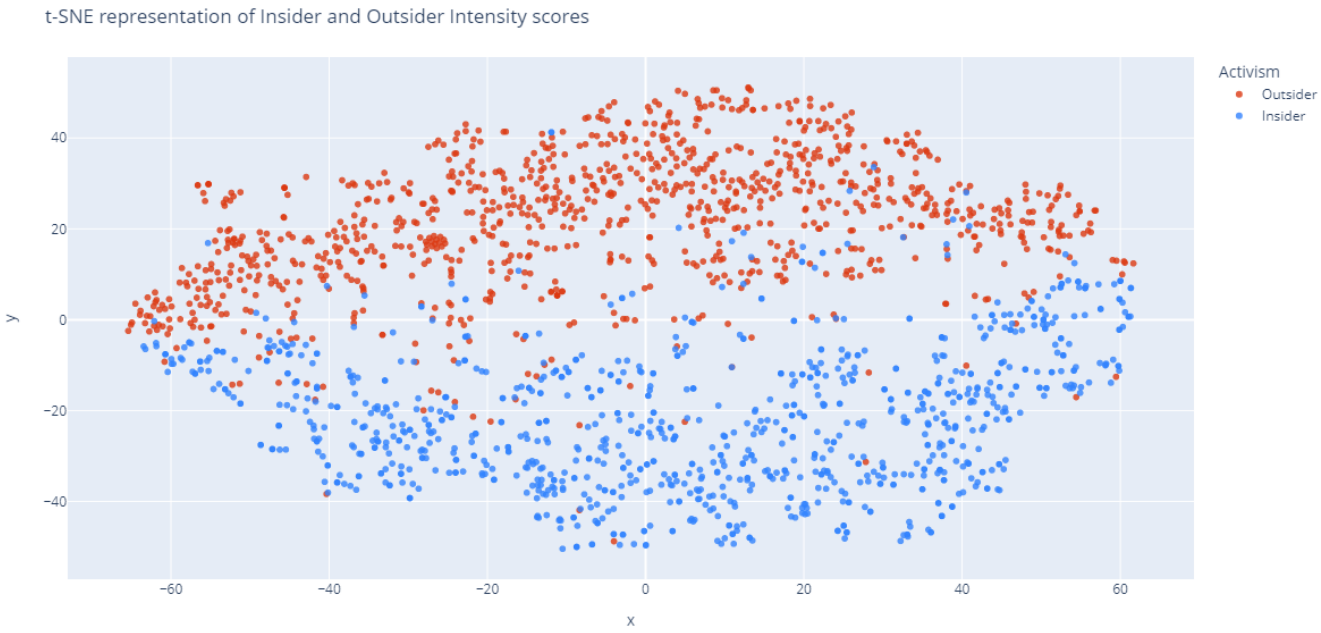

Figure 13: t-SNE scatter plot of Insider and Outsider Intensity scores

Table 6. Classification report for the logistic regression model

Class	Precision	Recall	F1-score	Support
External	0.96	0.95	0.95	280
Internal	0.96	0.97	0.97	391
Avg	0.96	0.96	0.96	671

Table 7. Logistic regression coefficients (Insider-Outsider classification)

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.0232	0.529	-1.935	0.053	-2.059	0.013
bias_fairness	10.9422	10.043	1.09	0.276	-8.742	30.626
bias_care	98.0759	10.465	9.372	0.0	77.565	118.587
bias_loyalty	-37.3682	9.473	-3.945	0.0	-55.934	-18.802
bias_authority	-84.4122	10.084	-8.371	0.0	-104.176	-64.649
bias_sanctity	9.9635	10.671	0.934	0.35	-10.951	30.878
int_fairness	122.0292	53.822	2.267	0.023	16.539	227.519
int_care	580.745	56.503	10.278	0.0	470.002	691.488
int_loyalty	548.8605	54.882	10.001	0.0	441.294	656.427
int_authority	-848.8223	74.289	-11.426	0.0	-994.426	-703.218
int_sanctity	-443.0206	60.944	-7.269	0.0	-562.469	-323.573
tweet_count	0.007	0.004	1.651	0.099	-0.001	0.015

Nonetheless, while the model might be good, we advance the hypothesis that the data employed for the Study 2 analysis is flawed at its definition. We will cover more about the potential pitfalls and data problems in the Discussion (see section 4).
We abstain ourselves from drawing conclusions about the target classes for this study.

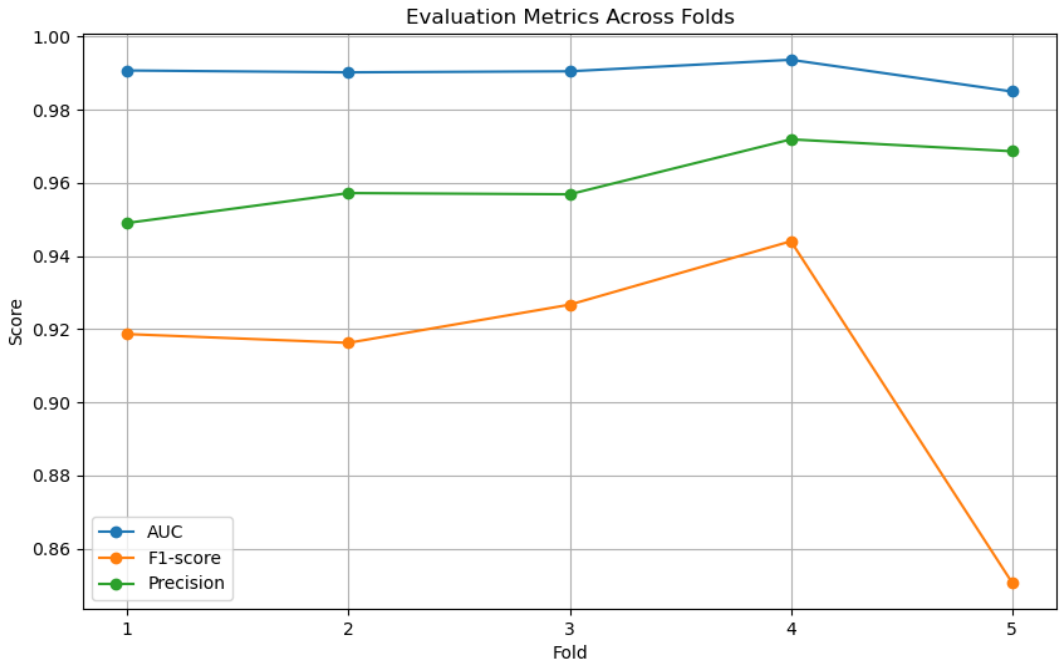

Figure 14: Evaluation Metrics for AUC, F-1-score and Precision across 5 folds

4. Discussion

Through the studies performed in our analysis we addressed psychological questions while testing an approach for predicting social media users' behavior.

In both studies we showed differences in the moral foundation usage between the chosen classes. Significant results emerged, especially when analyzing tweets related to specific topics such as political and social issues.

However, we have also shown how our findings fail to replicate results found in previous studies. This raises some uncertainties about the validity of the methodological approach we have harnessed. In this section I will summarize the findings and discuss the epistemological value of this research method, after systematizing the psychological and theoretical insights that can be drawn from our trials.

4.1 Results

In both studies we observed differences in the usage of moral foundation between the considered classes (section 3).

In particular, Study 1 showed larger differences between Conservatives and Labor users' moral profile when specifically considering tweets discussing a relevant topic such as Brexit.

While these results provided support to the body of work that states that there are differences in how liberals and conservatives see the world (Graham, Haidt, & Nosek, 2009; Hunter, 1991; Jost, 2006; Lakoff, 2004), they don't completely align to previous studies findings in *how* moral foundations are being endorsed (e.g. Graham et al., 2009).

Several potential factors can be addressed for this.

First, most of the previous studies on moral foundations differences between liberals and conservatives were designed and conducted in controlled environments, making use of self-report methodologies such as questionnaires (Frimer, 2019).

Our studies have been conducted by gathering social media conversations, using a scoring algorithm (Mokhberian et al. 2020) to observe moral foundation usage among individuals.

However, studies that investigated moral foundations through language (e.g. Study 4 in Graham et al. 2009; Frimer 2019) also show different findings from the one we obtained.

As our data stems from Twitter users' conversations (tweets), it may carry more ecological and organic value but also may not necessarily represent the real population (Chen, Duan Yan, 2021). Hence, we should hesitate to generalize our conclusions to anything outside subsets of the Twitter population.

Second, our class definitions are based upon the act of following a set of arbitrarily chosen Twitter accounts and rely on the unproven assumption that the act of following a Twitter account reflects a positive attitude towards it.

In Study 2, we found very large and significant differences between accounts following Insider activism related accounts and those following Outsider ones. We suggest that this result was mainly due to our class definitions. In fact, in Study 2, we defined Insider and Outsider classes through Twitter accounts with rather low

follower count (see section 2.4.1). Using niche accounts could've led us to select extremely polarized users that do not reflect the reality of an average user. Furthermore, it has been proved how algorithmic political communication has been exploited during US elections in 2016 (Howard, Kollanyi, & Woolley, 2016). Hence, we might have inadvertently spoiled our data by including politically slanted bots into our analysis.

Despite our hesitancy in drawing theoretical conclusions about the differences in the usage of moral foundations among the considered classes, we believe that our research approach may become - with the correct adjustments - a useful tool for predicting users ideological orientations and preferences from media posts.

Our classification report for the Conservatives and Labor classes (see section 3.1.2) shows that a logistic regression model can correctly predict with reasonably good accuracy between left and right wing users when collecting conversations about the Brexit topic. However, it is important to show possible hindrances and factors that could lead to bad models. In section 3.2.1 we've reported suspiciously high differences between the Insider and Outsider activism classes. Here we list some potential pitfalls we've learned while carrying out our research that could provide alternative explanations of our results, especially during the class definitions part. We came up with guidelines through our intuition and experience, yet we did not formally test for these through proper scientific trials, which we intend to perform in the future.

- Account Selection: Classes must be defined by accounts that faithfully represent their essence. Solid domain knowledge is mandatory to create an accurate representation of the group of interest.

- Number of Accounts: While classes ought to be defined by representative accounts, it is also important to take into consideration the number of accounts for class definition. Too few accounts would probably be not enough to be able to sample users that are actual members of the class (social group) of interest; too many could potentially select an overly extreme subset of the class.

- Followers Imbalances in class definitions: We suggest taking into consideration the number of followers of definition accounts. Our intuition is

that the action of following a niche account carries more information about the attitude of the following user towards the account than the act of following a popular one. Hence, taking into account users that simultaneously follow a list of obscure accounts might produce unrealistic results.

Our intention was to showcase and test a new possible methodology for analyzing and predicting social media behaviors inside the Moral Foundation Theory framework. More tests and replications ought to be done in order to gain a clearer picture of the epistemological, theoretical and practical value of our approach.
About the latter, model evaluation (section 3.1.3) suggests that regardless of the direction of the results, our methodology is efficient in finding significant differences between users.
Regardless of the content and orientations of the differences in the usage of moral language, a simple classification model such as logistic regression seems to be able to discriminate users between different classes.

4.2 Research implications

In the Introduction we talked about how the MFT can be used in communication to enhance messages' persuasiveness through Moral Reframing (Feinberg, Willer, 2013).
Haidt (2012) discusses how most of the disagreements between individuals might be due to underlying differences in the moral matrix.
As our results suggest that it can be possible to predict users' ideology through our approach, it can be then worthwhile considering to use this method to reduce the moral gap between individuals of different ideologies (and hence different moral foundations). It can then be worthwhile to consider the implementation of our method to predict users' ideology and target them with messages that suit their moral foundations pattern.
This can be applied in many fields such as social marketing, environment communication, health, policies and many others.
On the other hand, it can raise concerns about users' data privacy, manipulation and malevolent intentions.

4.3 Future steps

As already mentioned, our future goals are to investigate the validity of our approach in yielding knowledge about the usage in moral foundations among groups, especially Liberals and Conservatives.

Particularly, we plan to discover whether results such as the moral foundation pattern found in Graham et. al (2009) and replicate on social media by using language-based approaches.

However, we do not intend to stop at studying differences on political issues.

We would like to investigate whether our approach can be able to discern users among opposing classes around any wedge or polarizing issue.

Twitter and social media seem to be a rich source of opportunities for gathering such data in huge quantities.

Along this line of thought, we consider the possibility of building pre-trained models able to predict social media users' positions on particular subject matters.

Nevertheless, as already pointed out, we hold concerns about the status of the internet and social media in the future.

While machine learning advancements will indeed create new opportunities and better methods for studying users' behaviors, usage of such technologies to produce and share fake human-like data is becoming increasingly harder to spot (Tang, Ruixiang, 2023).

This not only can hugely impact research of social phenomena through social media platforms, but can also contribute to the already increasing polarization of opinions around wedge issues.

Thus, we believe results obtained from analyzing social media data will need further control.

Nonetheless, new discoveries in the field of machine learning will eventually offer new tools and research methodologies for psychologists and social scientists to study and predict behavior on social media. For example, studies that employed Large Language Models (LLMs) have shown that such technologies (such as ChatGPT-4) are equally if not more accurate and unbiased than human annotators in detecting the political affiliation of a post on Twitter (Törnberg, Petters, 2023). Thus, we believe that psychologists and research scientists would highly benefit from the employment of machine learning technologies as research tools in their studies.

4.4 Ethical considerations

In this research we employed data gathered from social media users to predict their follow preferences and therefore make inferences about their position on the polarization spectrum around a topic. Important considerations ought to be made around the implementation and usage of such data to conduct social research.

A first aspect to consider is about the informed consent of using data gathered from social media users. In our research, users were unaware of their data being gathered and analyzed for academic purposes. Nonetheless, we kept sensitive information such as the link between users' ID and related tweets encrypted.

A critical point that could be made against our research comes from the potential harm that could arise from applying our methodology for malicious and manipulative purposes. We've already experienced how this kind of research could be employed for machiavellian and undemocratic purposes in 2018 by witnessing the Cambridge Analytica scandal. Targeting users with specific content to manipulate their view in order to win elections is an unfair and unethical practice, as it exploits users' indecision to majorly increase power of already rich and influential entities, skewing the distribution of power.

It can instead be made an argument for the implementation of such practices to improve decision making in topics that may improve the overall well being of the individuals and the society. Examples of these topics might be related to health policies, climate change actions and education. Such examples fall under the category of nudge and "libertarian paternalism" (Thalers, Sustain, 2009). Moreover, as already stated in the introduction of this article, such methodologies might be implemented to narrow existing gaps between individuals holding different ideologies. This can help avoid phenomena such as radicalization and extreme polarization, which often leads to conflicts and even violent episodes.

However, more often the line between doing good and bad is blurred and not so clear and requires universally shared definitions of such concepts. Nudging is already a quite controversial topic among scholars (Hausman, Welch, 2010) and it's not our job to contribute to this discussion.

Nonetheless, we believe that, regardless of the positive or negative impacts of such methodologies, it can be argued that discovering and sharing new potential ways through which manipulation can be performed on social media can help either prevent

the malicious implementation of them and/or sparkle new ideas on how to exploit them for the good of society. We consider it important to identify ways through which technology can exploit psychological weaknesses to threaten democracy and it is in our best interest to point them out, for the safety of our society.

5. Bibliography

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right. *Psychological Science*, *26*(10), 1531–1542. https://doi.org/10.1177/0956797615594620

Briscoe, F., & Gupta, A. (2016). Social activism in and around organizations. *The Academy of Management Annals*, *10*(1), 671–727. https://doi.org/10.5465/19416520.2016.1153261

Caceres, M. M. F., Sosa, J. P., Lawrence, J. A., Sestacovschi, C., Tidd-Johnson, A., Rasool, M. H. U., Gadamidi, V. K., Ozair, S., Pandav, K., Cuevas-Lou, C., Parrish, M., Rodriguez, I., & Fernández, J. P. (2022). The impact of misinformation on the COVID-19 pandemic. *AIMS Public Health*, *9*(2), 262–277. https://doi.org/10.3934/publichealth.2022018

Chen, K., Yang, S., & Duan, Z. (2021). Twitter as research data. *Politics and the Life Sciences*, *41*(1), 114–130. https://doi.org/10.1017/pls.2021.19

Day, M. V., Fiske, S. T., Downing, E. L., & Trail, T. E. (2014). Shifting liberal and conservative attitudes using Moral Foundations theory. *Personality and Social Psychology Bulletin*, *40*(12), 1559–1573. https://doi.org/10.1177/0146167214551152

DemTech | Bots and Automation over Twitter during the First U.S. Presidential Debate. (n.d.). https://demtech.oii.ox.ac.uk/research/posts/bots-and-automation-over-twitter-during-the-first-u-s-presidential-debate/

Falkenberg, M., Galeazzi, A., Torricelli, M., Di Marco, N., Larosa, F., Sas, M., Mekacher, A., Pearce, W., Zollo, F., Quattrociocchi, W., & Baronchelli, A. (2022). Growing polarization around climate change on social media. *Nature Climate Change*, *12*(12), 1114–1121. https://doi.org/10.1038/s41558-022-01527-x

Feinberg, M., & Willer, R. (2012). The moral roots of environmental attitudes. *Psychological Science*, *24*(1), 56–62. https://doi.org/10.1177/0956797612449177

Feinberg, M., & Willer, R. (2019). Moral reframing: A technique for effective and persuasive communication across political divides. *Social and Personality Psychology Compass*, *13*(12). https://doi.org/10.1111/spc3.12501

Fielding, K. S., & Hornsey, M. J. (2016). A Social Identity Analysis of climate change and environmental attitudes and behaviors: Insights and opportunities. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00121

Finkel, E. J., Bail, C. A., Cikara, M., Ditto, P. H., Iyengar, S., Klar, S., Mason, L., McGrath, M. C., Nyhan, B., Rand, D. G., Skitka, L. J., Tucker, J. A., Van Bavel, J. J., Wang, C. S., & Druckman, J. N. (2020). Political sectarianism in America. *Science*, *370*(6516), 533–536. https://doi.org/10.1126/science.abe1715

Frimer, J. A. (2020). Do liberals and conservatives use different moral languages? Two replications and six extensions of Graham, Haidt, and Nosek's (2009)

moral text analysis. *Journal of Research in Personality*, *84*, 103906. https://doi.org/10.1016/j.jrp.2019.103906

Frimer, J. A., Skitka, L. J., & Motyl, M. (2017). Liberals and conservatives are similarly motivated to avoid exposure to one another's opinions. *Journal of Experimental Social Psychology*, *72*, 1–12. https://doi.org/10.1016/j.jesp.2017.04.003

Gidron, N. (2019). *Toward a comparative research agenda on affective polarization in mass publics*. https://papers.ssrn.com/abstract=3391062

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029–1046. https://doi.org/10.1037/a0015141

Gunther, R., Beck, P., & Nisbet, E. C. (2019). "Fake news" and the defection of 2012 Obama voters in the 2016 presidential election. *Electoral Studies*, *61*, 102030. https://doi.org/10.1016/j.electstud.2019.03.006

Haidt, J. (2012). *The righteous mind: Why Good People Are Divided by Politics and Religion*. National Geographic Books.

Hausman, D. M., & Welch, B. F. (2010). Debate: To nudge or not to nudge*. *Journal of Political Philosophy*, *18*(1), 123–136. https://doi.org/10.1111/j.1467-9760.2009.00351.x

Hong, S., & Kim, S. H. (2016). Political polarization on twitter: Implications for the use of social media in digital governments. *Government Information Quarterly*, *33*(4), 777–782. https://doi.org/10.1016/j.giq.2016.04.007

Hunter, J. D. (1991). *Culture wars: The Struggle To Define America*. [New York] : BasicBooks.

Kahneman, D. (2012). *Thinking, fast and slow*. Penguin UK.

Kubin, E., & Von Sikorski, C. (2021). The role of (social) media in political polarization: a systematic review. *Annals of the International Communication Association*, *45*(3), 188–206. https://doi.org/10.1080/23808985.2021.1976070

Kwak, H., An, J., Jing, E., & Ahn, Y. (2021). FrameAxis: characterizing microframe bias and intensity with word embedding. *PeerJ*, *7*, e644. https://doi.org/10.7717/peerj-cs.644

Milesi, P., & Alberici, A. I. (2016). Pluralistic morality and collective action: The role of moral foundations. *Group Processes & Intergroup Relations*, *21*(2), 235–256. https://doi.org/10.1177/1368430216675707

Milfont, T. L., Davies, C., & Wilson, M. (2019). The moral foundations of environmentalism: care- and Fairness-Based morality interact with political liberalism to predict Pro-Environmental actions. *Social Psychological Bulletin*, *14*(2). https://doi.org/10.32872/spb.v14i2.32633

Mitchell, T. (2021, July 12). *U.S. Public Views on Climate and Energy | Pew Research Center*. Pew Research Center Science & Society. https://www.pewresearch.org/science/2019/11/25/u-s-public-views-on-climate-and-energy/

Mokhberian, N., Abeliuk, A., Cummings, P. J., & Lerman, K. (2020). Moral framing and ideological bias of news. In *Lecture Notes in Computer Science* (pp. 206–219). https://doi.org/10.1007/978-3-030-60975-7_16

Ozola, A. (2015). The Role Of Public Participation And Environmnetal Activism In Environmental Governance In Latvia. *Vide. Tehnoloģija. Resursi*. https://doi.org/10.17770/etr2011vol2.994

Pew Research Center. (2023, March 1). *Terms of use | Pew Research Center*. https://www.pewresearch.org/about/terms-and-conditions/

Ribeiro, M. H. (2019). Auditing Radicalization Pathways on YouTube. *arXiv.org*. https://arxiv.org/abs/1908.08313

Stewart, B. D., & Morris, D. S. (2021). Moving morality beyond the In-Group: liberals and conservatives show differences on Group-Framed moral foundations and these differences mediate the relationships to perceived bias and threat. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.579908

Tang, R., Chuang, Y., & Hu, X. (2023a). The Science of Detecting LLM-Generated Texts. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2303.07205

Tang, R., Chuang, Y., & Hu, X. (2023b). The Science of Detecting LLM-Generated Texts. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2303.07205

Törnberg, P. (2023). ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2304.06588

Tucker, J. A., Guess, A. M., Barberá, P., Vaccari, C., Siegel, A. A., Sanovich, S., Stukal, D., & Nyhan, B. (2018). Social Media, Political polarization, and Political Disinformation: A review of the Scientific literature. *Social Science Research Network*. https://doi.org/10.2139/ssrn.3144139

Van Der Maaten, L. (2008). *Visualizing Data using t-SNE*. https://www.jmlr.org/papers/v9/vandermaaten08a.html

Van Vliet, L. (2021). Moral Expressions in 280 characters or less: An analysis of politician tweets following the 2016 Brexit referendum vote. *Frontiers in Big Data*, *4*. https://doi.org/10.3389/fdata.2021.699653

Wiant, F. M. (2002). Exploiting factional discourse: Wedge issues in contemporary American political campaigns. *The Southern Communication Journal*, *67*(3), 276–289. https://doi.org/10.1080/10417940209373236

Wolsko, C., Ariceaga, H., & Seiden, J. (2016). Red, white, and blue enough to be green: Effects of moral framing on climate change attitudes and conservation behaviors. *Journal of Experimental Social Psychology*, *65*, 7–19. https://doi.org/10.1016/j.jesp.2016.02.005