You are what you browse: A robust framework for uncovering political ideology

Rohit Ram
University of Technology Sydney
Sydney, Australia
rohit.ram@uts.edu.au

Marian-Andrei Rizoiu
University of Technology Sydney & Data61, CSIRO
Sydney, Australia
marian-andrei.rizoiu@uts.edu.au

ABSTRACT
In democratic countries the latent ideology landscape is foundational to individual and collective political action; conversely, fringe ideology drives Ideologically Motivated Violent Extremism (IMVE). Therefore, quantifying ideology is a crucial first step to an ocean of downstream problems, such as; understanding and countering IMVE, detecting and intervening in disinformation campaigns, and broader empirical opinion dynamics modeling. However, online ideology detection faces two significant hindrances. Firstly, the ground truth that forms the basis for ideology detection is often prohibitively labor-intensive for practitioners to collect, requires access to domain experts and is specific to the context of its collection (i.e., time, location, and platform). Secondly, to circumvent this expense researchers generate ground truth via other ideological signals (i.e. hashtags, politicians, etc.), however, the bias this introduces has not been quantified and often this still requires expert intervention. In this work, we present an end-to-end ideology detection pipeline applicable to large-scale datasets. We construct context-agnostic and automatic ideological signals from widely available media slant data; show the derived pipeline is performant, compared to pipelines of common ideology signals and SOTA baselines; employ the pipeline for left-right ideology, and (the more concerning) detection of extreme ideologies; generate psychosocial proxies of the inferred ideological groups; and, generate insights into their morality and preoccupations.

1 INTRODUCTION
Investigating online ideologies is like peering into a hall of mirrors, each reflecting a different facet of society. It’s a kaleidoscope of beliefs and opinions that shape the way we see the world and interact with each other. In this digital age, the internet has become a megaphone that amplifies the voices of these ideologies. They can spread quickly, influencing the thoughts and actions of countless individuals; breeding division, fueling tribalism, entrapping us in echo chambers and perpetuating filter bubbles. A reliable measure of online ideology is the first step in important downstream tasks including opinion dynamics modeling and detecting disinformation campaigns. Tracking online ideology is particularly important in the detection of extreme voices that can spread harmful and false information, which can lead to dangerous and even deadly outcomes. Ideology is canonically (and inexactly) projected onto a left-right spectrum; where the left is associated with equality and reform, and the right is associated with authority and tradition. Recently, the far-right, a prominent archetype of extreme ideologies associated with ultranationalism and opposition to multiculturalism, has risen in popularity internationally. Worryingly, this has led to an increase in ideologically motivated violent events. For the subset of voices involved in Ideologically Motivated Violent Extremism (IMVE), ideology detection can serve as a lead indicator of violent offline behaviors, fortifying individual and collective security. In brief, understanding ideology is essential in understanding how a society thinks and behaves.

Ideological detection is conventionally difficult for several reasons; namely the activity of users is often sparse and their discussions do not directly signal their ideology. In this work we focus on the difficulty of context-dependant ideological activity: the signals of ideology – here dubbed ideological proxies – will often change in time and differ between countries, social media platforms, and subcultures. Fig. 1 illustrates the difficulty further. We observe that only some ideological proxies are consistent across only some contexts (represented by the dashed green boxes). For example, #RoboDebt is not relevant to America, and did not exist before 2016; and, although @MittRomney signaled right-wing ideology in 2012, the right have shifted since Trump’s election. Prior ideology detection techniques fail to easily context-switch and cannot be readily applied to several distinct domains.

Accordingly, three core research questions emerge concerning ideology detection.

The first relates to an automatic and end-to-end architecture for large-scale ideological detection. The primary hindrance to
ideology detection that prevents broad applicability across datasets in prior work is a failure to context-switch. They struggle for three reasons; firstly, they require access to in-depth expert knowledge of the ideological signals within any particular domain. Secondly, they require laborious expert labeling of these ideological proxies (i.e. posts, users, or hashtags). This can vary from extensive and laborious labeling of users [39], to the labeling of hashtags [31], or simply labeling inferred clusters [15, 32]. Note that even in lighter labeling workloads (which are often less robust), access to experts is a non-trivial resource. Finally, they require supplementing the available data with network relations [39] (e.g., follower network) or additional posts [11, 32] (e.g., timeline tweets), which (on platforms like Twitter) is time-consuming, expensive, prone to unstable APIs, and ultimately inhibits broad usage by practitioners for large-scale datasets. Whenever a new dataset (i.e., context) is proposed, these reasons, access to knowledge, laborious labeling, and collection of supplementary data, compound the difficulty.

From Fig. 1 we see that contexts rarely share proxies. Followers of political parties might be largely consistent across time, however, the users cannot be linked across social platforms. Politicians can be linked across platforms, however, elected politicians change with time (and their online presence is not actively tracked). Hashtags may also be relatively consistent across platforms, however, they fall out of fashion. Furthermore, all three aforementioned proxies are significantly disjointed across countries. Alternatively, a media proxy, based on the politically slanted media sources that users share online, is largely consistent across contexts, however, prior literature has failed to exploit this insight.

Ideology detection must be applicable at scale, so that insights are representative. Some prior works apply to only small samples of active users, however, this can lead to potentially biased insights. Additionally, some architectures simply don’t scale to large datasets. For example, neural graph approaches [39] are prohibitively memory intensive for large networks.

Overcoming such issues requires an end-to-end architecture (i.e., no human intervention) applicable at scale. As such our first research question is, can we build a large-scale end-to-end ideology detection pipeline?

The second question relates to the bias and performance of ideological pipelines derived from ideological proxies. Prior research suggests that the choice of proxy can lead to bias in inferred groups [4, 14]. Researchers commonly assume weak ideological proxies as gold-standard ground truths [15, 29, 39], however, the bias associated with such assumptions has not been quantified. Furthermore, it is unclear how the state-of-the-art [15, 32, 39] performs on a balanced gold standard. As such our second research question is, can we determine an ideological proxy that is readily available, generalizes to a variety of contexts, and is performant?

The last question relates to psychosocial characteristics of ideological groups for entire online populations. Understanding the ideological groups at the level of values and beliefs is instrumental in modeling their movements, and potential radicalization. There is rich literature delineating right and left ideologies in terms of moral values [18], and moderates and extremes based on psychosocial language [34]. However, it is unclear whether these findings extend to online populations. Online and offline populations differ demographically [6], but there is a lack of understanding of how they differ psychosocially. Several prior works apply psychosocial analysis to left-right [26, 30] and extremist online populations, however an investigation into the joint ideological spectrum is rare [3]. Other works [3, 30] apply psychosocial analysis to only predefined sets of users, which were limited in size and unrepresentative. As such our last research question is, can we build salient and representative profiles of the ideological groups at scale, and delineate them in terms of psychosocial characteristics?

Our solution rests on two hypotheses; firstly, readily available and rich external sources of ideological knowledge are sufficient for unbiased and performant detection of both left-right and far-right ideologies, and secondly, ties based on homophily (the tendency of similar individuals to associate) allow for the propagation of this ideological knowledge and inference of unlabeled users.

We address the above research questions in two parts; we introduce an end-to-end ideology detection pipeline and we generate psychosocial profiles of the inferred ideological groups.

We address the first research question in Section 3, where we introduce our end-to-end ideology detection consisting of four components; datasets, ideological proxies, homophilic lenses (see Section 3.1), and the inference architecture. In Section 4, we describe and profile five social media datasets, containing millions of users and spanning a variety of social domains. We apply our pipeline at the level of individual users and normalize datasets to this format. In Section 3.2, we infuse ideological knowledge into the pipeline by employing external sources via proxies. We operationalize four left-right and two far-right proxies leveraging behaviors such as emitting politically-charged hashtags, following political parties, endorsing politicians, and sharing media websites. In Section 3.1, we build ties between users, based on homophilic lenses, and propagate our ideological knowledge through these ties via an inference architecture. We generate three homophilic lenses based on: language, endorsements, and topics.

We address the second research question in Section 5, where we evaluate the generalisability and performance of the pipeline constructed through various ideological proxies. We begin by constructing gold-standard benchmarks for left-right and far-right classification. We evaluate the pipeline weakly supervised by each proxy against the benchmarks to determine optimal proxies. We further perform an ablution study of homophilic lenses, to determine the optimal combination. Next, we emphasize the need for an end-to-end pipeline by showing that a model trained in one domain does not generalize well to another. Finally, we compare the performance of the pipeline to state-of-the-art baselines: TIMME [39], UUS [15], and UUS+ [32].

We address the last research question in Section 6 where we characterize the inferred ideological groups through two psychosocial profiling instruments (see Section 2); the Moral Foundations Theory operationalized via FrameAxis and the Grievance Dictionary threat-assessment tool. We critically evaluate the seminal hypotheses of Moral Foundations Theory as they apply to online users at scale. We find an appropriate delineation of left from right ideology and a delineation of moderate from extreme ideologies.

The main contributions are as follows:
• We construct a context-agnostic, end-to-end and large-scale pipeline for performing both left-right and far-right ideology detection.
• We quantify the bias associated with common ideological proxies and show SOTA performance.
• We generate psychosocial profiles of the ideological groups within online populations, and draw insights into these ideologies.

2 PSYCHOSOCIAL MEASUREMENT
Moral Foundations Theory (MFT) [18, 19] aims to explain variations in human moral reasoning through the five modular foundations: care, fairness, loyalty, authority, and sanctity. It further espouses that liberals express individualizing foundations (care and fairness) while conservatives express binding foundations (loyalty, authority, and sanctity) relatively more. This explanation has been empirically verified in several settings, including: religious sermons [18], by political elites [13], and on online social content [26, 30]. Conversely, recent studies refute the strength of these claims [3, 17], and whether it applies to all contexts (e.g., elites vs laypeople) [38].

FrameAxis. MFT is commonly operationalized through dictionary-based approaches, where sets of words are curated for the virtues (e.g., care) and vices (e.g., harm) of each foundation, respectively. These dictionaries have progressively expanded (via MFD [18], MFD2 [16], and eMFD [20]) to better categorize text into the foundations (especially short social media text). Recently, FrameAxis [24] has been applied to moral foundations [26], utilizing clusters of word embeddings generated via the dictionaries. Briefly, this technique generates a vice embedding and a virtue embedding for each foundation: defining a foundation’s axis. The cosine similarities of a user’s word embeddings to each foundation’s axis are used to compute the bias (i.e., average similarity) and intensity (i.e., the variance of similarity with respect to the entire corpus). Intuitively, the bias corresponds to whether a user’s language is closer to the vice or virtue pole, respectively, while intensity corresponds to the relatedness of a user’s language and a foundation.

Grievance Dictionary. The grievance dictionary [34] is a psychosocial dictionary (similar to LIWC and MFD), curated with the purpose of threat assessment, including categories such as fixation, violence, and paranoia. The dictionary is extensively validated, including on social media data, and provides useful features for distinguishing extremist texts.

3 IDEOLOGY DETECTION PIPELINE
In this section, we introduce our ideology detection pipeline in two parts. We begin by broadly delineating the pipeline into four components, and describe their interaction. We then provide details for the operationalization of these components.

There are two guiding desiderata of the pipeline; it is end-to-end (because in practice context-switching is common, and ideology models trained in one context are often narrowly applicable to that context) and it scales to large populations (to form representative profiles of entire online populations).

In brief, we construct an end-to-end scalable pipeline by employing weak supervision signals from other external sources of political information and carefully curated features. The pipeline has four components illustrated in Fig. 2: datasets, ideological proxies, homophilic lenses, and an inference architecture. The first component and the pipeline’s input are datasets, which are conceptualized here as a set of unlabelled users. Each dataset represents a different set of users with a separate social context. The second component, ideological proxies, is used to infuse external political knowledge into the pipeline through weak supervision. We assign labels to a subset of seed users based on particular political behaviors (sharing hashtags, following political parties, endorsing politicians, or sharing media websites). These labels can be in either left-right or far-right flavors. We describe the operationalization of several ideological proxies in Section 3.2. The third component, homophilic lenses, is used to characterize user similarity in ideologically meaningful ways: drawing ties between users. We describe three homophilic lenses in Section 3.1. The last component, the influence architecture, propagates the labels from seed users to the remaining unlabelled users. Here we use LightGBM [21], an efficient tree-based classifier, which effectively assigns the same labels to similar users. We also use FlaML [37], a system that infers hyperparameters based on dataset characteristics. We fix hyperparameter n_estimators to 200 for left-right detection to prevent overfitting, but allow it to be inferred for far-right detection due to the sparsity of far-right users. We further set the is_unbalanced flag due to the disproportionate number of left-leaning users.

The remainder of this section describes the implementation of the homophilic lenses and ideological proxies.

3.1 Homophilic Lenses
Homophily is the tendency of similar users to form ties and consequently be similar in multiple ways. For example, users similar in geography are likely to be politically similar. We assume that if users are similar, through carefully chosen lenses, then they will be similar in ideology. We utilize a classifier that exploits user similarity, measured through homophilic lenses, to infer the ideologies of unlabelled users. This section details three homophilic lenses with known correlations (and intuitive associations) with ideology; lexical lens [12], resharing lens [36], and hashtag lens [8]. We validate these lenses in Section 5, where we show the performance of these ideologically salient lenses in predicting ideology.

Lexical Lense (USE). Language is a strong indicator of one’s community and political ideology [12]; since one’s sociolect and ethnolect are formed through associations with others and language politics often reveal one’s political leaning. Following recent state-of-the-art stance detection approaches [29], we employ the Universal Sentence Encoder (USE) [10] to generate embeddings of users’ posts. USE is a transformer-based model trained for the semantic textual similarity task, which learns expressive representations of sentences. We begin by preprocessing text by removing URLs, hashtags, and mentions, preventing potential data leaks. Next, we concatenate each user’s tweets and encode the result as a 512 dimensional vector via USE. The choice of USE is arbitrary, and other state-of-the-art stance detection pipelines have also utilized large transformer-based language models [32].

Hashtag Lense (HT). Hashtags signal users’ interests through the discussion topics they participate in. Intuitively, users who share similar interests are likely to be politically similar. We construct the
hashtags as the Term-Frequency Inverse Document Frequency (TF-IDF) embedding of users (i.e., documents) via the hashtags (i.e., words) they use. Hashtags that occur at least 10 times are included.

**Resharing Lense (RT).** Resharing is a signal of endorsement. We assume users who endorse the same people likely share the same ideology [36]. We generate a multi-hot encoding for users based on the 1000 most reshared posts. We represent a user \( u_i \) as \( h_i \in \mathbb{R}^{1000} \), where \( h_i[j] = 1 \) if \( u_i \) reshares the \( j \)-th most reshared post \( (h_i[j] = 0 \) otherwise).

### 3.2 Generating Ideological Proxies

An ideological proxy is an external source of political knowledge used to automatically label a sample of users, for later weak supervision of the classifier. There are several types of ideological proxy, requiring varying amounts of labor to generate and correlating with ‘true ideology’ to varying degrees. We construct proxies to generate both left-right and far-right labels via; sharing hashtags, endorsing politicians, following political parties, and sharing media websites.

Proxies assume that specific user behaviors are correlated with ideologies; users who follow a political party likely share the ideology of the party they follow; resharing signals endorsement and users who endorse a politician likely share their ideology; users who emit politically charged hashtags signal their political alignment through them; and, users who spread media publications likely share the slant of the publications they emit. This section operationalizes these proxies, describes the relative advantages between them and describes the gold-standard ground truth, for both left-right and far-right classification. We begin by introducing the four left-right proxies.

**Hashtags** proxy requires an expert to inspect the 1,000 most common hashtags within a dataset and label the political lean of the hashtag: −1 if left-leaning, 0 if non-partisan, and 1 if right-leaning. A user’s political lean is the average of the labeled hashtags they emit, and their ideology label is the sign of this lean.

**Party Followers** proxy requires collecting the followers of the major political parties for a dataset’s corresponding country. The political parties are coded by their ideology and followers, who follow only a single party and are users within the dataset, are assigned this same ideology label.

**Politician Endorsers** proxy requires a dataset of politicians, their political affiliations, and their social media handles, such as the Twitter Parliamentarian Database [35]. Politicians are coded by their party’s ideology (where independents are excluded). Users are labeled via majority vote of the ideologies of the politicians they retweet within a dataset.

**Left-Right Media** proxy requires a dataset of media websites with their political slants. We utilize a survey [27, 28] of news consumption behavior conducted by Reuters, which includes a self-reported political lean by news consumers, to generate media slant scores. We encode the seven-point self-reported political lean onto a numerical scale from −3 to 3. The survey results include reports by Australian, Canadian, United Kingdom, and American participants collected in 2020 and 2021. We compute the media slant, for each year and country, of a publication as the weighted mean political lean of participants who consume the publication, where participant weights are the inverse of the number of publications they consume. Since countries’ perspectives on what constitutes left-leaning and right-leaning differ, we calibrate the scores with the AllSides Media Bias Ratings [5]. We encode AllSides Media Bias Ratings five-point scale onto a numerical scale from −1 to 1. We shift each country’s scores, such that we minimize the sum of squared differences, between a country’s scores and AllSides scores for overlapping publications. Finally, we generate media slant scores, for each publication, as the average slant over all countries and years. We associate media publications (and their slants) with their website domains, averaging where a domain is shared; compute a user’s political lean as the average lean of the media domains they share, and their ideology label as the sign of this lean.

Next, we introduce two far-right proxies based on media sharing. **Far-Right Media** proxy is constructed from the media slant scores of mainstream media as generated above, where a user is labeled far-right if their political lean is greater than 0.5, and moderate otherwise.

**MBFC** proxy is constructed from the Media Bias Fact Check [40] dataset, including both media slant and veracity, and containing conspiratorial and fake news sources. Users who share media from sources classified as ‘right’ are labeled as far-right.

For our purposes, proxies are broadly delineated by: whether they require expert annotation, remain relevant in time, are applicable in many social domains, and how well they represent ‘true ideology’. This delineation describes how well proxies generalize to arbitrary datasets, and ultimately how much effort is required by practitioners. Hashtags, although the canonical approach [31] to generating partisanship labels, requires access to an expert with intimate domain knowledge for labeling, which is expensively time-consuming. Furthermore, hashtags fall out of usage quickly and are specific to the dataset/domain they’re generated from. Party Followers and Politician Endorsers leverage databases of political parties and politicians, and accordingly, require little manual annotation beyond coding political parties. However, politicians will change with elections and the followers of political parties change.
over time. Furthermore, recoding is required for each new country, and politicians do not participate equally in all domains. Media-based proxies (Left-Right Media, Far-Right Media, and MBFC) have a relative advantage in all categories; they leverage readily available sources of media slants such that no manual annotation is required and labeling is completely automatic; media slants are generally consistent through time; and, media sharing behaviors are applicable in virtually all datasets/domains.

The Gold-Standard labels are generated via expert labeling for both left-right and far-right classification. We first generate gold standard left-right labels for the #QandA dataset (see Section 4). For each proxy; we train our pipeline, apply it to the entire dataset, and extract 100 left and 100 right users (for which the classifier is most confident). The combined users are shuffled, and duplicated or unavailable users are removed, resulting in 695 users. These users are labeled, via profile inspection, by an expert as left, right, far-right, or indeterminable. Generating gold-standard labels of extremism is difficult, due to their relative sparsity within datasets. We utilize a list of far-right users generated through a complex manual procedure within the Australian context by Bailo et al. [7]. They started with a 'seed' user; recovered the 'lists' (a Twitter feature for documenting similar users) that this user belonged to; used the intersection of all members of these lists with the users within their dataset; and, had a domain expert determine if these were far-right users. We find 686 (out of 1,496) of these users in #QandA and label them far-right.

## 4 DATASETS

We utilize a variety of datasets, shown in Table 1, that span several domains. These datasets showcase; the relative ease of applying the pipeline, the breadth of domains where the pipeline is applicable, and the scale of users we can characterise. There are three Australian and two American datasets; one comes from Parler, another a mixture of Facebook and Twitter, and the remainder are Twitter-based. In this section, we briefly describe the social context and profile of the activity of users for each dataset.

### #QandA features discussions related to the Australian panel show, Q+A (formerly Q&A); where panelists (public figures, politicians, and experts) answer and discuss curated questions from online viewers and the studio audience. Twitter participation is encouraged and highlighted throughout the airings. The show covers a range of topics relating to Australian politics and issues of concern. #QandA was collected using the filter keyword *qanda* over the entire 2020 interval.

### #Ausvotes features discussions related to the 2022 Australian Federal election; tracking the lead-up to the election and the aftermath. It follows the major parties and their leaders; the Australian Labor Party led by Anthony Albonese, and the Liberal-National Coalition led by Scott Morrison. #Ausvotes was collected by filtering for the keywords *auspol* and *ausvotes*, and for mentions of *@ScottMorrisonMP*, *@AlboMP*, and *@AusElectoralCom*, during the May 9 to June 15 2022 interval (where election day was May 21).

### #Socialsense [9, 23] features discussions related to the Australian Black Summer bushfires, which gathered political discourse concerning climate change. It contains far-right opinions and some misinformation spreading. #Socialsense, as generated in [9], contains 90 days of Twitter and Facebook discussions (from 1 November to 29 January 2020) of bushfires and climate change.

### Riot [22] features discussions related to the January 6th U.S. Capitol Insurrection, including topics of election fraud and insurrection. The dataset spans 6 January to 1 February 2021, and was collected with the filter keywords *TrumpRally*, *Democracy*, *USCapitol*, *Capitol*, *DCProtests*, and *AshliBabbit*.

### Parler [2] features discussions about the U.S. Capitol Insurrection from the Parler (right-leaning) platform. Here we use all posts from 6 January 2021, with no filtering.

The datasets described above represent a diverse set of contexts, with varying levels of politicalness, user activity, community connectedness, and topic distributions.

Fig. 3 shows the distribution of activity for users for each dataset. It shows long-tailed activity distributions and the proportion of low-activity users. Riot shows a significant proportion of low-activity users, who’re often difficult to classify.

## 5 MODEL EXPERIMENTS

### Baseline Comparison. There are a plethora of approaches to ideology (a.k.a. stance) detection, utilizing different features, architectures, and supervision setups. In the pursuit of performance, these approaches enforce limiting assumptions. Firstly, they require specific data-generating processes, such as utilizing discussions surrounding a controversial topic (e.g., gun control) or datasets where users have clear partisanship (i.e., politicians). Secondly, they require supplemental data such as the followership network or users’
prior posting history, which are often prohibitive to acquire. Lasty, they require some degree of human intervention to garner labels, whether that is labeling users, proxies, or inferred clusters. It is unclear if this performance trade-off for lack of flexibility is required. Here we evaluate, for comparison, three state-of-the-art stance detection techniques; UUS [15], UUS+ [32], and TIMME [39]. We utilise the left-right gold-standard for evaluation. UUS is an unsupervised procedure that encodes the $k$ most active users, applies dimensionality reduction, and finally clusters these embeddings. The clusters are assigned to stances via inspection by an expert. Darwish et al. [15] perform hyperparameter tuning of $k$ active users, features (based on retweets, retweeted accounts, and hashtags), dimensionality reduction schemes, and clustering schemes; recommending the use of a reweeted account encoding of the 1000 most active users, with UMAP [25] and Mean-Shift. In practice, the recommended setup does not lead to delineated clusters. We apply every combination of features, encoded by UMAP and clustered by Mean-Shift, reporting the average for each metric. In addition, UUS only reports labels for the most active users, however, the gold-standard users are not the most active. We utilize the inference methods of both UMAP and Mean-Shift to acquire labels for these users. UUS+ is an extension that finetunes a BERT model using the labels of active users inferred by UUS and the text of these users’ posts. The model can then be applied to all users to infer stances. We fine-tune for 10 epochs, which takes some minutes. We report the average for each metric over every UUS TIMME is a supervised multi-task multi-relation neural graph technique employing the followership, retweet, reply, mention, and quote networks, to embed and classify users. We use all relations except the followership network, as it is prohibitive to acquire and this version provides a fair comparison. Training this model has significant GPU memory requirements.

Figure 4: The ideology detection ROC-AUC performance when trained on one proxy (y-axis) and tested on another (x-axis) for (a) far-right proxies and (b) left-right proxies, and (c) trained on one dataset (y-axis) and tested on another (x-axis). In (a) and (b) the dataset is fixed to QandA, while in (c) the proxy is fixed to Left-Right Media.

|        | UUS     | UUS+    | TIMME   | Ours    |
|--------|---------|---------|---------|---------|
| F1-Macro | 0.6038 ± 0.2324 | 0.6123 ± 0.2605 | 0.8819 | 0.9187 |
| ROC-AUC | -       | 0.7587 ± 0.1524 | 0.8889 | 0.9526 |

Section 5 shows the F1-macro scores for each technique for the gold-standard labels. We see that our approach outperforms all others except UUS+. We note that UUS, and therefore UUS+, are not robust on our datasets, with most implementations resulting in no clustering. Furthermore, they require expert intervention to label the result clusters. Conversely, our approach is robust: producing proxy consistent and gold-standard consistent inferred labels. Accordingly, our approach can be reliably applied across a variety of contexts with no expert intervention and performs at a state-of-the-art level.

**Model Performance.** Several studies compare the relative importance of textual, relational, and topical features [1, 15]; however, the conclusions are often conflicting. Darwish et al. [15] report best performances with retweeted account features; Aldayel and Magdy [1] conclude that a combination of network and textual features is best. Architecture is likely a contributing factor when considering the performance of these classifiers. Table 2 shows area-under-the-receiver-operating-curve (ROC-AUC) scores for the pipeline trained with each proxy (columns), utilizing each combination of the lexical, endorsement, and topical lenses (rows). The gold standard is evenly split into validation and test, with the test metric reported. A higher ROC-AUC score is better with a maximum score of 1 and a random baseline of 0.5. The most performant pipeline for left-right detection achieves a ROC-AUC of 0.95 using the lexical and retweet lens with the Left-Right Media proxy. For far-right detection the pipeline utilizing the MBFC and all lenses achieves a ROC-AUC of 0.78. In both setups, no lens combination consistently dominates the others, however, we observe that in general pipelines including the lexical lens outperform their peers (with a few exceptions). This supports the literature that points to the robust superiority of contextual text embeddings [32]. We recommend and hereafter employ the pipeline utilizing the lexical lens, and the Left-Right Media and MBFC proxies, for ideology detection. In addition to being performant, these proxies require no expert intervention and make the pipeline fully automated.

**Cross Proxy Generalization and Validation.** The political inclination of online users is not readily observable; instead studies must infer ideological labels, either through an expert or via a proxy. However, studies [4, 14] have shown that the choice of proxy often
Cross Dataset Generalization. One might expect that political language from one social context would transfer to another. For example, we might expect that ideological groups from the same country or social media platform, use similar language. We assume that each dataset represents a distinct social context. Fig. 4c shows the 5-fold cross-validation ROC-AUC performance of left-right classification models, trained and tested on each combination of dataset, respectively, utilizing Left-Right Media as the proxy. Foreseeably, models perform best when trained and tested on the same dataset; however, this extends to broader contexts. Models trained in the Australian context perform relatively well only when tested within the Australian context (with a noticeable underperformance when transferred poorly to implicitly political contexts. These observations confirm that the signals of political ideology differ within social contexts. These observations confirm that the signals of political ideology differ within social contexts. We count these hypotheses for both the bias and intensity, and for both right and far-right (i.e., each cell contains four hypotheses). This effectively delineates the population into users of vice and virtue. In Fig. 5a, ideological classification techniques that work well within one context might not transfer well to others. For example, explicit political contexts might render one performance while transferring poorly to implicitly political contexts.

### Table 2: Ablation of Features & Proxies
The table shows the ROC-AUC for each combination of features, trained on each proxy, followed by validation and testing on an even-split the gold standard, respectively.

| Features       | Hashtags | Left-Right Media | Politician Endorsers | Party Followers | Far-Right Media | MBFC  |
|----------------|----------|------------------|----------------------|-----------------|----------------|-------|
| use            | 0.880952 | 0.946363         | 0.788069             | 0.867870        | 0.690548       | 0.773385 |
| ht             | 0.872580 | 0.848509         | 0.812140             | 0.875720        | 0.559088       | 0.632652 |
| rt             | 0.840136 | 0.879383         | 0.751701             | 0.844061        | 0.538484       | 0.668365 |
| use+ht         | 0.948980 | 0.939037         | 0.869963             | 0.878598        | 0.714673       | 0.784933 |
| use+rt         | 0.879906 | **0.952643**     | 0.784929             | 0.820774        | 0.665831       | 0.762458 |
| ht+rt          | 0.903977 | 0.937467         | 0.798796             | 0.914443        | 0.570337       | 0.632258 |
| use+ht+rt      | 0.949503 | 0.929095         | 0.853741             | 0.874935        | 0.713392       | 0.784844 |

### Table 3: Hypotheses Win-Loss
The table counts the number of times the hypotheses (that the left use individualising foundation and the right use binding foundations) are shown to be significant, for a Wilcoxon Rank Sign Test (95%) with Holm adjustment for family-wise error, in each dataset. We count these hypotheses for both the bias and intensity, and for both right and far-right (i.e., each cell contains four hypotheses).

| Foundation     | Qanda | Ausvotes | Social sense | Riot | Parler | Total |
|----------------|-------|----------|-------------|------|--------|-------|
| Fairness       | 2     | 2        | 2           | 2    | 2      | 10/20 |
| Care           | 2     | 4        | 3           | 1    | 3      | 13/20 |
| Loyalty        | 2     | 0        | 1           | 1    | 2      | 6/20  |
| Authority      | 2     | 1        | 2           | 2    | 2      | 9/20  |
| Sanctity       | 2     | 0        | 1           | 2    | 3      | 8/20  |

Total          | 10/20 | 7/20     | 9/20        | 8/20 | 12/20  | 46/100|

### 6 PSYCHOSOCIAL EXPERIMENTS

#### Ideological stability of Moral Foundations
A core allure of moral foundations theory is its explanation of the division of political ideologies via moral language [18]. Namely, it posits liberals favor individualising foundations and conservatives favor binding foundations. Despite its support in psychological survey data [18], and a handful of online studies [26, 30], it is not clear how consistently this explanation is supported by online social data [3, 38]. Table 3 counts support for each moral foundations hypothesis (rows) in each dataset (columns). We count support for both left/right and left/far-right hypotheses, and both bias and intensity measures. The hypotheses are supported only 44% of the time, with support marginally favoring the individualizing over the binding hypotheses. This suggests that the moral foundations theory’s explanation of ideological division should not be applied to online social data.

#### Right-wing commonalities
The moral foundations theory’s hypotheses are not consistently observed throughout our datasets (as we discern above), nor in prior literature investigating online moral language [33, 38]. Still, a moral language explanation of ideological division is desirable. For each moral foundation, we assign a user virtue/vice score equal to their intensity, if their bias is positive/negative, respectively, otherwise it is 0. This effectively delineates the population into users of vice and virtue. In Fig. 5a,
Figure 5: In (a) and (b) the differences in the distribution of psychosocial properties with respect to ideological groups are shown for #QandA and #Ausvotes, respectively; where ideological groups are represented by color, the y-axis shows psychosocial categories. (a) Vices-Virtues. The mean difference of each group ideological group from neutral for Moral Foundations vice and virtue categories. (b) Grievances. The signed-KL divergence of each group ideological group from neutral for grievance categories.

we plot each foundation’s mean vice and virtue scores for each ideological group in the #Socialsense dataset. We observe that a larger proportion of right-wing users partake in the language of vice than of virtue, compared to left-wing users. In the supplementary material appendix, we show that this is relatively consistent across all datasets. Notably, this is also observable in prior literature [33]. This provides a more consistent signal for delineating the left from right in the online social context.

Differentiating extremist from moderates. For various stakeholders, identifying extremist ideological groups is often of greater concern given their propensity to violence; however, moral language doesn’t sufficiently distinguish extremes from moderates. We utilize the Grievance dictionary [34] to quantify users’ language usage with respect to threat assessment categories. In Fig. 5b, we plot the Kuller-Leibach divergence (signed by mean difference) between the distribution of each ideological group from the neutral group for each threat assessment category in the #Ausvotes dataset. We observe that the far-right differs significantly from the other ideological groups in all categories, and generally uses more grievance language. Notably, in the #Ausvotes dataset, the far-right use honour and god type language less than other groups. There are two implications of these observations; first, grievance language signals far-right ideology; and second, there is an overlap between the language of the far-right and a threat assessment indicator (strengthening the case for monitoring this group for public safety).

7 CONCLUSION

Ethical Considerations. We note there are some privacy concerns for inferring political affiliations, and tools like ours should only be used by researchers to understand population characteristics. Where public safety outweighs other ethical concerns, tools like ours should only be used for early warning systems and further investigation into users is required. It is important to note that we only use expert inferred political affiliation, and not private self-reported political indicator data, as our gold standard. The tools we develop here are simply more efficient inference mechanisms.

Conclusions. In this work, we propose a robust ideological detection pipeline, which addresses several fundamental problems. Assuming that users’ media-sharing behaviors betray their political ideology, we present a robust ideological proxy that is context-independent and requires no manual labeling. We compare the media proxy to other common choices, and how our pipeline outperforms state-of-the-art approaches. Finally, we present psychosocial profiles of entire online populations and generate unique insights into the values of online ideological groups. Such research provides a reliable understanding of our online political landscape and is integral to downstream tasks.

REFERENCES

[1] Abeer Aldayel and Walid Magdy. 2019. Your stance is exposed! analysing possible factors for stance detection on social media. Proceedings of the ACM on Human-Computer Interaction 3, CSCW (2019), 1–20.

[2] Max Alipapoulos, Emmi Bevensee, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, and Savvas Zannettou. 2021. A Large Open Dataset from the Parler Social Network. https://doi.org/10.5281/zenodo.4442460

[3] Meryam Alizadeh, Ingmar Weber, Claudio Cioffi-Revilla, Santo Fortunato, and Michael Macy. 2019. Psychology and morality of political extremists: evidence from Twitter language analysis of alt-right and Antifa. EPJ Data Science 8, 1 (2019), 1–35.

[4] Kenan Alkiew, Bohan Zhang, and David Jurgena. 2022. Classification without (Proper) Representation: Political Heterogeneity in Social Media and Its Implications for Classification and Behavioral Analysis. In Findings of the Association for
Computational Linguistics. ACL 2022. 504–522.

[5] AllSides. 2022. AllSides Media Bias Ratings. https://www.allsides.com/media-bias-ratings.

[6] Brooke Auxier and Monica Anderson. 2021. Social media use in 2021. Pew Research Center 1 (2021), 1–4.

[7] Francesco Bailo, Amelia Johns, and Marian-Andrei Rizoiu. 2022. Ruling Information Crises: The Australian Far-right community in the Twitterosphere. Under Review.

[8] Leticia Bode, Alexander Hanna, Ben Sayre, JungHwan Yang, and Dhavan V Shah. 2013. Mapping the political Twitterverse: Finding connections between political elites. (2013).

[9] Pio Calderon, Rohat Ram, and Marian-Andrei Rizoiu. 2022. Opinion Market Model: Stemming Far-Right Opinion Spread using Positive Interventions. arXiv preprint arXiv:2208.06620 (2022).

[10] Daniel Cey, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limitiaco, Rohmi St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. arXiv preprint arXiv:1803.11175 (2018).

[11] Souvik Chakraborty, Pawan Goyal, and Animesh Mukherjee. 2021. Fast Few Shot Self-attentive Semi-supervised Political Inclination Prediction. In International Conference on Asian Digital Libraries. Springer, 3–20.

[12] Aleksandra Cichońcka, Michał Bilewicz, John T Jost, Natalia Marroch, and Marta Witkowski. 2016. On the grammar of politics—or why conservatives prefer nouns. Political Psychology 37, 6 (2016), 786–815.

[13] Scott Clifford and Jennifer Jerit. 2013. How words do the work of politics: Moral foundations theory and the debate over stem cell research. The Journal of Politics 75, 3 (2013), 659–671.

[14] Raviv Cohen and Derek Buth. 2013. Classifying political orientation on Twitter: It’s not easy! In Proceedings of the International AAAI Conference on Web and Social Media, Vol. 7. 91–99.

[15] Kareem Darwish, Peter Stefanov, Michael Anpetit, and Preslav Nakov. 2020. Unsupervised user stance detection on twitter. In Proceedings of the International AAAI Conference on Web and Social Media, Vol. 14. 141–152.

[16] J Frimer, JHaidt, J Graham, M Dehghani, and R Boghrati. 2017. Moral foundations theory and the debate over stem cell research. The Journal of Politics 75, 3 (2013), 659–671.

[17] Jerome A Frimer. 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 (2020), 103906.

[18] Jesse Graham, Jonathan Haidt, and Brain A Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. Journal of personality and social psychology 96, 5 (2009), 1029.

[19] Jonathan Haidt and Jesse Graham. 2007. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. Social Justice Research 20, 1 (2007), 98–116.

[20] Frederic R Hopp and René Weber. 2021. Reflections on extracting moral foundations from media content. Communication Monographs 88, 3 (2021), 371–379.

[21] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems 30 (2017).

[22] Daniel Kerchner and Laura Wrubel. 2021. U.S. Capitol Riot and #TrumpRally Tweet IDs. https://doi.org/10.7910/DVN/I5WC7O

[23] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems 30 (2017).

[24] Daniel Kerchner and Laura Wrubel. 2021. U.S. Capitol Riot and #TrumpRally Tweet IDs. https://doi.org/10.7910/DVN/I5WC7O

[25] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems 30 (2017).

[26] Negar Mokhberian, Andrés Abeluk, Patrick Cummings, and Kristina Lerman. 2020. Moral framing and ideological bias of news. In International Conference on Social Informatics. Springer, 206–219.

[27] Nie Newman, Richard Fletcher, Anne Schulz, Simone Andi, Craig T Robertson, and Rasmus Kleis Nielsen. 2021. Reuters Institute digital news report 2021. Reuters Institute for the Study of Journalism.

[28] Sora Park, Caroline Fisher, Kieran McGuinness, Jee Young Lee, and Kerry McCal- lum. 2021. Digital news report: Australia. News and Media Research Centre (UC).

[29] Ammar Rashed, Muzahid Kutlu, Kareem Darwish, Tamer Elsayed, and Canun Bayrak. 2021. Embeddings-Based Clustering for Target Specific Stances: The Case of a Polarized Turkey. In Proceedings of the International AAAI Conference on Web and Social Media, Vol. 15. 537–545.

[30] Markus Reiter-Haas, Simone Kopinek, and Elisabeth Lex. 2021. Studying Moral-based Differences in the Framing of Political Tweets. In ICWSM. 1085–1089.

[31] Marian-Andrei Rizoiu, Timothy Graham, Rui Zhang, Tifei Zhang, Robert Ackland, and Lexing Xie. 2018. # DebateNight: The Role and Influence of Socialbots on Twitter During the 1st 2016 US Presidential Debate. In ICWSM.