Proceedings of the Nineteenth International AAAI Conference on Web and Social Media (ICWSM 2025)

Practical Guidelines for Ideology Detection Pipelines and Psychosocial Applications

Rohit Ram[1,2], **Emma Thomas**[3], **David Kernot**[4], **Marian-Andrei Rizoiu**[2]
[1] Thaum
[2] University of Technology Sydney
[3] Flinders University
[4] Defence Science and Technology Group
rohit@thaum.io,marian-andrei.rizoiu@uts.edu.au,
emma.thomas@flinders.edu.au,david.kernot@defence.gov.au

Abstract

Online ideology detection is crucial for downstream tasks, like countering ideologically motivated violent extremism and modeling opinion dynamics. However, two significant issues arise in practitioners' deployment. Firstly, gold-standard training data is prohibitively labor-intensive to collect and has limited reusability beyond its collection context (i.e., time, location, and platform). Secondly, to circumvent expense, researchers employ ideological signals (such as hashtags shared). Unfortunately, these signals' annotation requirements and context transferability are largely unknown, and the bias they induce remains unquantified. This study provides guidelines for practitioners requiring real-time detection of left, right, and extreme ideologies in large-scale online settings. We propose a framework for pipeline constructions, describing ideology signals by their associated labor and context transferability. We evaluate many constructions, quantifying the bias associated with signals and describing a pipeline that outperforms state-of-the-art methods (0.95 AUC ROC). We showcase the capabilities of our pipeline on five datasets containing more than 1.12 million users. We set out to investigate whether the findings in the psychosocial literature, developed for the offline environment, apply to the online setting. We evaluate at scale several psychosocial hypotheses that delineate ideologies concerning morality, grievance, nationalism, and dichotomous thinking. We find that right-wing ideologies use more vice-moral language, have more grievance-filled language, exhibit increased black-and-white thinking patterns, and have a greater association with national flags. This research empowers practitioners with guidelines for ideology detection, and case studies for its application, fostering a safer and better understood digital landscape.

Code — github.com/behavioral-ds/ideology_prediction

1 Introduction

Ideologies are the collection of beliefs and opinions about the ideal arrangement of society (Cohrs 2012). Tracking extreme ideologies is particularly important in detecting extreme voices that can spread harmful and false information, leading 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. There has been a recent increase in fringe and extreme-leaning worldviews, including the far-right – a prominent archetype of extreme ideologies associated with ultranationalism and opposition to multiculturalism. Worryingly, this has increased Ideologically Motivated Violent Extremism (IMVE) (Carr et al. 2022) – a term coined to encompass religious, political and nationalist extremism. Ideology detection is a lead indicator for IMVE, fortifying individual and collective security. It facilitates understanding these ideological groups' values and beliefs, which helps design interventions, build political bridges and tackle radicalization.

Radicalization can occur in a matter of weeks (McCauley and Moskalenko 2008; Booth et al. 2024), both offline (face-to-face) and online (forums and social media platforms). To combat this, practitioners – such as law enforcement and national security agencies – need practical, real-time ideology detection tools that minimize human effort and can be applied across diverse contexts. Despite the significant existing literature, practical and effective detection guidelines remain scarce. This study establishes a framework for ideology detection pipelines, examining diverse constructions and demonstrating practical implementations using off-the-shelf components. Our first aim is to identify practical pipelines that reduce annotation efforts while maintaining transferability across different contexts. Our second aim is to validate insights into the psychosocial asymmetries of ideologies. We leverage five large datasets, totaling 1.12 million profiles, and test several hypotheses from the psychosocial literature at scale, mainly developed in offline laboratory setups. We answer two specific research questions.

The first question involves *ideological proxies* – measurable user behavior signals correlating with "true" ideology – that minimize annotation labor and are transferable across contexts. We define a *context* as the tuple (topic, time, geography, platform). Prior works rely on various sources of ideological knowledge, including manually labeling users, labeling ideological proxies, and detecting group behavior differences. However, these approaches have limitations: the former two require extensive expert labeling – an expensive resource – and often fail to transfer across contexts. The latter often lacks robustness. Of the three, ideological proxies are the most common approach to reducing labor; however, they vary in reusability. See Section 2 for a complete discussion.

Furthermore, few users partake in direct ideological activity, and some actively avoid disclosure. Consequently, many proxies reveal only the vocal subset of users, biasing downstream analyses (Alkiek, Zhang, and Jurgens 2022; Cohen and Ruths 2013). Despite this, prior works commonly use proxies as ground truth (Darwish et al. 2020; Rashed et al. 2021; Xiao et al. 2020) without quantifying the bias this entices. Our first research question is: **Which ideological proxies minimize annotation labor, maximize context transferability, and reduce bias?**

The second question involves the psychosocial asymmetries of ideologies. Understanding the values and beliefs of ideological groups is instrumental in modeling their polarization and user radicalization. Ideological asymmetry studies are abundant in relevant disciplines (Tomkins 1963; Jost 2017); often shown via offline surveys. For example, moral values delineate left-from-right ideologies (Graham, Haidt, and Nosek 2009); and grievance/grudge language delineate moderate-from-extreme ideologies (Stankov 2021; Van der Vegt et al. 2021). Many of these hypotheses were developed with offline populations, and there is limited evidence for online populations. We know online and offline populations differ demographically (Auxier and Anderson 2021), but we do not understand their psychosocial differences. We ask **can we build psychosocial profiles of ideological groups and employ them to evaluate hypotheses related to the psychosocial traits of these groups?**

Solution Outline. First, we evaluate ideological proxies. We make the widely adopted assumption that homophily – the tendency of similar individuals to associate – propagates ideology. We build a framework to construct ideology detection pipelines. We qualitatively evaluate proxies, by their minimization of labelling efforts and how readily they transfer to new contexts, and quantitatively evaluate proxies on their prediction of human-annotated ground truth. We show that a pipeline constructed through a proxy based on media consumption and a lens based on text, is both qualitatively advantageous and quantitatively performant. Second, we use a pipeline to test hypotheses of the psychosocial asymmetries of ideology at scale.

We address the first question in Section 4. We introduce our pipeline framework, consisting of four components: dataset, ideological proxy, homophilic lens, and inference architecture. We use five *social media datasets*, collected from three platforms (Parler, Facebook and Twitter), containing 1.12 million users, and spanning social domains such as TV shows, elections, climate change, antivaccination and the January 6th US Capitol Insurrection. We frame the problem as user classification: the left-right detection as ternary classification (left, right, and neutral), and the far-right detection as binary classification. We limit our scope to Anglo-centric, English-speaking contexts with a dominant uniaxial political spectrum[1]. We explore four left-right and two far-right *ideology proxies*, leveraging behaviors such as posting politically-charged hashtags, following political parties, endorsing politicians, and sharing media websites. We build three *homophilic lenses* based on language, endorsements, and topics. We use the ideology proxies and homophily lenses to build ideology pipelines with an *off-the-shelf classifier*. See Fig. 1.

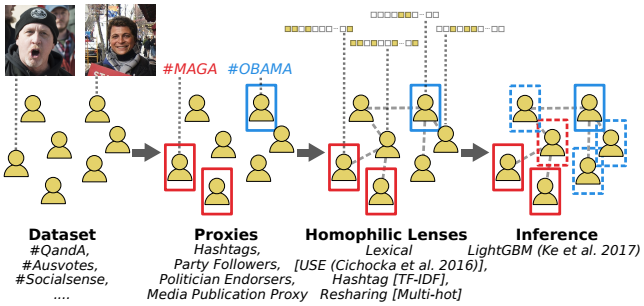

Figure 1: **The schema conceptualizes the four components of the pipeline**; (1) the datasets contain information about users (two examples are shown), (2) the ideological proxies assign labels on some of the users based on external information (here #MAGA indicates right-leaning, while #OBAMA indicates left-leaning), (3) the homophilic lenses build numeric descriptions for user and a way to measure their similarity, and (4) inference architecture predicts the likely labels of all other users in the dataset.

In Section 6 we evaluate the performance of ideology detection pipelines. We construct gold-standard benchmarks for left-right and far-right classification via human annotation and use them to evaluate bias introduced by ideological proxies. Furthermore, we assess various combinations of ideology proxy and homophilic lens to observe interaction effects and find the best performing combination. Finally, we compare this pipeline to state-of-the-art baselines: TIMME (Xiao et al. 2020), UUS (Darwish et al. 2020), and UUS+ (Samih and Darwish 2021) and achieve the best area-under-the-receiver-operating-curve (AUC ROC) of 0.95, an improvement of 6.7% over the next best, TIMME.

We address the second question in Section 7. We evaluate psychosocial hypotheses relating to morality, grievance, nationalism, and dichotomous thinking. For *morality*, we evaluate the seminal Moral Foundations Theory (Graham, Haidt, and Nosek 2009) hypotheses, operationalized via FrameAxis (Kwak et al. 2021) (see Section 3). In its two subsets of hypotheses, individualizing and binding; we find relatively more support for the prior. However, only 46% of hypotheses are supported overall. As alternative hypotheses, we find that the right uses the language of vice more than the left, with statistical significance. For *grievance*, following literature that theorized that grudges and grievances are requirements for radicalization (Stankov 2021), we find large-scale proof that the far-right uses grievance language more than moderates. We operationalize via the Grievance Dictionary threat-assessment tool (Van der Vegt et al. 2021) (see Section 3). For *nationalism*, we show that the right exhibit nationalism via flag emojis, adding validity to our inferred grouping. Finally, for *dichotomous thinking*, we apply a dictionary-based approach, showing that the far-right, followed by the right,

[1]This is not to discount the need for ideology detection in other regions, like the Global South. Nor to suggest that a uniaxial spectrum is sufficient to encompass the complex politics of global communities. See Section 8 for a discussion.

exhibits more black-and-white thinking (supporting prior work).

The main contributions of this work are as follows:

- An ideology detection pipeline applicable in large-scale online setups, that minimizes labor requirements and improves transferability to multiple contexts.
- The most comprehensive discussion and analysis of ideological proxies (to our knowledge); quantifying their bias independently and jointly with homophilic lenses. One construction outperforms state-of-the-art methods.
- Evaluation of psychosocial hypotheses concerning ideologies in a large-scale online setting.

Glossary. For readability, we collocate and define terms here. *Ideological Proxy*: measurable user behaviors correlating with ideology (e.g., emitting hashtags, following ideological users, sharing ideological media, etc.). *Homophilic Lens*: a representation of users highlighting specific behaviors under the homophilic assumption (users who act similarly are likely to share similar ideological beliefs). *Inference Architecture*: a classifier used to infer user connections in a latent space.

2 Related Work

Two corpora relate to our study; ideology detection and psychosocial asymmetries. Our primary concern, for the prior, is pipeline delineation criteria and, for the latter, is evidence bases for hypotheses.

2.1 Ideology Detection Delineation

Ideology detection is becoming popular and relevant for researchers and practitioners across the computer, social, and political sciences. We delineate prior work by population scope, homophilic lenses, and ideological proxies.

Population Scope describes *who the technique applies to?* Many works limit their scope to a population subset: legislators, elites (Xiao et al. 2020), the politically active (Darwish et al. 2020), or everyone (Samih and Darwish 2021). Subsets offer clearer ground truth and easier inference, but lack representativeness of the population, leading to biases when applied broadly (Alkiek, Zhang, and Jurgens 2022; Cohen and Ruths 2013) and constraining the representativeness of correlational analyses (Alizadeh et al. 2019). This work applies to all users, providing representative downstream analysis.

Homophilic Lenses describe *which features are utilized to infer ideology?* Underlying detection is the homophilic assumption – people who act similarly are likely to share similar ideological beliefs. Prior works operationalize this via several lenses: content (including metadata, images (Xi et al. 2020), and text (Preoţiuc-Pietro et al. 2017)), network (such as followership and resharing (Xiao et al. 2020)), or a combination (Chakraborty, Goyal, and Mukherjee 2022). In political science, the modus operandi is Ideal Point Estimation (Poole and Rosenthal 1985) using homophily via legislator voting behavior. Ideal Point Estimation techniques are largely unsupervised and rely on distinct behavioral patterns but are used in most political science ideology measurement work (Gu et al. 2016; O'Hagan and Schein 2023). In particular, Barberá (2015) utilize the *following of politicians* on Twitter to estimate user ideal points, and their work is employed in correlation analysis (Badaan et al. 2023). Given the host of behaviors that portray ideology, novel lenses continue to emerge, including media sharing (Cann, Weaver, and Williams 2021; Eady et al. 2020), and community participation (Ravi, Vela, and Ewetz 2022). Prior works commonly engineer salient lenses and seek their optimal combination (Darwish et al. 2020; Aldayel and Magdy 2019); however the complexity of data context, inference architecture, and ideological proxy choices often make the conclusions unclear. For example, Darwish et al. (2020) recommend a retweet lens, while Aldayel and Magdy (2019) recommend a network and lexical lens combination. The ideological salience of lenses and their combinations is not our work's focus. We implement three homophilic lenses previously shown to be ideologically salient, to limit interaction effects with ideological proxies concerns.

Ideological Proxy describes *what is the source of ideological knowledge?* Prior work utilizes three paradigms for detection: supervised, unsupervised, and weak supervision. Each employs distinct ideological knowledge sources– dubbed *ideological proxies*. In this study, we focus on both the proxies' performance and their expert annotation labor requirement when used across multiple contexts. We delineate proxies by (1) the extent to which they require expert annotation, (2) are transferable to different contexts, and (3) how well they represent *true* ideology. These criteria describe how well proxies generalize to arbitrary datasets and how much manual effort is required for switching contexts.

Direct user annotation for supervised learning (Thomas et al. 2022; Xiao et al. 2020) is simple, the most representative, and accommodates fine-grained distinctions between ideologies (Liu et al. 2023); however, it is also the most restrictive, requiring laborious expert evaluation of users, across every new context. Conversely, unsupervised approaches need little annotation and, in theory, are applicable in any context. Some apply embedding and clustering techniques (Darwish et al. 2020; Samih and Darwish 2021; Rashed et al. 2021). Others utilize matrix factorization to jointly learn representations of users and their behaviors (Lai et al. 2022; Lahoti, Garimella, and Gionis 2018). These methods are not robust in practice, require highly polarized contexts, fail on homogenous user sets, and depend heavily on lenses. Furthermore, they require expert knowledge in post-analysis (e.g., identifying clusters) (Darwish et al. 2020), and clusters do not always align with ideology. Weak supervision trades-off between the high labor of supervised and the instability of unsupervised methods. It employs an ideological proxy, a user behavior strongly correlated with ideology. Prior work utilizes a range of ideological proxies, including; politically-charged hashtags (Rizoiu et al. 2018), political party relationships (Eady et al. 2020), politician relationships, community participation (Lai et al. 2022), and news media sharing (Jiang, Ren, and Ferrara 2023; Badawy, Lerman, and Ferrara 2019; Bailo, Johns, and Rizoiu 2023). We assess proxies' labor minimization and context transferability qualitatively in Section 4.2 and assess their representativity quantitatively in Section 6.2.

Related Ideology Detection. Jiang, Ren, and Ferrara (2023) use text and retweet features, and a combined media-hashtag

proxy which they validate. However, they limit scope to active users who retweet and require hashtag proxy labeling.

2.2 Psychosocial Profiling of Ideological Groups.

Many social science works detail the nuanced profiles of fine-grained ideological groups and highlight the asymmetries between ideologies (Tomkins 1963; Jost 2017; Rao 2017; Rao, Morstatter, and Lerman 2022), often requiring painstaking surveys and ethnographic inquiry. In this work, we supply large-scale online evidence for hypotheses surrounding psychosocial asymmetries of ideologies, relating to morality, grievance, nationalism, and dichotomous thinking.

Morality. Moral Foundations Theory (Graham, Haidt, and Nosek 2009) is an explanation of moral values variations between liberals and conservatives (see Section 3). Despite its support in psychological survey data (Graham, Haidt, and Nosek 2009), and a handful of online studies (Reiter-Haas, Kopeinik, and Lex 2021; Mokhberian et al. 2020), online social data inconsistently supports this explanation (Wang and Inbar 2021; Alizadeh et al. 2019).

Grievance and Grudge are linked to extreme ideologies in psychological theory;Van der Vegt et al. (2021) link grievance to extremism, and Stankov (2021) link grudge to the far-right.

Nationalism is definitionally associated with right-wing politicians. Prior work has shown that flags are associated with nationalism (Kemmelmeier and Winter 2008), emojis hold identity and semantics information (Li et al. 2020), and that flag emojis are significant in right-leaning political communication (Kariryaa et al. 2022). However, this research is limited to politicians in a US context.

Dichotomous Thinking is a cognitive distortion in people with internalizing disorders, is tied to language (Bathina et al. 2021), and is associated with the right (Meyer 2020).

Our concern is evaluating hypotheses in large-scale online populations in various contexts. Accordingly, we limit our scope to automated techniques using online metadata alone. Prior work, online, analyses left-right (Reiter-Haas, Kopeinik, and Lex 2021) or extremist asymmetries, but rarely both (Alizadeh et al. 2019). Additionally, they analyze small and non-representative samples. This work analyzes left, right, and far-right ideologies in several large-scale online contexts.

3 Preliminaries

Our study relies on several techniques from prior work.

Encoding Techniques are employed to implement homophilic lenses; the Universal Sentence Encoder (USE) (Cer et al. 2018) for our lexical lens (a mature, off-the-shelf, transformer-based model), Term-Frequency Inverse Document Frequency (TF-IDF) for our hashtag lens, and a multi-hot encoding for our resharing lens. We utilize simple encoding techniques, as they are not our work's main focus.

Inference Architecture Implementation. We use LightGBM (Ke et al. 2017) – an efficient tree-based classifier – and FlaML (Wang et al. 2021), a system that infers hyperparameters based on dataset characteristics in pipelines.

Moral Foundations Theory (MFT) (Graham, Haidt, and Nosek 2009) explains variations in moral reasoning through five modular foundations.It espouses that liberals express individualizing foundations (care and fairness) while conservatives express binding foundations (loyalty, authority, and sanctity) relatively more. We characterize users' language with FrameAxis (Kwak et al. 2021), a dictionary embedding technique, to identify a user's value for each foundation. It supplies measures, *bias* and *intensity*. Importantly, dictionary-embeddings are generally a refinement over dictionaries alone, particularly for smaller documents, however they do not capture the complexities of human language. For example, such approaches will not handle negations (for example, "I do not care") and do not consider the context around word usage. Large-language model (LLM) approaches may improve these deficits; however, LLMs introduce their own complexities (Liscio et al. 2023) and the dictionary/embeddings approaches are better validated.

Grievance Dictionary (Van der Vegt et al. 2021) is curated for threat assessment, including categories such as fixation, violence, and paranoia. It is validated on social media data, and provides features for distinguishing extremist texts.

State-Of-The-Art Baselines. In Section 6, we compare our approach to three state-of-the-art detection approaches. *UUS* (Darwish et al. 2020) encodes the k most active users, applies dimensionality reduction, clusters these embeddings, and assigns clusters stances via expert annotation. The authors tune parameters including; k, features (based on retweets, retweeted accounts, and hashtags), dimensionality reduction schemes, and clustering schemes. They recommend encoding 1000 users via retweets, then applying UMAP and Mean-Shift. *UUS+* (Samih and Darwish 2021) extends *UUS* by finetuning BERT with *UUS*-labels; applying it to remaining users. Finally, *TIMME* (Xiao et al. 2020) is a supervised multi-task multi-relation deep graph method using five user relationships to embed and classify users.

4 Ideology Framework and Implementation

In this section, we describe our ideology pipeline framework in two parts; Section 4.1 partitions pipelines into four components and Sections 4.2 and 4.3 provides component implementation details.

4.1 Pipeline Constructions Framework

In this section, we abstract four components of ideology detection, shown in Fig. 1: the dataset, ideological proxy, homophilic lens, and inference architecture.

The Dataset is a set of unlabelled users and their activity metadata within a context. It has an underappreciated effect on observed pipeline performance. Section 5 discusses classification *difficulty* and introduces our evaluation datasets.

The Ideological Proxy infuses ideological knowledge via weak supervision. A user subset is labeled (left, right, or far-right) via ideology-correlated behaviors, such as sharing hashtags, following political parties, endorsing politicians, or sharing news media. See Section 4.2 for details.

The Homophilic Lens characterises ideologically salient user similarity. Section 4.3 describes three homophilic lenses: the lexical lens, the hashtag, and the resharing lens.

The Inference Architecture propagates labels from a user subset to the remaining unlabelled users. We train a classifier on the ideology-proxy-labeled users represented via

Proxy	AL	CT	AV
HASHTAGS	*	*	*
COMMUNITY PARTICIPATION	**	**	*
POLITICIAN ENDORSERS	**	**	**
PARTY FOLLOWERS	***	**	***
MEDIA	***	****	****

Table 1: **Ideology Proxy Qualitative Comparison** for application by practitioners based on three-part criteria; annotation labor minimization (AL), context transferability (CT), and Availability (AV). Criteria are rated out of four-stars.

homophilic lenses. We use LightGBM with FlaML as our classifier[2]. The remainder of this section enumerates the ideological proxies (Section 4.2) and homophilic lenses (Section 4.3) evaluated, and their implementations.

4.2 Implementating Ideological Proxies

Here, we qualitatively compare proxies and describe the implementations of the proxies evaluated in our study.

Proxy Qualitative Comparison. We conduct an assessment of proxies, based on their utility for practitioners. Based on our reading of the thematic review presented in Section 2, we qualitatively build three criteria to assess each proxy. The criteria are designed to partially order proxies as a guide to practitioners. Therefore, we apply a four-star rating (one star is lower) for each criterion, as shown in Section 4.2.

The first criterion we construct is *labor minimization* (AL) defined as the extent to which expert labor is required to generate the proxy. Proxies which require human experts to perform the entire contruction will score one star, whereas an approach with no human intervention scores four stars. The second criterion is *context transferability*[3] (CT), defined as the number and diversity of contexts in which a proxy can be applied. If a proxy is only available in a given context it will score one star, whereas if the proxy is available with no restrictions, it will score four stars. The third criterion is *availability to practitioners*[4] (AV), defined as the extent to which a proxy or its ingredients are openly available, either for ideology detection or independent tasks.

HASHTAGS shared is commonly used as a proxy, but requires domain knowledge and is time-consuming to generate (one star on AL), and generally requires reannotation for every dataset (* for CT). Furthermore, not all social media platforms use hashtags therefore it has a low availability (* for AV). COMMUNITY PARTICIPATION uses user activity in ideological communities (e.g., subreddit posting). The communities tend to be fewer and more persistent (** on AL) and there is some detectable overlap of the communities between platforms (** on CT). However, they are unavailable on some platforms (e.g., Twitter/X) and are inconsistent across countries (* on AV). Furthermore, it requires experts for annotation, and datasets linking communities to ideologies are few. PARTY FOLLOWERS and POLITICIAN ENDORSERS leverage databases of political parties and politicians with their online profiles, which are intermittently available (** on AV). Such databases are usually country- and period-specific – political parties emerge, change, and become relegated in time. The advantage of these proxies is their stability and non-ambivalent nature during the studied context (** on CT). Furthermore, databases do not encode all ideologically relevant information, such as the lean of parties or specific politicians, requiring an expert instead (** and *** on AL, respectively). MEDIA proxies utilize users sharing news media, which often have known political slants. They leverage available and well-maintained data on media slants (*** on AL), which have intrinsic value in communication studies, the news ideology detection task, and general consumer value. There is strong evidence linking news readership (Garimella et al. 2021; Bakshy, Messing, and Adamic 2015) and sharing (An, Quercia, and Crowcroft 2014) to ideology. Media slants are fairly consistent across time, media-sharing behaviors occur on most platforms (**** on AV), and media tend to be ideologically consistent across topics (**** on CT). There are limitations to the media proxy (see Section 8), but it outperforms its alternatives in terms of annotation labor, context transferability, and general availability.

Left-Right ideological proxies. We build four proxies.

HASHTAGS proxy requires experts to code hashtags. We qualitatively inspect the $1,000$ most common hashtags in our datasets and label their political lean; -1 if left-leaning, 0 if non-partisan, and 1 if right-leaning. We quantify a user's political lean as the mean of the labeled hashtags they emit and their ideology label as the sign of this lean.

PARTY FOLLOWERS proxy requires collecting the followers of the major political parties' online accounts for each target country (i.e., Australia and USA). We code the political parties by their ideology. The users in the dataset who follow a single party receive the party's ideology label.

POLITICIAN ENDORSERS proxy requires a dataset of politicians, their political affiliations, and social media handles. We use the Twitter Parliamentarian Database (van Vliet, Törnberg, and Uitermark 2020). We code the politicians using their party's ideology (where independents are excluded). Note that independents' exclusion reduces the proxy representativity, but this is preferable to manually labeling all independents. We label users who retweet politicians using the majority vote of the politicians' ideologies.

LEFT-RIGHT MPP (Media Publication Proxy) requires a dataset of media websites with their political slants. We utilize an extensive survey (Park et al. 2021; Newman et al. 2021) of news consumption behavior within English-speaking countries (Australia, New Zealand, UK and the USA), collected in 2020 and 2021 by Reuters. Participants indicated the news media they read and self-reported their political leaning ranging from -3 (extreme left) to 3 (extreme

[2]The hyperparameter `n_estimators` is inferred for the far-right detection due to the sparsity of labeled users; it is fixed to 200 for left-right detection to prevent overfitting. We set the `is_unbalance` flag due to label imbalance.

[3]Note that context transferability has a multiplier effect on annotation labor since a failure to transfer requires reannotation.

[4]We do not discount prior work labor. However, we recognize that availability differs, independently of ideology tasks, and proxies' maintenance should be considered in practitioner guidelines.

right). We compute a publication's slant as the weighted mean political lean of the participants who consume that publication, where each participant is weighted by the inverse number of publications they consume. Since countries' perspectives on what constitutes left- and right-leaning differ, we calibrate scores across countries with the AllSides Media Bias Ratings (AllSides 2022). We encode the ratings' five-point scale onto a numerical scale from -1 to 1. We align each country's scores, minimizing the sum of squared differences between a country's scores and AllSides scores for overlapping publications. Finally, we generate slant scores for each publication as the average slant over all countries and years. We associate publications (and their slants) with their website domains, averaging where a domain is shared. We present the media organizations and their constructed slant scores in online appendix (Appendix 2024). We 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.

Far-Right ideology proxies. We build two proxies.

FAR-RIGHT MPP is constructed from the media slant scores of mainstream media built for LEFT-RIGHT MPP. Next, we label users 'far-right' if their political lean exceeds 0.5 or as 'moderate' otherwise.

MBFC MPP is constructed from the Media Bias Fact Check (Zandt 2022) dataset, including both media slant and veracity, and containing conspiratorial and fake news sources. We label users sharing the right-most media category as 'far-right'.

4.3 Homophilic Lenses

Homophily is the tendency of similar users to be similar (McPherson, Smith-Lovin, and Cook 2001) and is commonly assumed in ideology detection. A *homophilic lens* is a user embedding that encodes ideologically relevant information. Here we convert content about user behavior into numerical vectors. This section details three lenses.

Lexical Lens (USE). Language is a strong indicator of one's political ideology (Cichocka et al. 2016); since a sociolect is formed through associations with others.

Hashtag Lens (HT). Hashtags signal users' interests and the discussion topics they participate in (Bode et al. 2013).

Resharing Lens (RT). Resharing is a signal of endorsement (Metaxas et al. 2015). We assume users endorsing the same people likely share similar ideologies (Van Vliet, Törnberg, and Uitermark 2021).

Implementation. For the *lexical lens*, we preprocess text, to prevent potential data leaks, by removing URLs, hashtags, and mentions. We concatenate each user's tweets and encode them as 512 dimensional vectors via the universal sentence encoder (USE) (Cer et al. 2018). The encoder choice is arbitrary and based on its prior user in literature for social media-originating text (Rashed et al. 2021). For the *hashtag lens*, we use the Term-Frequency Inverse Document Frequency (TF-IDF) of users (i.e., documents) via hashtags (i.e., words) they use if used at least 10 times. TF-IDF is a refinement over the bag-of-words model that weights terms used by their occurrence within a corpus, providing a simple but salient vector representation. Finally, for the *resharing lens*, 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 jth most reshared post ($h_i[j] = 0$ otherwise). In summary, there are three representations of users; lexical $\mathbb{R}^{512}$, hashtag $\mathbb{R}^{|\#\text{hashtags}|}$, reshare $\mathbb{R}^{1000}$.

Dataset	#Users	#Posts	Country	Hopkins
#QandA	103,074	768,808	AUS	0.2624
#Ausvotes	273,874	5,033,982	AUS	0.2445
#Socialsense	49,442	358,292	AUS	0.2591
Riot	574,281	1,067,794	US	0.1490
Parler	120,048	603,820	US	0.3016

Table 2: **The datasets used in this work:** source, profiling, and country of origin (AUS and US refer to Australia and USA, respectively). The last column represents the Hopkins statistics (Hopkins and Skellam 1954) for the lexical lens.

5 Contexts, Datasets, and Ideology Labels

This section introduces datasets and their contexts. Section 5.1 describes the five datasets, and Section 5.2 shows how we qualitatively construct ideology ground truth, used to evaluate proxies and pipelines' performance.

5.1 Contexts and Datasets

Section 5.1 summarizes the datasets; there are three Australian and two American datasets; one originates from Parler, another is a mixture of Facebook and Twitter, and the remainder are Twitter-based. In prior work datasets, ideology correlates with explicit user behavior (e.g., discussion topics); this simplifies detection but rarely holds in practice. Here, we use data where detection is difficult, as one would likely encounter in the wild. We quantify the *detection difficulty* using Hopkin's statistics (Hopkins and Skellam 1954) of the lexical lens, indicating the clustering tendency of data, ranging from 1 (highly clustered, easy detection) to 0 (uniformly distributed, difficult detection). Hopkin's statistic is common measure of clustering tendency, effectively characterizing the probability that embeddings are drawn from a uniform distribution. We assume that embeddings with a high clustering tendency are easier to classify. Note clusters do not necessarily align with classes, however they often do in real-world data; baselines, like UUS and UUS+, directly employ this axiom to infer labels (relying heavily on the underlying clustering tendency of the data). Quantifying detection difficulty of datasets is uncommon in literature and prior work often vary dataset difficulty be construction (Macià, Orriols-Puig, and Bernadó-Mansilla 2008) or require class labels to infer it (Lorena et al. 2019). We employ Hopkin's Statistic as a simple quantification of difficulty (which is not the focus of our work). It is likely related to the decision boundary aspect of classification complexity (Lorena et al. 2019). Section 5.1 shows values $\in [0.14, 0.3]$ indicating no clustering tendency.

Briefly, the datasets are: *#QandA* [Twitter/X] surrounding a political panel show with audience questions; *#Ausvotes* [Twitter/X] surrounding the 2022 Australian Federal Election; *#Socialsense* [Twitter/X and Facebook] (Calderon, Ram, and Rizoiu 2024) surrounding the Australian Black Summer

Bushfires; *Riot* [Twitter] (Kerchner and Wrubel 2021) and *Parler* [Parler] (Aliapoulios et al. 2021) both surrounding the US capitol insurrection. See (Appendix 2024) for details.

5.2 Build a Ground Truth

We qualitatively annotate a subset of #QandA users to generate both a left-right and far-right ground truth.

Left-Right Ground Truth. Due to the imbalance and sparsity of some ideological classes[5], we employ the proposed pipeline to construct a candidate set of users for manual annotation. Platforms such as X/Twitter have been shown to lean-left, and the imbalance in datasets (such as Q+A which attracts a left-leaning audience) can be substantial. While it can be argued that using the pipeline to generate a ground truth to train future pipelines may skew the data selection, it has advantages over the alternatives. For example, (1) conducting a manual search through a random candidate set and generating a proportionately low-volume of right-leaning users is prohibitively expensive, and (2) employing a proxy directly as our ground truth (following the baselines we compare against) defeats the purpose of evaluating the proxies and introduces significant biases.

We generate the candidate set using the following four components; (1) we select each of the four proxies (HASHTAGS, PARTY FOLLOWERS, POLITICIAN ENDORSERS, LEFT-RIGHT MPP), (2) using labels derived from the selected proxy, we train the classifier to predict user labels (since even proxy do not necessarily produce sufficient volumes of right-leaning users), (3) we apply to proxy-trained classifier to the entire #QandA dataset (including those already labelled), (4) finally, we extract the 100 left- and 100 right-leaning users with the highest classifier confidence (estimated through the classifier sigmoid scores). We collect the pool of 800 users in one set, deduplicate, shuffle it, and remove users who are unavailable (either private or suspended). This results in 695 users; we sample 200 users, inspect their profiles and categorize them as left-leaning, right-leaning, far-right, or indeterminable.

Next, two experts manually labeled each profile. The experts both had extensive knowledge of the Australian political context, and were native English speakers. They were given examples of left, right, far-right, and indeterminable user profiles for context. They were instructed to use any signals of ideological-alignment they observed to make their assessments (see (Appendix 2024) for details). Finally, they were given links to each user profile and instructed to categorize them. They achieved moderate inter-annotator agreement i.e., Cohen's κ of 0.515 As a result, our left-right ground truth contains 103 left- and 74 right-leaning users.

Far-Right Ground Truth. Bailo, Johns, and Rizoiu (2023) snowball sample Australian far-right users, starting with a 'seed' user and recovering 'lists' (a Twitter feature documenting similar users) they belong to. They intersected the sample with their dataset, manually validated their far-right status, crawled this validated set's followers, and manually coded these too. They obtained 1,496 users, of which 686 are in #QandA, and serve as our far-right ground truth.

[5]Predicted label counts show this imbalance (Appendix 2024).

	Left-Right				Far-right	
	Hashtags	Party Follow.	Pol. Endors.	L.R. MPP	F.R. MPP	MBFC MPP
USE	0.881	0.868	0.788	0.946	0.691	0.773
HT	0.873	0.876	0.812	0.849	0.559	0.633
RT	0.840	0.844	0.752	0.879	0.538	0.668
USE+HT	0.949	0.879	<u>0.870</u>	0.939	<u>0.715</u>	<u>0.785</u>
USE+RT	0.880	0.821	0.785	<u>0.953</u>	0.666	0.762
HT+RT	0.904	<u>0.914</u>	0.799	0.937	0.570	0.632
all	<u>0.950</u>	0.875	0.854	0.929	0.713	0.785
Prec.	0.889	0.873	0.797	**0.892**	0.516	**0.530**
Recall	0.857	0.820	0.794	**0.902**	0.540	**0.557**
F1	0.855	0.821	0.766	**0.893**	0.636	**0.720**

Table 3: **Determine the optimal proxy and lens combination.** *(top)* AUC ROC for each combination of lenses (rows) and proxy (columns). The underlines show the best lens for a given proxy. *(bottom)* The precision, recall and macro-F1 for each proxy averaged over all lens combinations. The bold show the best-performing proxy.

6 Proxy Bias, Baselines, and Validation

In this section, we first quantify proxy bias (i.e., representativity) and homophilic lens interaction effects, by enumerating all pipeline constructions, in Section 6.1. Next, we present a pipeline construction that outperforms three state-of-the-art methods in Section 6.2. Finally, we evaluate transfer learning across contexts, illustrating 'in-context' training superiority, and test cross-proxy performance in Section 6.3.

To avoid confusion, Section 6.1 employs both ground-truth and Section 6.2 uses the left-right ground-truth constructed in Section 5.2 for the #QandA dataset. Section 6.3 does not utilize the constructed ground truth. In its first segment it trains on labels derived from one proxy and tests on labels derived from another, with fixed dataset #QandA. In its second segment it trains on users from one dataset and tests on users from another, with fixed proxy LEFT-RIGHT MPP.

6.1 Quantifying Proxy Bias

Here we jointly assess ideological proxy and homophilic lens combinations and their performance against our ground truths, to infer proxy representativity. The top section of Section 6 shows all combinations. The columns represent proxies, and the rows show the seven possible concatenations of our lens implementations. We use the respective ground truth for validation and testing in a $50\% : 50\%$ split, employing the validation set for threshold calibration (for converting continuous scores to discrete predictions), and removing neutral ideologies from training, as they do not appear in testing. Cells show AUC ROC scores for pipelines trained with respective proxy and lens combinations. A higher AUC ROC score is better with a maximum score of 1 and a random baseline of 0.5. The bottom section of Section 6 shows the precision, recall and F1, averaged over all lens combinations. The purpose is to quantify how well proxies represent 'true' ideology, approximated via our ground truth.

Results. There are two main conclusions. First, Section 6

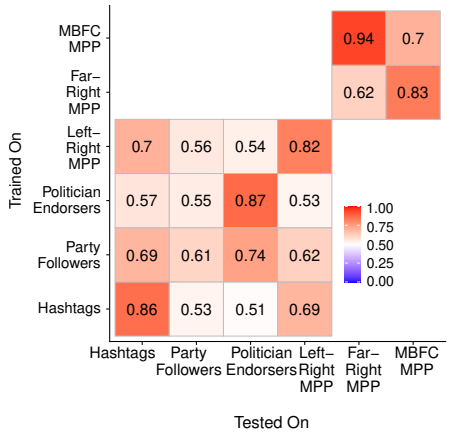

Figure 2: **Self- and cross-proxy generalization.** The AUC ROC of ideology detection on #QandA when trained on one proxy (y-axis) and tested on another (x-axis) for left-right far-right proxies.

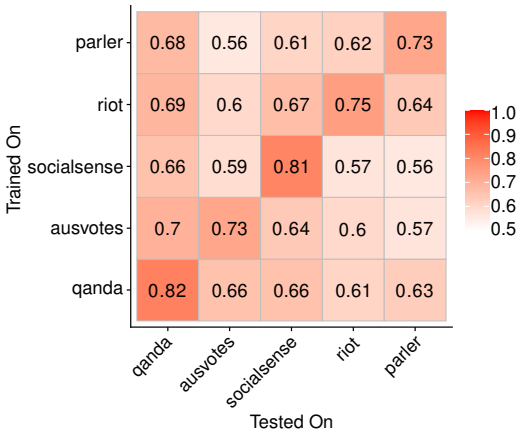

Figure 3: **Context generalization.** AUC ROC of LEFT-RIGHT MPP trained on one dataset (y-axis) and tested on another (x-axis).

Method	*UUS*	*UUS+*	*TIMME*	Ours
Macro-F1	$0.60_{\pm 0.23}$	$0.61_{\pm 0.26}$	0.88	0.92
AUC ROC	–	$0.76_{\pm 0.15}$	0.89	0.95

Table 4: **Baselines.** Left-right classification performance of baselines vs. our pipeline on the ground truth. We report the mean and standard deviation over all setup combinations for *UUS* and *UUS+*. Note that *UUS* does not produce a score, only labels; therefore, AUC ROC cannot be computed for it.

(bottom) shows that MPP consistently outperforms other proxies for left-right detection. In order of representativity, we have LEFT-RIGHT MPP, HASHTAGS, PARTY FOLLOWERS, and POLITICIAN ENDORSERS. The MBFC MPP is the most performant for far-right ideology. This is significant, as we have shown that media-based proxies are both qualitatively advantageous and optimal for representativity; providing clear guidelines for practitioners. Second, Section 6 (top) shows that no homophilic lens dominates all others and the best-performing lens combination changes for each proxy. This may explain unclear conclusions within the literature, where lens optimization is performed in isolation of other pipeline components (e.g., proxies). Despite the lack of a dominating lens, we observe that pipelines containing the lexical lens generally outperform their peers, and USE by itself (first row) has competitive performances. In addition, USE is the only platform-independent lens.

6.2 Prediction Performance Against Baselines

Baselines. We evaluate a pipeline construction against three state-of-the-art stance detection techniques: *UUS* (Darwish et al. 2020), *UUS+* (Samih and Darwish 2021), and *TIMME* (Xiao et al. 2020) – detailed in Section 2. For *UUS*, the authors' recommended setup (UMAP+Mean-Shift, retweet features, and 1000 active users) does not produce any clusters on #QandA. To render UUS competitive, we enumerate the setups similarly to their work. We fix the dimensionality reduction to `UMAP` and clustering to `Mean-Shift` following their recommended setup. We use the default *scikit-learn* settings (*n_neighbors=15*, *min_dist=0.1*, *n_components=2*, *metric=cosine*), and do not enumerate different hyperparameters to (1) faithfully replicate their work, and (2) simulate the experience of a time-poor practitioner. Furthermore, we implement setups for every combination of features (retweets, retweeted accounts, and hashtags) and number of active users (500, 1000, and 5000). In addition, *UUS* only reports the most active users' labels; however, our ground truth users are not the most active. Instead, we use UMAP and Mean-Shift inference methods to acquire labels for these users. For *UUS+*, we use the same set of *UUS* setups. Following the authors, we utilize $\text{BERT}_{\text{base multilingual}}$, using the HuggingFace implementations with PyTorch. We fine-tune BERT by adding a fully-connected dense layer followed by a softmax output layer. We minimize the cross-entropy loss over the training data. As it is not specified by the authors, we choose to fine-tune for 10 epochs (a sufficient quantity for our data volume). Finally, for *TIMME*, we use all relations except the followership network, which is prohibitive to acquire.

Predicting Human-Annotated Ideology. We evaluate performance using the left-right ground truth (see Section 5.2) with a 5-fold cross-validation (where applicable). For this task, we use the pipeline constructed from the LEFT-RIGHT MPP and the USE+RT homophilic lens (the best-performing combination from Section 6). Section 6.2 shows the F1-macro and AUC ROC scores for each technique. We make several observations. First, our approach consistently outperforms all baselines, with the next best being *TIMME*. Second, *UUS* and *UUS+* show low mean performance and high standard deviation. Most setups failed to cluster users and were removed before computing the mean and standard deviation. Furthermore, the clusters required an expert for labeling. Our pipeline construction has practical advantages over these baselines and outperforms them.

6.3 Cross Proxy and Context Generalization

Cross Proxy Generalization. Here, we characterize the robustness of ideological proxies through their self- and cross-consistency. *Self-consistency* indicates how well the pipeline predictions trained with a given proxy align with the same proxy on a test set. We evaluate self-consistency using a 5-fold cross-validation. *Cross-consistency* indicates that two proxies capture similar ideological signals. We evaluate the directed cross-consistency of a source $\longrightarrow$ target proxy by deploying a pipeline with the source proxy to predict the ideology of every user in the #QandA dataset and testing against the ideology labels set by the target proxy. We report the performance over users whom the target proxy labels, and use a one-vs-one scheme to adjust to the multiclass setting. For a given proxy, we deploy the pipeline with the best lens combination as per Section 6.

Section 6.2 shows the AUC ROC performance for every pair of source $\longrightarrow$ target proxy for both left-right and far-right ideology detection. *The self-consistency* (main diagonal) is high for all left-right pipelines, except PARTY FOLLOWERS. It is worth noting, POLITICIAN ENDORSERS has high self-consistency but a low prediction performance against the ground truth (see Section 6). This suggests that politician endorsement behavior is distinct from prototypical ideological behavior. Note, far-right proxies have relatively low self-consistency, perhaps due to the sparsity of far-right users.

Section 6.2 shows *cross-consistency* of left-right pipelines is relatively low, except for LEFT-RIGHT MPP and HASHTAGS. This supports prior work (Cohen and Ruths 2013; Alkiek, Zhang, and Jurgens 2022) arguing that different proxies confer diverse ideology prototypes. The LEFT-RIGHT MPP and HASHTAGS proxies generalize well to each other and the ground truth (see Section 6), suggesting they accurately represent *true ideology*. Both far-right proxies generalize well on each other, but their performance on the ground truth is relatively weak. This indicates they represent similar behaviors not fully aligned with ideology.

In-Context Dominance. Researchers often implicitly suggest political signals from one context transfer to others. Here we demonstrate the importance of 'in-context' training. Each dataset is typically associated with a distinct context (see Section 4.1). We evaluate transfer-learning across contexts by training a pipeline (constructed with the LEFT-RIGHT MPP proxy and the USE+RT lenses) on one dataset and testing on another dataset. Section 6.2 shows the 5-fold cross-validation AUC ROC performance of left-right ideology detection for every pair of datasets. Intuitively, models perform best when trained and tested on the same dataset (i.e., in-context). However, we observe a significant performance drop-off with transfer learning (off-diagonal). Despite this, we see relatively better transfer learning between contexts that share traits. Models trained in Australian contexts perform better when tested within the Australian context, and noticeably underperform when tested in US contexts. Moreover, a further reduction is observed when training or testing with the Parler dataset (i.e., a different social platform context). These observations indicate that signals of ideology differ between contexts. While transfer learning performs better in similar contexts, 'in-context' training is significantly more effective.

	#QandA	#Ausvotes	#Socialsense	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	$^{10}/_{20}$	$^{7}/_{20}$	$^{9}/_{20}$	$^{8}/_{20}$	$^{12}/_{20}$	$^{46}/_{100}$

Table 5: **Moral Foundations Hypotheses testing.** The number of times the MFT hypotheses tests are significant for each foundation (rows) and dataset (columns).

7 Psychosocial Analysis of Ideology Cohorts

In this section, we test four hypothesis sets for psychosocial asymmetries of ideologies, relating to morality, grievance, nationalism, and dichotomous thinking. This serves two purposes: an application case study for practitioners and to supply online evidence bases for conclusions of prior work. We use pipelines constructed from the MBFC MPP and LEFT-RIGHT MPP proxies, alongside the USE lens for its applicability across all datasets. The first pipeline labels users as 'far-right'. If users are not labeled 'far-right', the second pipeline assigns them as 'left', 'neutral', or 'right'. For most analysis below, we highlight results on a single dataset, however we produce the relevant plots for all datasets and label distributions in the supplementary material (Appendix 2024).

Testing Moral Foundations. We begin by evaluating MFT hypotheses. There are five hypotheses relating to individualizing (liberal) and binding (conservative) foundations. We use a Wilcox Rank Sign Test (95%), with Holm adjustment for family-wise error, to evaluate the support for each moral foundations hypothesis in each dataset. We test these hypotheses with the *bias* and *intensity* measures and both "left vs. right" and "left vs. far-right" (i.e., each combination has four hypotheses). Section 7 shows the number of statistically significant tests for each moral foundations hypothesis in each dataset. Overall, only 46% of hypotheses are supported, marginally favoring the individualizing over the binding hypotheses. This inconsistency, seen in prior work (Wang and Inbar 2021; Thomas et al. 2022), suggests MFT applies differently online than it does offline.

Next, given the lack of support for MFT, we test an alternative hypothesis, that *right-leaning users, relative to left-leaning users, exhibit vice over virtue foundations*. For each moral foundation, we assign each user a virtue/vice score equal to their intensity, if their bias is positive/negative, respectively. This segregates the population into vice or virtue users. In Fig. 4a, we plot each foundation's mean vice and virtue scores for each ideological group in the #QandA dataset. We observe that a significant proportion of right-leaning users partake in the language of vice rather than virtue compared to left-leaning users. We apply the Wilcox Rank Sign Test (95%) between the means of ideological groups,

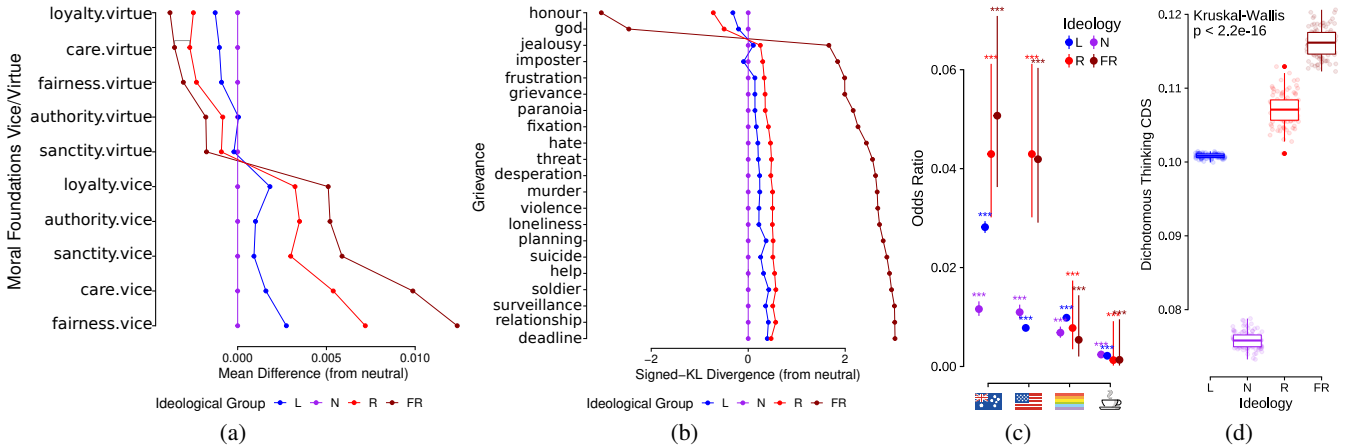

Figure 4: **(a)(b) Distribution of psychosocial properties** for ideological groups for #QandA and #Ausvotes, respectively. Line color represents ideological groups, and the y-axis shows psychosocial categories. *(a) Vices-Virtues.* The x-axis is the mean difference for each ideological group from neutral, for Moral Foundations vice and virtue categories. *(b) Grievances.* The x-axis is the signed-KL divergence of each group ideological group from neutral for grievance categories. (c) **Emoji Nationalism.** The odds (y-axis) of observing an emoji (x-axis) for a user given their ideological group (color), for #QandA. The odds are determined via logistic regression with no reference group. (d) **Dichotomous Thinking.** The bootstrapped prevalence distribution of dichotomous thinking CDS (y-axis) in tweets by users from ideological groups (x-axis), for #QandA.

for each category, and find all are significantly different[6]. We show that this is relatively consistent across all datasets in the online appendix (Appendix 2024), This provides a consistent moral asymmetry in the online context.

Testing Extremists' Association With Grievance. Early signals of extremism are of particular concern to national security and law enforcement practitioners. Prior work suggests that *extreme ideologies hold more grievance and grudge beliefs than moderates.* We use the Grievance dictionary (Van der Vegt et al. 2021) to quantify users' grievance and grudge language. In Fig. 4b, we plot the Kuller-Leibach divergence (signed by mean difference) between the distribution of each ideological group from the neutral group for each category with the #Ausvotes dataset. We apply the Kruskal-Wallis Test (95%) between ideological groups, for each category, and find all are significantly different. We observe that the far-right users differ significantly from the other ideologies in all categories and generally use more grievance language. Notably, in the #Ausvotes dataset, the far-right users use *honor* and *god* type language less than other groups. In the online appendix (Appendix 2024), we show that this hypothesis holds for most datasets. A takeaway for practitioners is that far-right language and threat assessment indicators overlap, suggesting a method to build effective public safety tools.

Testing Nationalism Via Emoji. Here we add online evidence that *the right-wing are associated with nationalism* via emojis. This hypothesis is widely accepted (and definitional), and supporting it validates our inferred ideologies.

Fig. 4c shows the odds of observing an emoji, given a user's ideological group in #QandA. Point ranges indicate the 95% confidence interval. The significance of the emoji in predicting ideological groups, via the Wald Test, is indicated with stars. We make several observations. First, 🇦🇺 is used more by ideological groups than neutral users. Second, the right (and far-right) use 🇦🇺 and 🇺🇸 significantly more than the left. Third, 🏳️‍🌈 is used marginally more by the left than other groups. Finally, we include ☕ as a control (showing no associations with any ideology). We conclude that nationalism, via national flags, is associated with our inferred right-leaning ideologies. The use of 🇺🇸 could be evidence of imported ideology from America to Australia. 🏳️‍🌈 is only marginally associated with the left.

Testing Dichotomous Thinking. Recent work suggests *the right-ideologies applying black-and-white thinking relatively more than left-wing ideologies*. Following (Bathina et al. 2021), we match n-grams relating to cognitive distortions schema (CDS) in user tweets in #QandA. We measure the prevalence – the empirical probability of observing a CDS n-gram in a tweet given an ideological group. Additionally, we utilize 100 bootstrap samples (i.e., repeated sampling of tweets) to estimate the prevalence distributions. Fig. 4d shows that all non-neutral ideologies exhibit a significantly higher prevalence of dichotomous thinking, with right-leaning higher than left-leaning and far-right higher than right-leaning. We perform T-Tests (95%) to compare group means and find all differ significantly from each other. These findings support prior literature (Meyer 2020), and extend it by showing that the far-right might engender an even greater extent of dichotomous thinking. Other cognitive distortions' prevalences are summarized in the online appendix (Appendix 2024).

[6]Except between the right and far-right in the *care-virtue* category, which is irrelevant to our conclusions

8 Conclusion

This work proposes a framework for ideology detection pipelines and quantifies biases introduced by ideological proxies. It tests hypotheses of the psychosocial asymmetries of ideological groups, in the online space. We present an evaluation of ideological proxies; qualitatively, indicating proxies that minimize labor, are transferable across multiple contexts, and are available; and, quantitatively, indicating the representativity and robustness of proxies. We find the media proxy advantageous, and a pipeline constructed from it and the lexical lens to be optimal, outperforming state-of-the-art approaches. Such research is essential for furnishing practitioners with actionable guidelines for ideology detection and its practical applications.

Limitations. The media proxy has several limitations. Firstly, it relies on the availability of up-to-date media slant data. Publication slant can shift over time, and publication emergence, acquisition and closure can hold significance (particularly on ideological fringes). Secondly, some users share media to refute it. Thirdly, article slants may differ from publication slants. Finally, it will not produce a perfectly representative user subset, although media sharing ubiquity makes it relatively competitive. Furthermore, our conceptualization of ideology is simplistic, and some political systems are complex requiring complex ideological proxies (which are largely unavailable).

Future Work. We limit our scope to English-speaking Anglocentric countries due to the expertise and language proficiency of the author team. However, the study could be applied broadly. Newman et al. (2021) provides data annually for 46 diverse countries, including segments of the Global South. Our study could be extended to any other uniaxial political setting with little amendment.

References

Aldayel, A.; and Magdy, W. 2019. Your stance is exposed! analysing possible factors for stance detection on social media. *CSCW*.

Aliapoulios, M.; Bevensee, E.; Blackburn, J.; Bradlyn, B.; Cristofaro, E. D.; Stringhini, G.; and Zannettou, S. 2021. A Large Open Dataset from the Parler Social Network.

Alizadeh, M.; Weber, I.; Cioffi-Revilla, C.; Fortunato, S.; and Macy, M. 2019. Psychology and morality of political extremists: evidence from Twitter language analysis of alt-right and Antifa. *EPJ DS*.

Alkiek, K.; Zhang, B.; and Jurgens, D. 2022. Classification without (Proper) Representation: Political Heterogeneity in Social Media and Its Implications for Classification and Behavioral Analysis. In *ACL*.

AllSides. 2022. AllSides Media Bias Ratings. https://www.allsides.com/media-bias/ratings. Accessed: 2022-04-08.

An, J.; Quercia, D.; and Crowcroft, J. 2014. Partisan sharing: Facebook evidence and societal consequences. In *COSN*.

Appendix, O. 2024. Supplementary Material: Practical Guidelines for Ideology Detection Pipelines and Psychosocial Applications. https://bit.ly/ideology_detection.

Auxier, B.; and Anderson, M. 2021. Social media use in 2021. *Pew Research Center*.

Badaan, V.; Hoffarth, M.; Roper, C.; Parker, T.; and Jost, J. T. 2023. Ideological asymmetries in online hostility, intimidation, obscenity, and prejudice. *Scientific reports*.

Badawy, A.; Lerman, K.; and Ferrara, E. 2019. Who falls for online political manipulation? In *WWW*.

Bailo, F.; Johns, A.; and Rizoiu, M.-A. 2023. Riding information crises: the performance of far-right Twitter users in Australia during the 2019–2020 bushfires and the COVID-19 pandemic. *Information, Communication & Society*.

Bakshy, E.; Messing, S.; and Adamic, L. A. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science*.

Barberá, P. 2015. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political analysis*.

Bathina, K. C.; Ten Thij, M.; Lorenzo-Luaces, L.; Rutter, L. A.; and Bollen, J. 2021. Individuals with depression express more distorted thinking on social media. *Nature Human Behaviour*.

Betz, M. 2016. Constraints and opportunities: what role for media development in countering violent extremism?

Bode, L.; Hanna, A.; Sayre, B.; Yang, J.; and Shah, D. V. 2013. Mapping the political Twitterverse: Finding connections between political elites.

Booth, E.; Lee, J.; Rizoiu, M.-A.; and Farid, H. 2024. Conspiracy, misinformation, radicalisation: understanding the online pathway to indoctrination and opportunities for intervention. *Journal of Sociology*.

Calderon, P.; Ram, R.; and Rizoiu, M.-A. 2024. Opinion Market Model: Stemming Far-Right Opinion Spread using Positive Interventions. In *ICWSM*.

Cann, T. J.; Weaver, I. S.; and Williams, H. T. 2021. Ideological biases in social sharing of online information about climate change. *Plos one*.

Carr, H. J.; Dancho, R.; Michaud, K.; Chiang, P.; Damoff, P.; Lloyd, D.; MacGregor, A.; McKinnon, R.; Noormohamed, T.; Schiefke, P.; Shipley, D.; Popta, T. V.; and Zuberi, S. 2022. Rise of Ideologically Motivated Violent Extremism in Canada. Technical report.

Cer, D.; Yang, Y.; Kong, S.-y.; Hua, N.; Limtiaco, N.; John, R. S.; Constant, N.; Guajardo-Cespedes, M.; Yuan, S.; Tar, C.; et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Chakraborty, S.; Goyal, P.; and Mukherjee, A. 2022. Fast Few Shot Self-attentive Semi-supervised Political Inclination Prediction. In *ICADL*.

Cichocka, A.; Bilewicz, M.; Jost, J. T.; Marrouch, N.; and Witkowska, M. 2016. On the grammar of politics—or why conservatives prefer nouns. *Political Psychology*.

Cohen, R.; and Ruths, D. 2013. Classifying political orientation on Twitter: It's not easy! In *ICWSM*.

Cohrs, J. C. 2012. Ideological bases of violent conflict.

Darwish, K.; Stefanov, P.; Aupetit, M.; and Nakov, P. 2020. Unsupervised user stance detection on twitter. In *ICWSM*.

Eady, G.; Bonneau, R.; Tucker, J. A.; and Nagler, J. 2020. News sharing on social media: Mapping the ideology of news media content, citizens, and politicians.

Garimella, K.; Smith, T.; Weiss, R.; and West, R. 2021. Political polarization in online news consumption. In *ICWSM*.

Graham, J.; Haidt, J.; and Nosek, B. A. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*.

Gu, Y.; Chen, T.; Sun, Y.; and Wang, B. 2016. Ideology detection for twitter users with heterogeneous types of links. *arXiv preprint arXiv:1612.08207*.

Hopkins, B.; and Skellam, J. G. 1954. A New Method for determining the Type of Distribution of Plant Individuals. *Annals of Botany*.

Jiang, J.; Ren, X.; and Ferrara, E. 2023. Retweet-BERT: Political Leaning Detection Using Language Features and Information Diffusion on Social Networks. *ICWSM*.

Jost, J. T. 2017. Asymmetries abound: Ideological differences in emotion, partisanship, motivated reasoning, social network structure, and political trust. *Journal of Consumer Psychology*.

Kariryaa, A.; Rundé, S.; Heuer, H.; Jungherr, A.; and Schöning, J. 2022. The role of flag emoji in online political communication. *Social Science Computer Review*.

Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; and Liu, T.-Y. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *NeurIPS*.

Kemmelmeier, M.; and Winter, D. G. 2008. Sowing patriotism, but reaping nationalism? Consequences of exposure to the American flag. *Political Psychology*.

Kerchner, D.; and Wrubel, L. 2021. U.S. Capitol Riot and #TrumpRally Tweet IDs.

Kwak, H.; An, J.; Jing, E.; and Ahn, Y.-Y. 2021. FrameAxis: characterizing microframe bias and intensity with word embedding. *PeerJ Computer Science*.

Lahoti, P.; Garimella, K.; and Gionis, A. 2018. Joint non-negative matrix factorization for learning ideological leaning on twitter. In *WSDM*.

Lai, A.; Brown, M. A.; Bisbee, J.; Tucker, J. A.; Nagler, J.; and Bonneau, R. 2022. Estimating the ideology of political youtube videos. *Political Analysis*.

Li, J.; Longinos, G.; Wilson, S.; and Magdy, W. 2020. Emoji and self-identity in Twitter bios. In *NLP+CSS*.

Liscio, E.; Araque, O.; Gatti, L.; Constantinescu, I.; Jonker, C. M.; Kalimeri, K.; and Murukannaiah, P. K. 2023. What does a text classifier learn about morality? An explainable method for cross-domain comparison of moral rhetoric. In *ACL*.

Liu, S.; Luo, Z.; Xu, M.; Wei, L.; Wei, Z.; Yu, H.; Xiang, W.; and Wang, B. 2023. Ideology Takes Multiple Looks: A High-Quality Dataset for Multifaceted Ideology Detection. In *EMNLP*.

Lorena, A. C.; Garcia, L. P.; Lehmann, J.; Souto, M. C.; and Ho, T. K. 2019. How complex is your classification problem? a survey on measuring classification complexity. *CSUR*.

Macià, N.; Orriols-Puig, A.; and Bernadó-Mansilla, E. 2008. Genetic-based synthetic data sets for the analysis of classifiers behavior. In *HAIS*.

McCauley, C.; and Moskalenko, S. 2008. Mechanisms of Political Radicalization: Pathways Toward Terrorism. *Terrorism and Political Violence*.

McPherson, M.; Smith-Lovin, L.; and Cook, J. M. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology*.

Metaxas, P.; Mustafaraj, E.; Wong, K.; Zeng, L.; O'Keefe, M.; and Finn, S. 2015. What do retweets indicate? Results from user survey and meta-review of research. In *ICWSM*.

Meyer, P. H. 2020. Political Ideology and Black-and-White Thinking.

Mokhberian, N.; Abeliuk, A.; Cummings, P.; and Lerman, K. 2020. Moral framing and ideological bias of news. In *SocInfo*.

Newman, N.; Fletcher, R.; Schulz, A.; Andi, S.; Robertson, C. T.; and Nielsen, R. K. 2021. Reuters Institute digital news report 2021. *Reuters Institute for the Study of Journalism*.

O'Hagan, S.; and Schein, A. 2023. Measurement in the Age of LLMs: An Application to Ideological Scaling. *arXiv preprint arXiv:2312.09203*.

Park, S.; Fisher, C.; McGuinness, K.; Lee, J. Y.; and McCallum, K. 2021. *Digital news report: Australia 2021*. News and Media Research Centre.

Poole, K. T.; and Rosenthal, H. 1985. A spatial model for legislative roll call analysis. *American journal of political science*.

Preoţiuc-Pietro, D.; Liu, Y.; Hopkins, D.; and Ungar, L. 2017. Beyond binary labels: Political ideology prediction of Twitter users. In *ACL*.

Radsch, C. 2016. Media Development and Countering Violent Extremism: An Uneasy Relationship, a Need for Dialogue. *Center for International Media Assistance*.

Rao, A.; Morstatter, F.; and Lerman, K. 2022. Partisan asymmetries in exposure to misinformation. *Scientific reports*.

Rao, A. R. 2017. Red, blue and purple states of mind: Segmenting the political marketplace. *Journal of Consumer Psychology*.

Rashed, A.; Kutlu, M.; Darwish, K.; Elsayed, T.; and Bayrak, C. 2021. Embeddings-Based Clustering for Target Specific Stances: The Case of a Polarized Turkey. In *ICWSM*.

Ravi, K.; Vela, A. E.; and Ewetz, R. 2022. Classifying the Ideological Orientation of User-Submitted Texts in Social Media. In *ICMLA*.

Reiter-Haas, M.; Kopeinik, S.; and Lex, E. 2021. Studying Moral-based Differences in the Framing of Political Tweets. In *ICWSM*.

Rizoiu, M.-A.; Graham, T.; Zhang, R.; Zhang, Y.; Ackland, R.; and Xie, L. 2018. # DebateNight: The Role and Influence of Socialbots on Twitter During the 1st 2016 US Presidential Debate. In *ICWSM*.

Samih, Y.; and Darwish, K. 2021. A few topical tweets are enough for effective user stance detection. In *ACL*.

Stankov, L. 2021. From social conservatism and authoritarian populism to militant right-wing extremism. *Personality and Individual Differences*.

Thomas, E. F.; Leggett, N.; Kernot, D.; Mitchell, L.; Magsarjav, S.; and Weber, N. 2022. Reclaim the Beach: How Offline Events Shape Online Interactions and Networks Amongst Those Who Support and Oppose Right-Wing Protest. *Studies in Conflict & Terrorism*.

Tomkins, S. 1963. Left and right: A basic dimension of ideology and personality.

Van der Vegt, I.; Mozes, M.; Kleinberg, B.; and Gill, P. 2021. The grievance dictionary: Understanding threatening language use. *Behavior research methods*.

van Vliet, L.; Törnberg, P.; and Uitermark, J. 2020. The Twitter parliamentarian database: Analyzing Twitter politics across 26 countries. *PLoS one*.

Van Vliet, L.; Törnberg, P.; and Uitermark, J. 2021. Political Systems and Political Networks: The Structure of Parliamentarians' Retweet Networks in 19 Countries. *International Journal of Communication*.

Wang, C.; Wu, Q.; Weimer, M.; and Zhu, E. 2021. FLAML: A fast and lightweight automl library. *MLSys*.

Wang, S.-Y. N.; and Inbar, Y. 2021. Moral-language use by US political elites. *Psychological Science*.

Wikipedia. 2023. Q+A (Australian talk show). https://en.wikipedia.org/wiki/Q%2BA_(Australian_talk_show). Accessed: 2023-04-27.

Xi, N.; Ma, D.; Liou, M.; Steinert-Threlkeld, Z. C.; Anastasopoulos, J.; and Joo, J. 2020. Understanding the political ideology of legislators from social media images. In *ICWSM*.

Xiao, Z.; Song, W.; Xu, H.; Ren, Z.; and Sun, Y. 2020. TIMME: Twitter ideology-detection via multi-task multi-relational embedding. In *KDD*.

Zandt, D. 2022. Media Bias/Fact Check. https://mediabiasfactcheck.com/about. Accessed: 2022-04-08.

Ethics Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, see the Ethics and Broader Impact Statement at the end of this checklist.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, see Contexts, Datasets, and Ideology Labels.
   (e) Did you describe the limitations of your work? Yes, see the Conclusion.
   (f) Did you discuss any potential negative societal impacts of your work? Yes, see the Ethics and Broader Impact Statement at the end of this checklist.
   (g) Did you discuss any potential misuse of your work? Yes, see the Ethics and Broader Impact Statement at the end of this checklist.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Yes.
   (b) Have you provided justifications for all theoretical results? Yes.
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes.
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes.
   (e) Did you address potential biases or limitations in your theoretical framework? Yes.
   (f) Have you related your theoretical results to the existing literature in social science? Yes.
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes.
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? No, we will include our git repository once the paper gets accepted. The repository will include the code and instructions.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes, where applicable.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? No, our study does not require significant compute resources.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes.
   (b) Did you mention the license of the assets? NA
   (c) Did you include any new assets in the supplemental material or as a URL? Yes.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? NA
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? NA
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? NA
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? NA
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA
   (d) Did you discuss how data is stored, shared, and deidentified? NA

Ethical Statement

This work introduces a powerful tool for inferring user ideology based on covert cues such as language patterns. We

demonstrate our tool for detecting far-right ideologies; however, it could, in theory, be used by oppressive regimes to infer the true ideologies of their citizens and expose their opponents (Radsch 2016). The Countering Violent Extremism (CVE) literature (Betz 2016) explores the ethical concerns of developing tools that can be used for oppressive ends and proposes mitigation strategies. There are also privacy concerns, as one's ideology can be viewed as an intimate and private trait that our tool can expose. Additionally, we show how to use our pipeline to profile entire online populations based on their psychosocial characteristics. We argue that the pipeline predictions are not prescriptive; they should be treated as an early warning system, requiring human expert investigation. We further note that we only use expert-inferred political affiliation as our ground truth, not private self-reported political indicator data.

Appendices

This document accompanies the submission. The information in this document complements the submission and is presented here for completeness reasons. It is not required to understand the main paper or reproduce the results.

A Dataset Collection Details

#QandA. We collect discussions related to the Australian panel show Q+A (Wikipedia 2023), where panelists (public figures, politicians, and experts) answer curated audience questions. Twitter participation is encouraged in airings. We collect #QandA using the filter keyword *qanda* during January-December 2020.

#Ausvotes. We collect discussions about the 2022 Australian Federal election, tracking the lead-up and aftermath. It follows the major parties and their leaders: the left-leaning Australian Labor Party led by Anthony Albanese and the right-leaning Liberal-National Coalition led by Scott Morrison. We collect #Ausvotes using the keywords *auspol* and *ausvotes*, and for mentions of *@ScottMorrisonMP*, *@AlboMP*, and *@AusElectoralCom*, between 9 May and 15 June 2022 (the elections occurred on 21 May).

#Socialsense (Calderon, Ram, and Rizoiu 2024) features discussions related to the Australian Black Summer bushfires, which gathered discourse concerning climate change, and contains far-right opinions. #Socialsense contains 90 days of Twitter and Facebook discussions, from 1 November to 29 January 2020.

Riot (Kerchner and Wrubel 2021) features discussions about the January 6th US Capitol Insurrection, including election fraud and insurrection topics. The dataset spans 6 January to 1 February 2021 and was collected with the filter keywords *TrumpRally*, *Democracy*, *USCapitol*, *Capitol*, *DCProtests*, and *AshliBabbit*.

Parler (Aliapoulios et al. 2021) features discussions about the US Capitol Insurrection from Parler. We collect all posts emitted during the day of 6 January 2021.

B All UUS/UUS+ Metrics

Thi section shows all possible runs for the *UUS* and *UUS+*. We notice that in many instances *UUS* fails to seperate clusters, and even in instances where seperation can be achieved many suffer from poor performance. This shows that these techniques lack robustness for more difficult datasets.

Table 6: **All Baseline Performances.** The table shows to performances for all combinations of the *UUS* and *UUS+* baselines.

Representation	Active Users	F1-Macro	AUC ROC	UUS F1-Macro
H	500	0.37	0.68	0.37
H	1000	-	-	-
H	5000	0.37	0.54	0.37
HR	500	-	-	-
HR	1000	-	-	-
HR	5000	0.89	0.92	0.85
R	500	-	-	-
R	1000	-	-	-
R	5000	0.93	0.93	0.87
T	500	-	-	-
T	1000	-	-	-
T	5000	-	-	-
TH	500	0.4	0.58	0.54
TH	1000	-	-	-
TH	5000	0.92	0.91	0.87
TR	500	-	-	-
TR	1000	-	-	-
TR	5000	-	-	-
TRH	500	-	-	-
TRH	1000	-	-	-
TRH	5000	0.41	0.75	0.36

C Left-Right Annotation Procedure

Ideology is the subject of considerable subjectivity, not only because experts have their own ideology, but because annotators are often unclear as to what evidence is permissible for use. For this task we issued the following guidelines to annotators:

> It is not always clear what should count as an ideological signal. For our purposes, we will include the following as signals of ideology:
>
> - If a target user promotes/retweets someone or an organisation with a known ideological affiliation, you may assume that the target endorse them. For example, if a target user retweets a labor MP then you can label the user as 'left'.
> - If a target user, has a stance against someone with a known ideological affiliation, then you might infer that the target user's ideology is the opposing ideology. For example, if a target user calls a labor MP an insult, then you can label the user as 'right'.
> - If a target user expresses a view about a issue related to an ideology, you can infer the user's ideology. For example, if a user supports LGBTQ or environmental issues, then (if there is enough evidence) you may label them as 'left'.

These guidelines aim to increase the clarity of the annotation task. In countries where political affiliation is obvert (e.g. the united states), this labelling task is often unambiguous; however, in Australia ideological signals are often implicit. The full annotation briefing material is available in the code repository [https://github.com/behavioral-ds/ideology_prediction].

Context-Transfer Illustration

Fig. 5 further illustrates the difficulty with utilizing particular proxies as ground truth. We observe that some ideological proxies are consistent across only some contexts (represented by the dashed green boxes). For example, `#RoboDebt` (in relation to an Australian incident) is not relevant to the USA and did not exist before 2016; and, although `@MittRomney` signaled right-wing ideology in 2012, the right has shifted

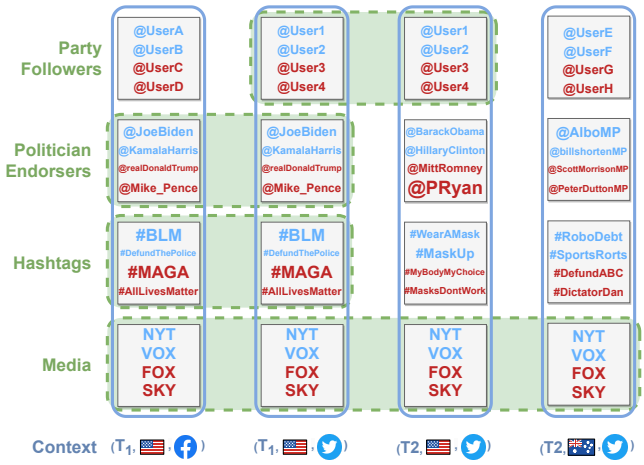

Figure 5: **Most ideology proxies do not generalize across contexts.** The x-axis shows four contexts that vary in time (T_1 and T_2), country (Australia and USA), and platform (Twitter and Facebook). The y-axis show four proxies: endorsing political parties or political figures, using politically charged hashtags and the consumed media slant. The green dashed boxes indicate whether a proxy is applicable across contexts.

since Trump's election.

Prior ideology detection techniques fail to easily *context-switch* and cannot be readily applied to multiple distinct domains.

Media Publication Slants

The media slant scores are shown in Fig. 6, where we observed publications like *Breitbart* and *Fox News* are extremely right-leaning, and *Vox* and *NYTimes* are left-leaning.

Cognitive Distortions Schemata Prevalence

Fig. 7 shows the prevalence of all twelve cognitive distortions in each of the ideological groups, for #QandA. Note that many CDS n-grams are extremely rare (or do not appear), namely; *emotional reasoning* and *mental filtering*. In several CDS the left exhibit higher prevalence, such as *catastrophizing*, *fortune-telling*, *disqualifying the positive*, and *should statements*.

Flag Emoji Hurdle Model

For completeness, we present the results of the hurdle model (used to model zero-inflated count data, such a tokens in a corpus). The hurdle model is a mixed model, comprised of a logistic regression to model the presense of no emoji, and a truncated poisson with log linkage, to model the count of the emoji. Fig. 8 shows the coefficients for each model, including the reference groups. 🇦🇺 is observed more for far-right users in both the zero and the count models. The count models for the other flags show mixed results and not significant.

Precision-Recall of Pipelines

Fig. 9 shows precision and recall for every lens combinations and proxy.

Table 7: **Distribution of Predicted Labels.** The number of users predicted to be in each class (rows) for each dataset (columns). Note that for many datasets there is a significant imbalance toward the left (except Parler hich is a right-leaning platform).

	#QandA	#Ausvotes	#Socialsense	Riot	Parler
Left	80,375	189,233	48,056	339,095	293
Neutral	21,176	79,221	604	227,839	48,829
Right	777	3,689	464	3,624	68,104
Far-right	746	1,731	318	3,723	2,822

Predicted Label Distribution

Table 7 shows the distribution of predicted labels, to provide context for the psychosocial analysis.

Dataset Profiling

Activity levels are often a concern for ideology detection frameworks, given that low-activity users reveal few signals of ideology. Fig. 10 shows the distribution of activity for users for each dataset. It shows long-tailed activity distributions and the proportion of low-activity users. Riot hows a significant proportion of low-activity users, who're often difficult to classify.

Exhaustive Psychosocial Analysis

Grievance

This section shows the difference between ideological groups in terms of grievance categories for all available datasets.

MFT

This section shows the difference between ideological groups in terms of moral foundations for all available datasets.

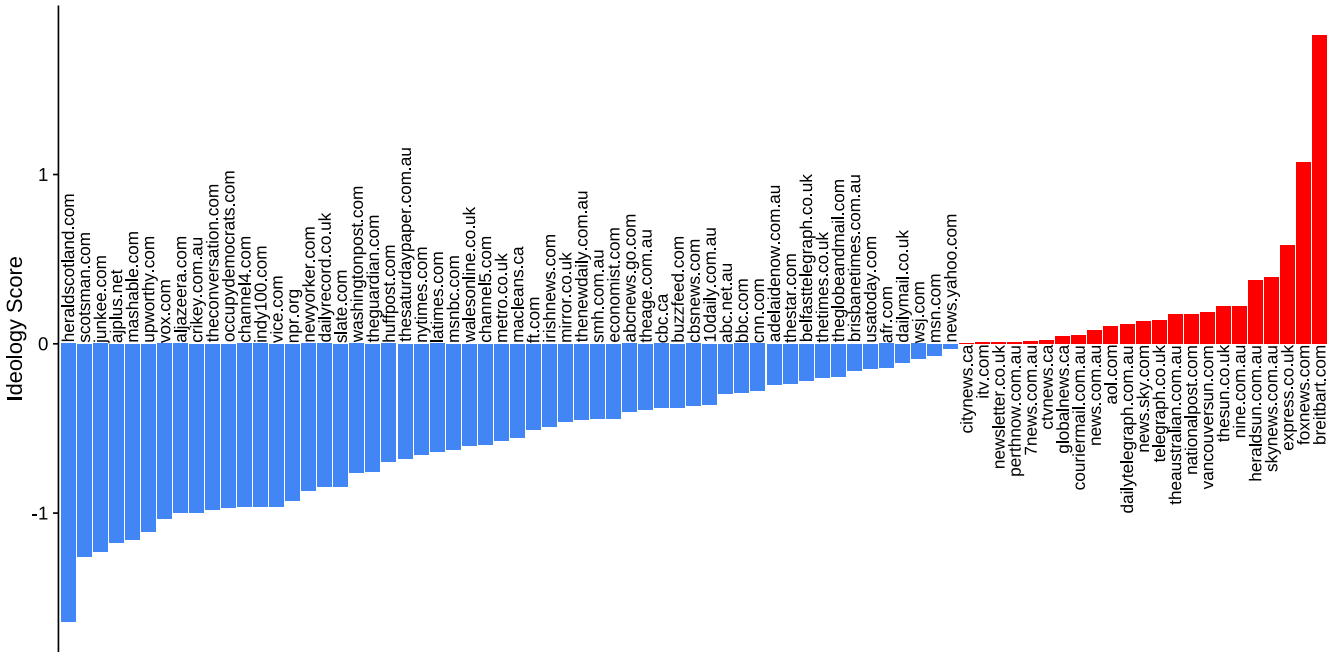

Figure 6: **Media Publication Slants.** The plot shows the slants of Media Publications, as averaged over the year, country, and source point estimates.

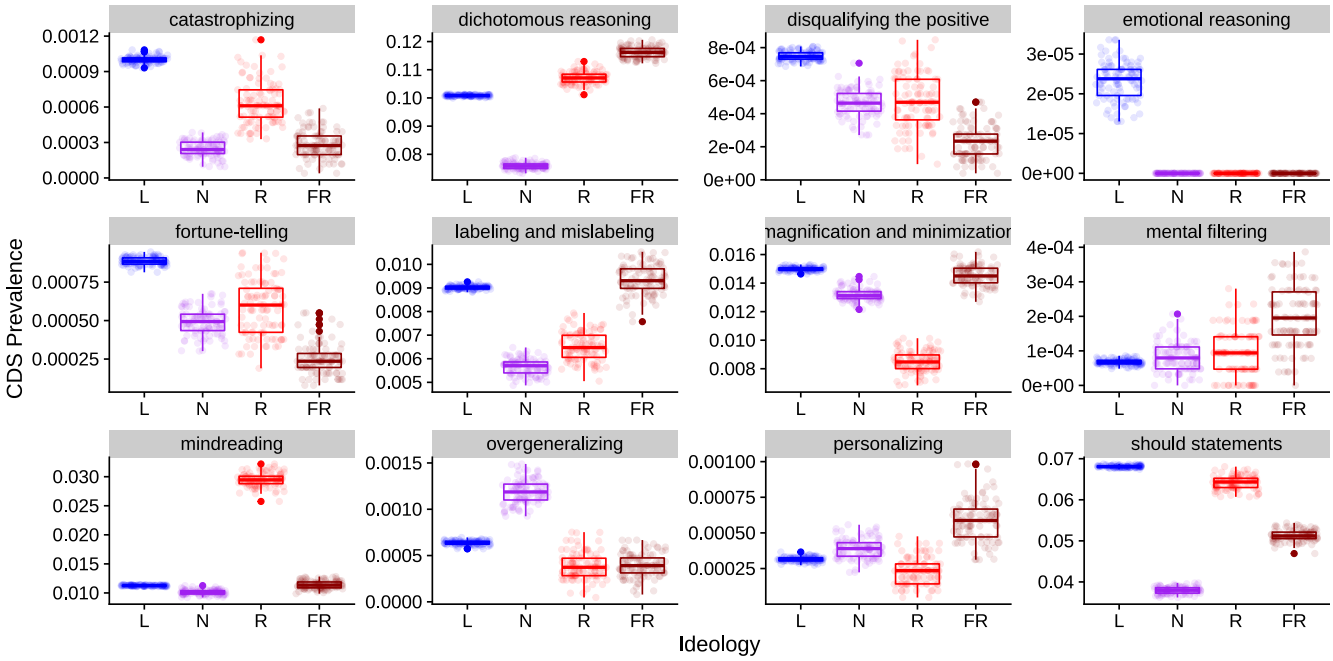

Figure 7: CDS Prevalence

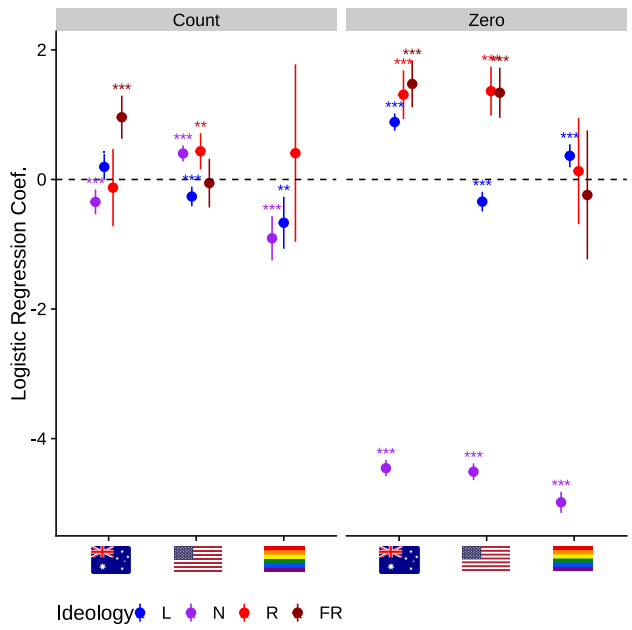

Figure 8: Hurdle Model

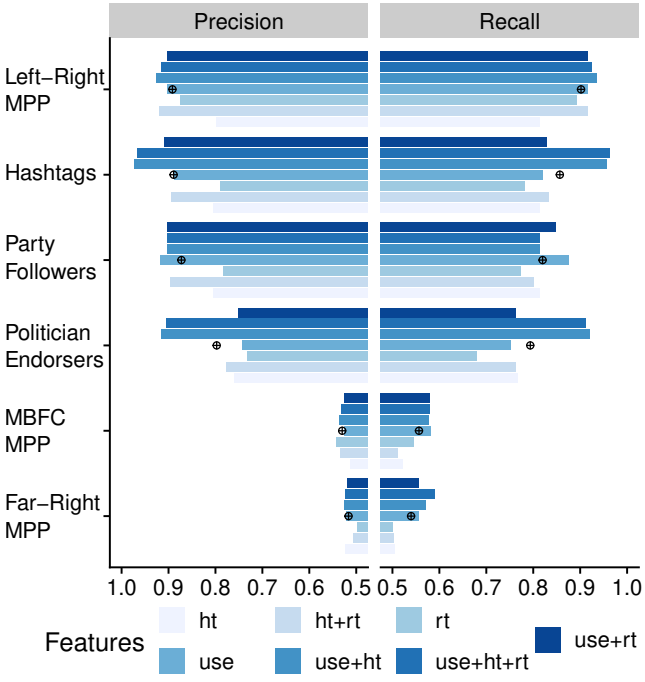

Figure 9: **Precision-Recall.** The plot shows the macro-averaged precision and recall of pipelines, trained with each proxy (y-axis) and each feature set (colors), probability calibrated with the hold-out validation set for F1-macro scores.

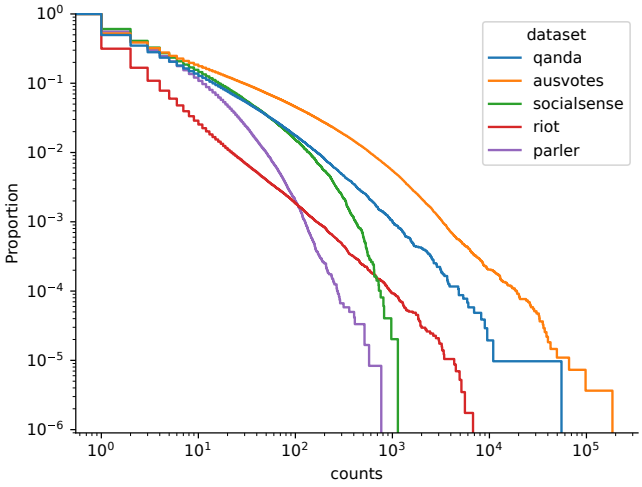

Figure 10: **Activity Distribution.** The log-log ECCDF distribution of activity (number of posts per user) for each dataset.

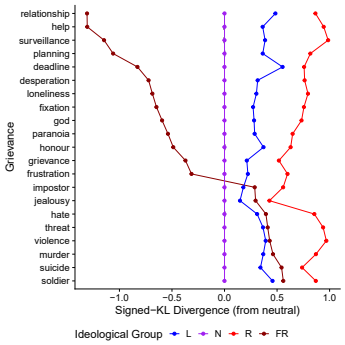

Figure 11: Grievance #QandA

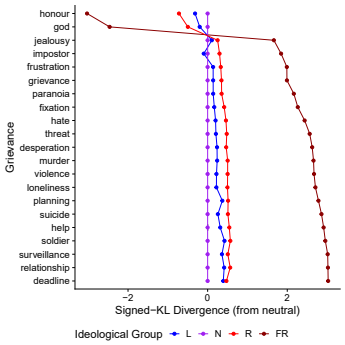

Figure 12: Grievance #Ausvotes

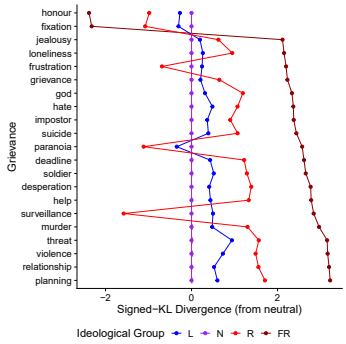

Figure 13: Grievance #Socialsense

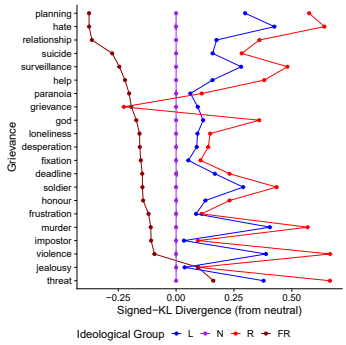

Figure 14: Grievance Riot

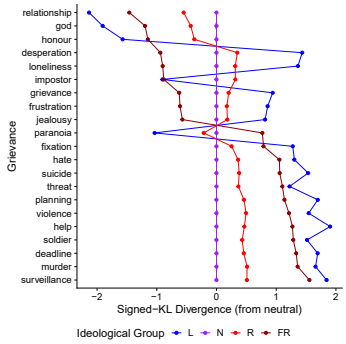

Figure 15: Grievance Parler

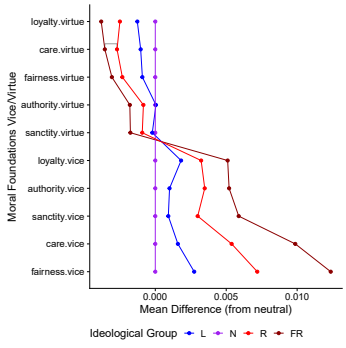

Figure 16: MFT #QandA

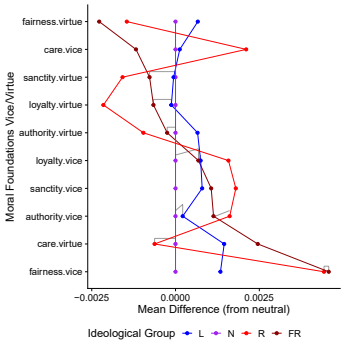

Figure 17: MFT #Ausvotes

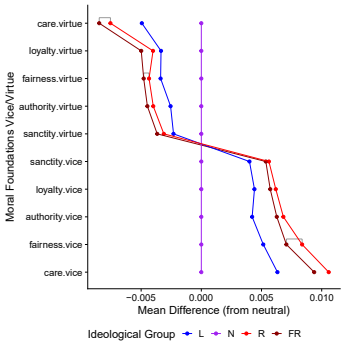

Figure 18: MFT #Socialsense

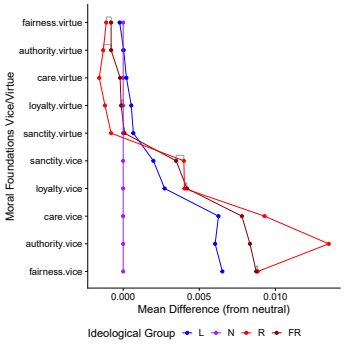

Figure 19: MFT Riot

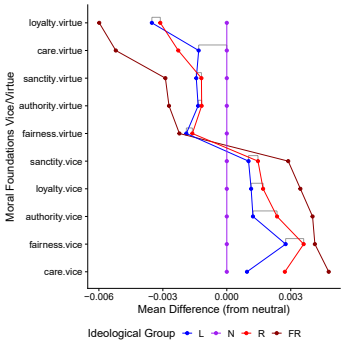

Figure 20: MFT Parler