Beyond Dataset Creation: Critical View of Annotation Variation and Bias Probing of a Dataset for Online Radical Content Detection

Arij Riabi, Virginie Mouilleron, Menel Mahamdi, Wissam Antoun, Djamé Seddah, Djamé Seddah

Beyond Dataset Creation: Critical View of Annotation Variation and Bias Probing of a Dataset for Online Radical Content Detection

Arij Riabi **Virginie Mouilleron** **Menel Mahamdi**
Wissam Antoun **Djamé Seddah**
Inria, Paris
{firstname,lastname}@inria.fr

Abstract

The proliferation of radical content on online platforms poses significant risks, including inciting violence and spreading extremist ideologies. Despite ongoing research, existing datasets and models often fail to address the complexities of multilingual and diverse data. To bridge this gap, we introduce a publicly available multilingual dataset annotated with radicalization levels, calls for action, and named entities in English, French, and Arabic. This dataset is pseudonymized to protect individual privacy while preserving contextual information. Beyond presenting our freely available dataset, we analyze the annotation process, highlighting biases and disagreements among annotators and their implications for model performance. Additionally, we use synthetic data to investigate the influence of socio-demographic traits on annotation patterns and model predictions. Our work offers a comprehensive examination of the challenges and opportunities in building robust datasets for radical content detection, emphasizing the importance of fairness and transparency in model development.

1 Introduction

Given the current debate on the influence of social media and their lack of moderation, making it a giant echo chamber for all kinds of ideologies, it is an understatement to say that the detection of radical content on online platforms has become an increasingly pressing concern. Indeed, radicalization, often driven by online propaganda, has contributed to recent terror attacks and public violence. For example, the United Kingdom experienced a baffling rise in racially motivated attacks[1], whose impact was amplified by the viral spread of many related videos. In this context, online expressions of radicalization pose a unique challenge as they can constitute a rallying point for potentially burgeoning communities and then provide direct access to such communities where extreme opinions can be further intensified (Bowman-Grieve, 2010; Nouh et al., 2019; Chatfield et al., 2015; Stephane Baele and Ging, 2024). Beyond the spread of ideas, online extremism can lead to *offline* dangers, including violent riots, terrorist attacks and so on (Farwell, 2014; Fernandez and Alani, 2021; Pellicani et al., 2023). An important point is to note that the rapid spread of information online, mainly through social media, enables extremist groups to disseminate radical content and recruit others to their cause. However, this is often the first step before these groups migrate to encrypted platforms, evading regulation and oversight. This is why trying to understand the interplay between radicalization process, social network dynamic and human interactions is a crucial challenge. Studies (Flaxman et al., 2016; Bakshy et al., 2015; de Kock, 2024) have explored how exposure to varying ideological perspectives online influences individuals, highlighting the importance of analyzing echo chambers in social media, where the most radicalized and polarized views tend to dominate the discourse (Roy et al., 2021). Few annotated datasets cover radical content from social media (Fernandez and Alani, 2021).

As data quality directly affects model performance and user trust, developing high-quality, consistently labeled data is essential. Our research addresses this challenge by offering a comprehensive analysis of the dataset creation process. We explore how variations in annotations and model training impact radical content detection in NLP. We present COUNTER, a novel pseudo-anonymized dataset created to tackle the complexities of radical content detection across multiple languages—English, French, and Arabic-and ideologies, from far-right to Jihadism. We release the dataset with multiple annotations (NER, `Radical Level`, `Call for Action`), annotators disagreement, all the guide-

[1] https://edition.cnn.com/2024/08/05/uk/uk-far-right-protests-explainer-gbr-intl/index.html

lines, and the synthetic data for bias analysis. We seek to understand the interplay between annotation biases, model generalization, and fairness in detecting radical content. First, we analyze how human label variations affect model outcomes, highlighting how it is heavily dependent on the aggregation method and the evaluation. Second, we conduct an in-depth study of the annotations to identify the most suitable experimental settings to improve model performance. Third, we introduce and evaluate synthetic data as a bias analysis tool, simulating socio-demographic attribute's influence on model predictions. Our results highlight the complexities in detecting radical content, especially given the inherent subjectivity in human annotations and the sociodemographic variations that influence data and model outcomes.

2 Online Radical Content Detection

Radical content can be defined as a signal used by an individual or group of individuals to express a radical perspective in opposition to a political, social, or religious system, and adopting a radical discourse could be followed by a progressive shift in social behavior, resulting in violence and even serious undermining of public safety (Fink, 2014). The definition of radicalization itself is fluid, evolving with the phenomena and associated events, making it difficult for detection algorithms to maintain effectiveness as the associated behaviors and language evolve (Berjawi et al., 2023; Schmid, 2016). This ongoing evolution complicates the definition of radicalization and diminishes the efficiency of detection models as the language and behaviors indicative of radicalization shift over time.

NLP for radical content detection NLP methods show potential for detecting radicalization but require further exploration, as indicated by literature (Mussiraliyeva et al., 2020; De Kock and Hovy, 2024). Analyzing radicalization mechanisms using NLP techniques has been mostly done in a supervised learning setting for different steps like propaganda, recruitment, networking, data manipulation, and disinformation (Hung et al., 2019; Torregrosa et al., 2021; Aldera et al., 2021a). However, existing datasets used for radicalization detection tend to have a narrow focus, often focusing on specific behaviors within particular extremist communities, thus lacking a broader perspective on radicalization across different groups (Hartung et al., 2017; Alatawi et al., 2021). The quality and availability of training and evaluation datasets are significant constraints in radicalization detection, and large datasets often suffer from biases and inadequate quality checks (Gaikwad et al., 2021). Many are gathered using simplistic rules, such as identifying users who employ specific lexicons or share certain content (Lara-Cabrera et al., 2017; Fernandez et al., 2018). These rules often rely on unverified assumptions, introducing noise and reducing dataset quality. Furthermore, when human annotators evaluate data, only a small subset of content is manually verified (Ashcroft et al., 2015; Agarwal and Sureka, 2015; Gaikwad et al., 2023), with annotations often performed through crowdsourcing platforms rather than domain experts, adding additional biases. We found that the literature on radicalization detection in NLP focuses on two primary objectives: detecting online radical content and identifying radicalized users and communities.

Investigating radicalized users allows researchers to detect early-stage radicalization by analyzing behavioral changes over time (El Barachi et al., 2022; Sakketou et al., 2022). However, challenges such as content deletion, account changes, and cross-posting complicate this task. On this topic, De Kock and Hovy (2024) emphasize the lack of research in NLP and propose a semi-supervised solution to bypass a "potentially biased human annotation step." They focus on sociolinguistic indicators like hostility, longevity, and social connectivity, using a lexicon for hostility and radical ideologies (Farrell et al., 2019).

Detecting radical content can also serve as an indicator for identifying potential radicalized users. Especially after linguistic analysis of online communication from high-risk groups revealed the existence of linguistic characteristics that distinguish them from general discourse (Winter et al., 2021; Mueller et al., 2022). Some datasets have been proposed for radical content detection, with the ISIS dataset from Twitter being the most common (Smedt et al., 2018). A significant challenge in building these datasets is defining adequate annotation schemes. Most datasets still treat radicalization as a binary state (Agarwal and Sureka, 2015), which oversimplifies its complex and gradual nature. Some works have attempted to refine this approach by differentiating between hostile and irrelevant content (Ashcroft et al., 2015; Abrar et al., 2019; Kaur et al., 2019) or by categorizing content into propaganda, radicalization, and recruit-

ment (Gaikwad et al., 2023). While most of the research focuses on English with a US-centered perspective, few works focusing on other languages can be found, such as Indonesian (Miranda et al., 2020) and Arabic (Aldera et al., 2021b), particularly in the context of jihadism. As far as we know, no large-scale multilingual works have been conducted in this domain. This is why to fully address the complexity of radical content across diverse contexts, a multilingual dataset with rich annotations from various sources is essential. This can enable the study of radicalization across cultural and linguistic boundaries, providing more nuanced insights into the detection of radical content.

3 COUNTER: Radical Content Dataset

3.1 Data collection

The dataset includes English, French, and Arabic posts from various sources that can be split between social media (Facebook, Twitter), platforms (Telegram), and forums either public, such as Reddit or banned from search engines (4chan) available via special software such as Tor (a set of tools that enables anonymous communication) enabling access to what is often referred as "Darkweb". The contractor carried out the data collection using a list of keywords inspired by relevant geopolitical events. The content posts cover two main ideologies (Jihadism and Far-right), each with different levels of sub-ideologies. Content that cannot be grouped in the previous two categories is put in a third category, which includes posts that do not directly align with Jihadist or Far-right ideologies but still exhibit radical tendencies. The category distribution varies by language, with Far-right dominant in English and French and Jihadism in Arabic. All the meta-data with the posts were kept when available, including the extraction date, post date, and interaction information. Images and video links were also collected (Cf. Appendix A for a detailed data statement (Bender and Friedman, 2018)).

	Arabic	English	French
#sentences	2499	2650	2650
#tokens	168.48K	100.73K	87.58K
Avg length	67.42	38.01	33.06
#NER entities	6579	6651	4884
# anonymized sents	1500	2650	2650
# anonymized entities	130	1615	649

Table 1: Dataset Overview.

3.2 Prescriptive Annotation

Annotating radicalized data requires "experts" in the domain. Annotators must have native-like knowledge of the target language and its linguistic and cultural aspects (Ex, recognizing puns based on specific cultural references like Klan Chowder). Therefore, a contractor with expertise in this task created the first dataset version. This version follows a prescriptive approach to annotation (Rottger et al., 2022), discouraging subjectivity and aiming for consistency by adhering to predefined guidelines. The contractor received the specifications and the needs to produce a dataset to train a detection model for the platform, which is supposed to be used to facilitate online moderation of radicalization content. The annotations are multi-label and multi-class. The main label is `Call for Action`, with five predefined levels based on the degree to which it motivates specific actions, ranging from "negative" to "very high". We also have annotations for `Radicalization Level` with six levels covering "Negative" or "Neutral" content, "Expression of Radical Views", "Using Radical Propaganda", "Associated with Radical Groups", "Dehumanizing the Other" and "Call for Action against others". (See Appendix B.1 for more details).

3.3 Descriptive Annotation

We lacked detailed information about the contractor's annotation process, which is crucial for investigating biases in subjective tasks. Therefore, we added double annotations for `Call for Action` Classification and `Radicalization Level` for English and French[2]. We adopted the descriptive paradigm (Rottger et al., 2022) for those annotations. We gave the annotators the contractor's task description and discussed the task and possible use cases. However, we relied on the lack of details for the annotation to encourage annotators' subjectivity. We wanted the annotations to represent a larger range of beliefs to extract related insights since correlations have been shown between socio-demographic factors and annotations for different tasks such as sentiment analysis (Diaz et al., 2018) and hate speech (Waseem, 2016; Sap et al., 2022). We recruited two trained linguists with different socio-demographic profiles. Both female annotators are between [25-30] and [40-45], have a master's degree as their highest completed education,

[2]Due to the potential psychological impact of annotating radical content, psychological support was made available to annotators to ensure their well-being throughout the process.

are native French speakers, and are fluent in English. Re-annotation guidelines were developed to provide more detailed instructions and accommodate the annotators' interpretations. After defining the guidelines, additional uncertainties were addressed at different times during the process. They double-annotated the French dataset and a large sample of the English dataset. When uncertain, they were encouraged to select the most appropriate class while we tracked those cases.

We report in Table 7 in Appendix B the inter-annotator agreement for `Radicalization Level` and `Call for Action`. The French dataset shows higher agreement, with moderate agreement for `Radicalization Level` at Fleiss' Kappa 0.50 and fair agreement for `Call for Action` at 0.43. In contrast, the English dataset shows lower agreement, with fair agreement for `Radicalization Level` at Fleiss' Kappa 0.26 and slight agreement for `Call for Action` at 0.13.

Pseudononymization and NER annotations were performed simultaneously by the annotators. The main goal was to preserve all semantic properties that can be extracted from the dataset. Our approach ensures the protection of sensitive information without losing critical data, which facilitates sharing the dataset for research purposes. We kept well-known events and public figures non-anonymized to leverage the model's embedded knowledge and maintain alignments within the text. We explain the detailed pipeline in Riabi et al. (2024). We analyzed the effect of the pseudonymization and found that training on both datasets gave comparable results (Appendix C.2).

3.4 Synthetic Data for Bias Analysis

While investigating how socio-demographic traits influenced model decisions, we faced challenges in directly extracting this information from the posts in our dataset, resulting in substantial time and resource constraints. To address this, we adopted the recent trend of using large generative models to create examples with socio-demographic information (Durmus et al., 2024; Aher et al., 2023), which reduces privacy risks and annotations cost (Argyle et al., 2023). This technique, referred to as "persona prompting", is often used to simulate survey participants (Jiang et al., 2022; Simmons and Hare, 2023) or annotators (Lee et al., 2023; Hu and Collier, 2024). The effectiveness of such techniques remains debatable among researchers (Santurkar et al., 2023; Bisbee et al., 2023; Grossmann et al., 2023), but promising results have been demonstrated in several cases (Simmons and Savinov, 2024; Jiang et al., 2024). While many works using generative models use in-context learning to guide the generation of examples (Simmons and Savinov, 2024), it was unsuitable for our case due to the lack of annotated examples with relevant socio-demographic variables, and using in-context learning would have influenced the model with inaccurate or incoherent profiles. Instead, we opted for a zero-shot prompting setup. Our approach is based on creating user profiles that include socio-demographic variables such as age and gender, income, education level, and more (See Appendix E.1 for a complete list of the variables and a prompt example). We tried to include as many protected characteristics as possible [3] as most studies focus on gender and "race"(Sotnikova et al., 2021). These profiles were used to prompt an uncensored LLM called *Wizard-Vicuna-13B-Uncensored* (WizardVicuna, 2023), which was trained on a dataset with its alignment guardrails intentionally removed. The model was created by combining the WizardLM (Xu et al., 2024) approach to training dataset generation with Vicuna's (Chiang et al., 2023) multi-turn conversational data, allowing for more open and flexible interactions. The LLM was used to generate posts annotated for `Radicalization levels` and `Calls to action` by the annotators who had previously re-annotated the original dataset. There was moderate agreement between annotators on a sample of 300 examples, with Cohen's Kappa scores of 0.51 and 0.40 for English and 0.54 and 0.47 for French on `Radicalization Level` and `Call for Action`, respectively. The process involves generating both "base profiles" and "variation profiles" by altering a few variables to maintain consistency while ensuring the authenticity of the generated content. Profiles were also created by taking inspiration from real profiles in the COUNTER dataset. The generated posts generally reflect socio-demographic variables but sometimes rely heavily on stereotypical keywords according to the annotators. The model accurately incorporates most variables but struggles with differentiating based on age and does not always consider the register.

[3] Protected characteristics correspond to attributes of people that anti-discrimination law mentions explicitly Cf. link

4 Results

Experimental setting. We use the MaChAmp v0.2 toolkit (van der Goot et al., 2021), a framework that allows the implementation of different transformers-based tasks and supports single-task and multi-task learning. In the multi-task setting, the encoder is shared between the tasks, which are jointly fine-tuned during training, while we have a different decoder per task. We split our datasets approximately to 70% train, 10% validation, and 20% test for each language with stratification across ideologies, `Call for Action`, and `Radicalization level` (More Details in Appendix C.1). We report the average Macro-F1 over five seeds on the test set and use the validation set to pick the best checkpoint.

Baseline. We fine-tune XLM-T (Barbieri et al., 2022), an XLM-R (Conneau et al., 2020) model, that has been fine-tuned for the MLM task on 200 million tweets in more than 30 languages, which makes it more adapted for social media data. We report in Table 2 the results for our main task, multi-class classification of `Call for Action`. The baseline results show that XLM-T performs reasonably well across all languages, with Macro-F1 scores ranging from 59.41 for Arabic to 65.65 for French.

	en	fr	ar
Baseline	64.63(±2.0)	65.65(±1.8)	59.41(±1.3)
+ Radical level	64.82(±2.5)	63.91(±6.5)	58.98(±2.4
+ Ideology pred	65.10(±1.7)	65.56(±8.6)	57.84(±2.6)
+ NER ID	64.64(±1.3)	61.98(±3.9)	-
+ NER OOD	66.47(±2.5)	63.74(±4.5)	58.37(±1.5)
MULTI-SAME	63.98(±1.9)	60.87(±4.7)	56.85(±2.3)
MULTI-DIFF	**66.65**(±4.3)	**68.77**(±3.2)	**59.48**(±4.3)

Table 2: Macro-F1 of `Call for Action` on the test set for XLM-T.

Does Adding Additional Features Improve the Performance? We examine whether incorporating additional features or tasks could improve model performance. Following Montariol et al. (2022), we perform multi-task training to assess whether these supplementary tasks provided helpful context (cf. Table 2).

Adding **Radicalization Level prediction** as an auxiliary task did not improve performance substantially, with slight variations across languages. In particular, performance in French dropped notably, suggesting that this feature may introduce noise rather than aid in prediction. This observation is coherent with the annotator's observations about the variations across datasets regarding the classification scale as they noted that the tone in the English dataset is far more crude, violent, and derogatory than in the French dataset.

When incorporating **Ideology Prediction** as an auxiliary task, we observed minor improvements in English, while performance in French and Arabic remained relatively stable or declined slightly. This suggests that the benefit of ideology prediction may depend on language-specific features or underlying data distribution.

Next, we added NER both within-domain (ID) (Arabic excluded as only 1500 out of 2500 examples were annotated.) and out-of-domain (OOD). For OOD data, we used CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003) for English, ANERcorp (Obeid et al., 2020) for Arabic, and FTB-NER (Ortiz Suárez et al., 2020) for French. The within-domain NER did not improve results much, with performance even dropping in French. However, NER OOD showed more promise, especially in English, where we observed the most significant improvement, indicating that OOD entities might offer valuable contextual signals for the model.

Finally, the **multilingual** experiments demonstrated that using separate classification layers for each language (MULTI-DIFF) led to the best performance across all languages. This setup outperformed the single-classifier approach (MULTI-SAME), suggesting that handling the nuances of each language separately is more effective than sharing a classifier across all languages.

5 Discussion

We recognize the importance of providing a clear overview of what to expect from models trained on our dataset. In this discussion, we focus on three key aspects. First, we examine how human label variations and annotator disagreements impact model performances. Second, we address fairness concerns, exploring how model predictions may disproportionately affect different demographic groups. Finally, we compare multi-class classification and regression approaches to better capture the nuances of radical content detection.

5.1 How about Human Label Variations?

Annotator disagreement has been observed for different tasks in literature (Jamison and Gurevych, 2015; Larimore et al., 2021; Peng et al., 2024). A

	Test		
Train	Contractor	MACE	Majority
Contractor	**65.65**(±1.8)	50.25(±1.7)	52.57(±1.8)
Mace	53.21(±1.7)	**57.92**(±1.4)	**57.63**(±0.7)
Majority	53.53(±1.4)	55.59(±1.1)	56.73(±2.0)
Repeated-lab	56.46(±1.6)	56.02(±0.8)	56.91(±1.6)
Annot-classifier	58.31(±4.8)	55.12(±1.2)	56.96(±1.3)

Table 3: Macro-F1 results on the test set for human label variation analysis for French set and XLM-T model.

relatively recent line of work called "perspectivist paradigm" (Fleisig et al., 2024) investigates the best way to handle the variability in annotations (Uma et al., 2022; Barrett et al., 2024). This variability can be caused by several reasons, such as task ambiguity, unclear guidelines, annotator expertise, data complexity, and the most challenging reason, subjectivity (Sandri et al., 2023). Therefore, we leverage our multiple annotations to explore how annotator disagreements impact model performance and whether incorporating diverse labels can mitigate biases. We adopt the concept of *Human label variation* as defined by Plank (2022), considering societal biases and interpretative disparities as the primary sources of disagreement in our task.[4] Multi-annotator techniques help measure uncertainty when annotations are inconclusive (Mostafazadeh Davani et al., 2022). This process is critical in the context of radicalization detection, where factors like ethnicity, gender, and age influence how different communities perceive radicalizing content, shaped by their cultural and social backgrounds (Vijayaraghavan et al., 2021; Fleisig et al., 2023). Varying interpretations of radical content may also stem from subjective perceptions, particularly when language includes coded messages or ideological allusions (Lee et al., 2024). Moreover, annotators' thresholds for identifying dangerous material can differ, further influencing their conclusions (Sap et al., 2022). We do not seek the best aggregation method, but we want to show the variability between the approaches and encourage the choice of the adequate option depending on the use of the model. We test four approaches of annotation aggregations:

- Inspired by(Plaza-del Arco et al., 2024), we use **MACE** (Hovy et al., 2013) aggregated labels, a Bayesian annotation tool that calculates two scores: the most likely label and the competence (reliability) of each annotator (i.e., the likelihood that an annotator selects the "true" label based on their expertise rather than speculating on one). MACE operates on unlabeled data and infers both variables using variational Bayesian inference. The aggregated MACE labels are usually more accurate than majority voting.
- **Repeated Labeling**: Treats each annotation as a separate instance. It captures the full range of crowd opinions by treating each label as a separate learning signal.
- **Majority** aggregated the annotations by taking the most frequent label and choosing randomly in case of a tie.
- **Annot-classifier** A single classification head models each annotator. The three predictions are aggregated using a majority vote.

We focus on French for this analysis as we have 3 annotations for all examples. We report the results in Table 3, considering three gold labels: expert's annotations, the majority vote, and MACE.

As expected, the models trained on the corresponding gold standard labels achieved the best results for each test set. Except for the Majority, the best results on majority-labeled test data were from the model trained on MACE aggregation. This highlights MACE's capacity to effectively model variations and capture the underlying consensus among multiple annotators more accurately than majority voting. The drop in performance when models were tested on different gold sets underscores the distinct perspectives each annotation method introduces. For instance, models trained on expert labels showed a notable decrease in performance when tested on MACE and majority labels, suggesting that it offers a more uniform and possibly stricter interpretation than the variability captured by methods like MACE and majority voting.

Cohen's kappa analysis further supports these findings, with high agreement between MACE and majority labels (0.89) and lower agreement between Contractor and MACE (0.70). These results underscore the importance of choosing the appropriate label aggregation method, depending on whether the model needs to align more closely with expert consensus or capture broader interpretations.

[4]We note that annotator disagreements may also arise from annotation errors or plausible variations, as discussed in (Weber-Genzel et al., 2024), but we leave the distinction between these cases to future work.

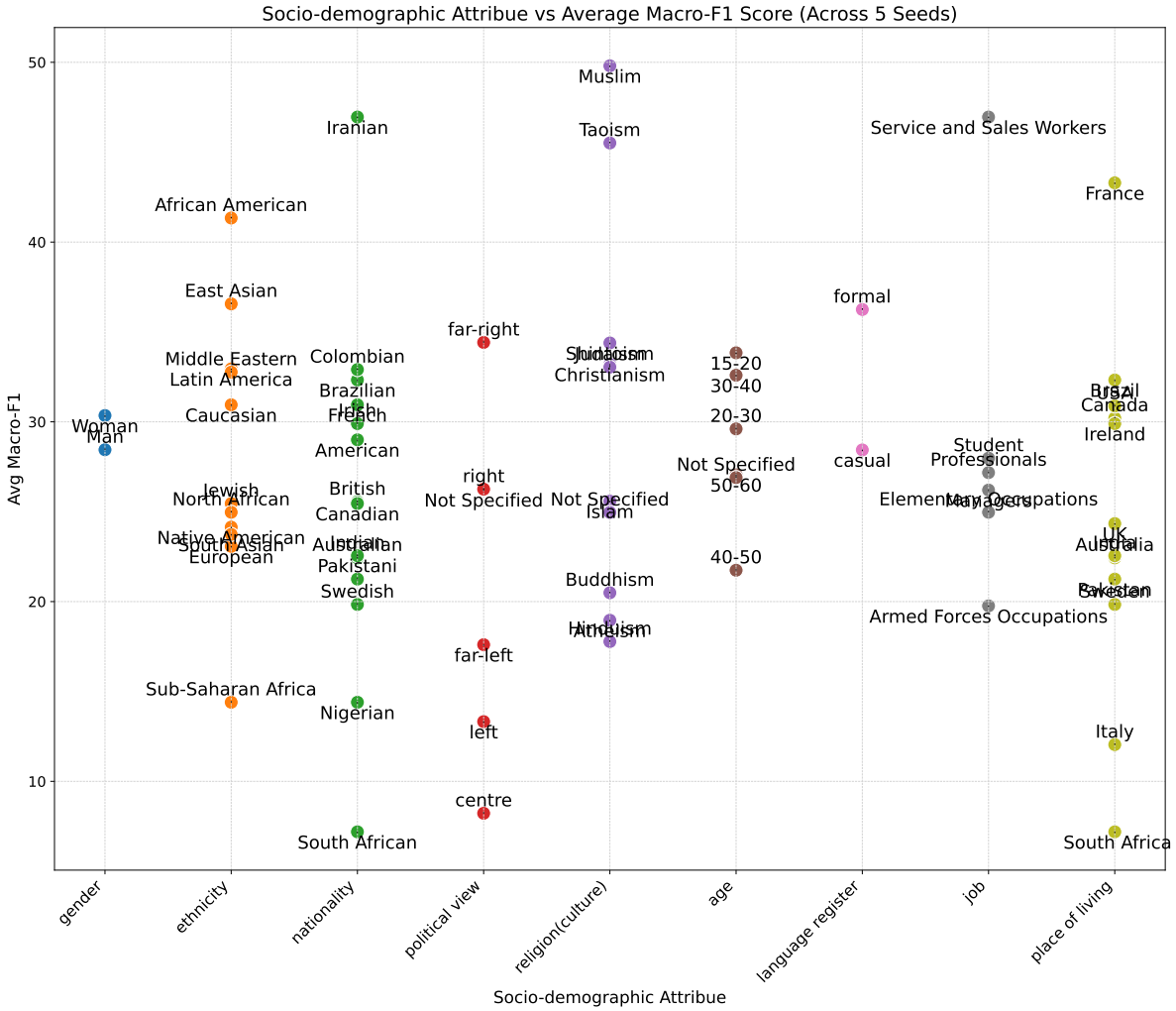


Figure 1: Average Macro-F1 variations for the attributes for XLM-T model for the synthetic English set.

5.2 Bias Analysis: What can we infer?

We compared the average Macro-F1 score per value for each socio-demographic attribute using our generated dataset to assess model bias. In addition to XLM-T we also trained two other models, XLM-R and MBERT, on both French and English datasets. The results of these models are reported in Appendix C.3, showing that XLM-T outperforms both XLM-R and MBERT in the English and French test sets. As expected, performance in synthetic data was lower than in the standard test set due to differences in distribution between the training data and the generated dataset. The purpose of this experiment was not to maximize performance, but to diagnose any systematic errors correlated with the attributes. Figure1 shows the plot of average Macro-F1 scores across categories for each attribute for XLM-T (we report the plots for MBERT and XLM-R in the appendix).

Sociodemographic biases impact performance We observed significant differences across attributes such as nationality, ethnicity, political views, and religion. For instance, all models displayed **substantial performance variation across political views and ethnicities, indicating a potential bias in how certain groups are represented or classified**. While XLM-T showed the highest overall performance, it exhibited larger disparities between categories than XLM-R and MBERT. In particular, XLM-T had more pronounced political views and nationality variations, while XLM-R and MBERT displayed slightly more balanced results but lower overall performance. **This suggests that while** XLM-T **captures certain patterns better, it may amplify biases in specific demographic categories.** This is likely due to XLM-T being trained on social media data such as tweets, which tend to reflect more polarized and informal language, compared to XLM-R and

MBERT, which were trained on broader corpora like Wikipedia and books, containing more formal and neutral text(See Figure 7 in Appendix E for the results of all models on the synthetic dataset).

Interestingly, for the French data, the differences between the three models XLM-T, XLM-R and MBERT are less pronounced than in English, with more consistent performance across socio-demographic attributes. However, the distribution of the gender attribute [5] shows a larger variation in French results compared to what we observed for English. This could be attributed to the fact that detecting the gender of a user from French texts is generally easier than from English, likely due to many gender-sensitive morpho-syntactic markers in French (Cf. Figure 8 and Figure 9 for the other plots).

	Demographic Parity		Equalized odds	
Attribute	en	fr	en	fr
Place of living	0.38	0.55	0.61	0.62
Ethnicity	0.46	0.37	0.68	0.35
Religion	0.32	0.24	0.57	0.31
Political view	0.24	0.21	0.23	0.24
Age	0.21	0.48	0.23	0.56
Gender	0.06	0.17	0.05	0.20
Job	0.24	0.60	0.41	0.62
Nationality	0.45	0.27	0.66	0.32

Table 4: Demographic Parity Difference and Equalized Odds Difference for XLM-T on synthetic data.

Challenges in assessing bias Metrics to evaluate bias in the dataset or model decisions are very challenging as they depend on the used aggregation and the task definition (Olteanu et al., 2019) "The choice of metrics shapes a research study takeaways". Most metrics quantify the extent to which an algorithm treats people differently and the extent to which an algorithm impacts different people differently. The metrics that are commonly applied assume that there are two outcomes: a favorable one and an unfavorable one. We have a multi-class classification, so we considered class zero as the positive outcome. We use the **demographic parity difference**, also called disparate impact; it measures the ratio of favorable outcomes between different groups to assess whether the model treats different groups equally. A demographic parity difference of 0 means that all groups have the same selection rate, which refers to the proportion of individuals in each group who receive a positive outcome. We also use **equalized odds difference**, which seeks that the predictions made by the model have equal true positive and false positive rates, regardless of the membership in sensitive groups (Cf. Section D in the Appendix for detailed definitions of both metrics).[6]

Results in Table 4 show that **overall Demographic Parity is generally smaller than Equalized Odds**, indicating that while the model may be relatively fair in terms of selection rates, it struggles more with ensuring consistency in True Positive Rates and False Positive Rates across different groups. The largest disparities in Demographic Parity are observed for attributes like *Place of Living* and *Job*, particularly in the French dataset, where the selection rates across groups differ more significantly than in the English dataset. For Equalized Odds, substantial disparities are observed again for *Place of Living*, *Ethnicity*, and *Nationality*, particularly in English.

Our results imply that the model's predictive performance is less balanced across these groups, with larger differences in accuracy and error rates. Interestingly, *Political View* and *Gender* show the smallest differences in both metrics, suggesting more consistent treatment for these attributes.

Language-wise, the model generally shows more fairness challenges in French, especially for demographic categories like *Place of Living* and *Job*, with larger disparities compared to English.

5.3 Multi-Class Classification or Regression?

We aim to investigate the complexity of detecting `Call for Action`, focusing on whether multi-class classification or regression yields better performance and evaluation reliability. We trained a regression model using Mean Absolute Error, which measures the average distance between the predicted values and the true labels. To compare this regression approach directly with a classification model, we rounded the regression predictions to the nearest integer, allowing us to calculate the F1 score for both models. Although our classes are **discrete**, ranging from 0 to 4, they can also be viewed as a continuous spectrum where **errors between adjacent classes** are **less severe than between more distant classes**. This suggests that a regression approach, which inherently **accounts**

[5] The data lacks fine-grained classifications of gender identities, focusing only on male and female categories. We do not consider this classification to fully represent gender identities and use it solely for analysis within this specific subset.

[6] We use the Fairlearn library to compute these two metrics

for such gradations, might be more suitable. However, classification allows the model to learn **more complex and non-linear patterns** specific to each class, which could capture subtle, class-specific nuances that a regression model might oversimplify. For example, the boundaries between classes 2 and 3 could involve different features or relationships than those between classes 1 and 2, which a classification model is better equipped to learn.

Lang	Classification		Regression	
	Macro-F1	Spearman	Macro-F1	Spearman
en	**64.63**(±2.0)	0.75(±0.03)	56.21(±1.5)	**0.78**(±0.01)
fr	**65.65**(±1.8)	0.71(±0.03)	53.15(±3.8)	**0.73**(±0.02)
ar	**59.41**(±1.3)	0.58(±0.02)	49.46(±1.8)	**0.63**(±0.02)

Table 5: XLM-T's Macro-F1 results trained for `Call for Action` as multi-class classification and regression.

We calculated the Spearman correlation coefficient for both models. Spearman correlation is particularly relevant here as it measures the strength and direction of the monotonic relationship between the predicted and actual class rankings. This metric is well-suited for our task because it evaluates how well the models preserve the ordinal nature of the classes, regardless of the predicted values. The results in table 5 show a trade-off between classification and regression models. The classification models achieve higher Macro-F1 scores for all three languages than regression, indicating better accuracy in discrete class prediction. However, regression models slightly outperform in preserving the ordinal structure, as evidenced by higher Spearman correlations, with the biggest improvement for Arabic. The low standard deviations suggest consistent performance across seeds for regression. The confusion matrices shown in Figure 2 provide deeper insights into the performance differences between the regression and classification models, highlighting the importance of error severity for radical online content. In all the languages, the regression model consistently misclassifies only between adjacent classes. This sensitivity to the ordinal relationships is a distinct advantage, as errors between adjacent classes are less severe than between distant ones. However, the model struggles with class separation due to the dominance of certain classes, especially in the French dataset, where class 1 is prevalent. While more accurate overall, the classification model occasionally makes more critical errors by misclassifying non-adjacent classes. All models face challenges with higher

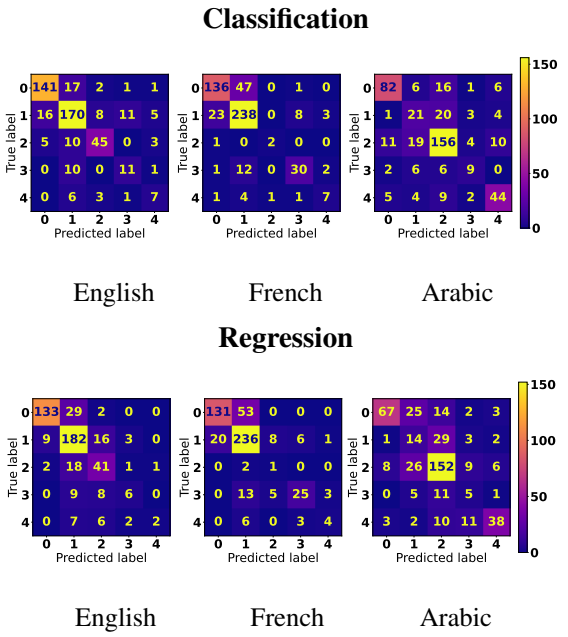


Figure 2: Confusion Matrix for `Call for Action`.

classes, but the regression model maintains better consistency, typically misclassifying into adjacent classes rather than across distant ones.

6 Conclusion

In this paper, we presented COUNTER, a novel, multilingual, pseudo-anonymized dataset for detecting online radical content, accompanied by an in-depth analysis of the dataset's biases and the impact of human label variations. We conducted extensive experiments to evaluate the performance of various models, considering factors such as radicalization levels, calls for action, and named entities. Our results highlight the complexities in detecting radical content, especially given the inherent subjectivity in human annotations and the sociodemographic variations that influence data and model outcomes. Furthermore, we investigated the use of synthetic data to explore biases related to sociodemographic traits, showing the potential of generative models to simulate realistic annotations. Regarding the classification challenges we faced, we plan to refine our approach by exploring techniques for handling unbalanced datasets, like weighted cross-entropy, and by incorporating multiple dimensions of radicalization behavior from our metadata to improve classification accuracy. Finally, our findings underscore the importance of understanding and accounting for bias in models and datasets, especially for sensitive tasks like radicalization detection. The COUNTER dataset is freely available for research.

Limitations

Even though our dataset has many advantages, certain drawbacks must be noted. First, there is a chance that the models and dataset contain underlying biases from the underlying data, which could lead to stereotypes or overrepresentation of particular groups, which could compromise the predictability of the results. The synthetic nature of the generated data presents further difficulties, since it might not accurately represent the complexity and diversity of radical content in the actual world, which restricts the applicability of the models in larger, multicultural, or international situations. Furthermore, the vocabulary and actions linked to radicalization change quickly. Over time, new slang or coded language not represented in the existing dataset may appear, diminishing the model's applicability.

As underlined by (Fernandez and Alani, 2021; De Kock and Hovy, 2024), there is still far too little cooperation between social and humanist researchers on one side and NLP and machine learning researchers on the other side, preventing the latter from benefiting from the theories, studies, and insights on radicalization from the former. Combining insights from different disciplines can help improve the models to include more features for the detection of radical content.

Ethics Statement

The development and use of machine learning models for radicalization detection raise necessary ethical concerns we have considered throughout our research. Given the ongoing discussions around the moral and ethical alignment of large language models (LLMs) (Liu et al., 2022), we remain cautious in our approach, particularly regarding using such models for subjective or sensitive tasks. LLMs may inadvertently propagate biases or reduce the richness of human judgment in contexts that require nuanced understanding, especially in areas as complex as radicalization detection. The risk of biased annotations and stereotypical outputs reinforces the need for a more thoughtful and transparent deployment of these models.

A motivation for our work is the ability to monitor discussions and identify at-risk users in online extremist communities. However, we recognize that this technology could conceivably be misused to profile individuals or preemptively prosecute them based on incomplete or inaccurate predictions. Since our evaluation demonstrates that the predictive models are not perfectly accurate, such actions would constitute a gross misuse of the technology. To mitigate this risk, we release our dataset only to researchers upon demand. Instead, we believe these models are best used as part of broader intelligence-gathering systems, providing context rather than determinative judgments, as discussed by Winter et al. (2021). Human oversight must complement any use of these technologies and be guided by stringent ethical standards to prevent abuse.

In this context, we also acknowledge the significant policy challenges and legal dilemmas highlighted by scholars like Jarvis et al. (2015), especially as governments wrestle with the need to counter terrorism while respecting individual rights and freedoms. Using algorithmic tools in sensitive areas such as policing and security has historically posed privacy risks and led to adverse social externalities (Byrne and Marx, 2011), including concerns over liberty and integrity. Research further reveals that individuals tend to perceive algorithms' decisions as less fair than those made by humans, which could erode public trust in automated systems (Hobson et al., 2021).

Moreover, we recognize that our work involves social datasets representing real people or groups, bringing it into human subjects research (Varshney, 2015). The ethical risks include potential privacy breaches or the reinforcement of harmful profiling based on race, socioeconomic status, or gender. We have taken care to minimize these risks by adhering to best practices in data handling and ensuring that our dataset respects the privacy and dignity of individuals.

Note that the whole annotation process was particularly challenging for our annotators due to the violent, if not borderline traumatizing in some cases, nature of the data, which had an impact on their psychological well-being. The team was provided with a mental health professional service and support from human resources services. A process dedicated to evaluating the psychological impact induced by annotating this content was put in place. Its results (through extensive surveys—similar in depth to PTSD evaluation forms—and debriefing interviews) are currently under evaluation at our institution.

In light of these considerations, we emphasize the importance of transparency, accountability, and the continuous scrutiny of our methodologies. Fu-

ture work in this domain must ensure that the deployment of these technologies is guided by rigorous ethical standards, striking a balance between the imperative to counter radicalization and the protection of individual freedoms.

Acknowledgments

We warmly thank Marine Carpuat for her feedback on this paper.

This work received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 101021607. The authors warmly thank the OPAL infrastructure from Université Côte d'Azur for providing resources and support.

References

Mohammad Fahim Abrar, Mohammad Shamsul Arefin, and Md Sabir Hossain. 2019. A framework for analyzing real-time tweets to detect terrorist activities. In *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pages 1–6. IEEE.

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69. PMLR.

Swati Agarwal and Ashish Sureka. 2015. Using knn and svm based one-class classifier for detecting online radicalization on twitter. In *International Conference on Distributed Computing and Internet Technology*, pages 431–442. Springer.

Gati V Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 337–371. PMLR.

Hind S. Alatawi, Areej M. Alhothali, and Kawthar M. Moria. 2021. Detecting white supremacist hate speech using domain specific word embedding with deep learning and bert. *IEEE Access*, 9:106363–106374.

S. Aldera, Ahmad Emam, Muhammad Al-Qurishi, Majed Alrubaian, and Abdulrahman Alothaim. 2021a. Online extremism detection in textual content: A systematic literature review. *IEEE Access*, 9:42384–42396.

Saja Aldera, Ahmed Emam, Muhammad Al-Qurishi, Majed Alrubaian, and Abdulrahman Alothaim. 2021b. Exploratory data analysis and classification of a new arabic online extremism dataset. *IEEE Access*, 9:161613–161626.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Michael Ashcroft, Ali Fisher, Lisa Kaati, Enghin Omer, and Nico Prucha. 2015. Detecting jihadist messages on twitter. In *2015 European intelligence and security informatics conference*, pages 161–164. IEEE.

Eytan Bakshy, Solomon Messing, and Lada A. Adamic. 2015. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Maria Barrett, Max Müller-Eberstein, Elisa Bassignana, Amalie Brogaard Pauli, Mike Zhang, and Rob van der Goot. 2024. Can humans identify domains? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2745–2765, Torino, Italia. ELRA and ICCL.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Omran Berjawi, Giuseppe Fenza, and Vincenzo Loia. 2023. A comprehensive survey of detection and prevention approaches for online radicalization: Identifying gaps and future directions. *IEEE Access*, 11:120463–120491.

James Bisbee, Joshua Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer Larson. 2023. Synthetic replacements for human survey data? the perils of large language models.

Lorraine Bowman-Grieve. 2010. The internet and terrorism: pathways towards terrorism and counterterrorism. -.

James Byrne and Gary Marx. 2011. Technological innovations in crime. *Cahiers Politiestudies Jaargang*, pages 17–40.

Akemi Takeoka Chatfield, Christopher G. Reddick, and Uuf Brajawidagda. 2015. Tweeting propaganda, radicalization and recruitment: Islamic state supporters multi-sided twitter networks. In *Proceedings of the 16th Annual International Conference on Digital Government Research*, dg.o '15, page 239–249,

New York, NY, USA. Association for Computing Machinery.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Christine de Kock. 2024. Jointly modelling the evolution of community structure and language in online extremist groups. *Preprint*, arXiv:2409.19243.

Christine De Kock and Eduard Hovy. 2024. Investigating radicalisation indicators in online extremist communities. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 1–12, Mexico City, Mexico. Association for Computational Linguistics.

Mark Diaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–14, New York, NY, USA. Association for Computing Machinery.

Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards measuring the representation of subjective global opinions in language models. *Preprint*, arXiv:2306.16388.

May El Barachi, Sujith Mathew, Farhad Oroumchian, Imene Ajala, Saad Lutfi, and Rand Yasin. 2022. Leveraging natural language processing to analyse the temporal behavior of extremists on social media. *Journal of Communications Software and Systems*, 18:195–207.

Tracie Farrell, Miriam Fernandez, Jakub Novotny, and Harith Alani. 2019. Exploring misogyny across the manosphere in reddit. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '19, page 87–96, New York, NY, USA. Association for Computing Machinery.

James P. Farwell. 2014. The media strategy of isis. *Survival*, 56(6):49–55.

Miriam Fernandez and Harith Alani. 2021. Artificial intelligence and online extremism: Challenges and opportunities. -.

Miriam Fernandez, Moizzah Asif, and Harith Alani. 2018. Understanding the roots of radicalisation on twitter. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '18, page 1–10, New York, NY, USA. Association for Computing Machinery.

Louis Fink. 2014. Understanding radicalisation and dynamics of terrorist networks through political-psychology. *International Institute for Counter-terrorism*.

Seth Flaxman, Sharad Goel, and Justin M. Rao. 2016. Filter Bubbles, Echo Chambers, and Online News Consumption. *Public Opinion Quarterly*, 80(S1):298–320.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore. Association for Computational Linguistics.

Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. 2024. The perspectivist paradigm shift: Assumptions and challenges of capturing human labels. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2279–2292, Mexico City, Mexico. Association for Computational Linguistics.

M. Gaikwad, Swati Ahirrao, Shraddha Phansalkar, and K. Kotecha. 2021. Online extremism detection: A systematic literature review with emphasis on datasets, classification techniques, validation methods, and tools. *IEEE Access*, 9:48364–48404.

Mayur Gaikwad, Swati Ahirrao, Shraddha Phansalkar, Ketan Kotecha, Shalli Rani, and Lorenzo Putzu. 2023. Multi-ideology, multiclass online extremism dataset, and its evaluation using machine learning. *Intell. Neuroscience*, 2023.

Igor Grossmann, Matthew Feinberg, Dawn C. Parker, Nicholas A. Christakis, Philip E. Tetlock, and William A. Cunningham. 2023. Ai and the transformation of social science research. *Science*, 380(6650):1108–1109.

Matthias Hartung, Roman Klinger, Franziska Schmidtke, and Lars Vogel. 2017. Ranking right-wing extremist social media profiles by similarity to democratic and extremist groups. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 24–33, Copenhagen, Denmark. Association for Computational Linguistics.

Zoë Hobson, Julia Yesberg, Ben Bradford, and Jonathan Jackson. 2021. Artificial fairness? trust in algorithmic police decision-making. *Journal of Experimental Criminology*, 19.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307, Bangkok, Thailand. Association for Computational Linguistics.

Benjamin W.K. Hung, Shashika R. Muramudalige, Anura P. Jayasumana, Jytte Klausen, Rosanne Libretti, Evan Moloney, and Priyanka Renugopalakrishnan. 2019. Recognizing radicalization indicators in text documents using human-in-the-loop information extraction and nlp techniques. In *2019 IEEE International Symposium on Technologies for Homeland Security (HST)*, pages 1–7.

Emily Jamison and Iryna Gurevych. 2015. Noise or additional information? leveraging crowdsource annotation item agreement for natural language tasks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 291–297, Lisbon, Portugal. Association for Computational Linguistics.

L. Jarvis, S. MacDonald, and T.M. Chen. 2015. *Terrorism Online: Politics, Law and Technology*. Routledge Studies in Conflict, Security and Technology. Taylor & Francis.

Hang Jiang, Doug Beeferman, Brandon Roy, and Deb Roy. 2022. CommunityLM: Probing partisan worldviews from language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6818–6826, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627, Mexico City, Mexico. Association for Computational Linguistics.

Armaan Kaur, Jaspal Kaur Saini, and Divya Bansal. 2019. Detecting radical text over online media using deep learning. *Preprint*, arXiv:1907.12368.

Raúl Lara-Cabrera, Antonio González Pardo, Karim Benouaret, Noura Faci, Djamal Benslimane, and David Camacho. 2017. Measuring the radicalisation risk in social networks. *IEEE Access*, 5:10892–10900.

Savannah Larimore, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. 2021. Reconsidering annotator disagreement about racist language: Noise or signal? In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 81–90, Online. Association for Computational Linguistics.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024. Exploring cross-cultural differences in English hate speech annotations: From dataset construction to analysis. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4205–4224, Mexico City, Mexico. Association for Computational Linguistics.

Noah Lee, Na Min An, and James Thorne. 2023. Can large language models capture dissenting human voices? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4585, Singapore. Association for Computational Linguistics.

Ruibo Liu, Ge Zhang, Xinyu Feng, and Soroush Vosoughi. 2022. Aligning generative language models with human values. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 241–252, Seattle, United States. Association for Computational Linguistics.

Eka Miranda, Mediana Aryuni, Yudi Fernando, and Tia Mariatul Kibtiah. 2020. A study of radicalism contents detection in twitter: Insights from support vector machine technique. In *2020 International Conference on Information Management and Technology (ICIMTech)*, pages 549–554.

Syrielle Montariol, Arij Riabi, and Djamé Seddah. 2022. Multilingual auxiliary tasks training: Bridging the gap between languages for zero-shot transfer of hate speech detection models. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 347–363, Online only. Association for Computational Linguistics.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Pia Mueller, Stefan Harrendorf, and Antonia Mischler. 2022. Linguistic radicalisation of right-wing and salafi jihadist groups in social media: a corpus-driven lexicometric analysis. *CrimRxiv*.

Shynar Mussiraliyeva, Milana Bolatbek, Batyrkhan Omarov, Zhanar Medetbek, Gulshat Baispay, and Ruslan Ospanov. 2020. On detecting online radicalization and extremism using natural language processing. In *2020 21st International Arab Conference on Information Technology (ACIT)*, pages 1–5.

Mariam Nouh, Jason R. C. Nurse, and M. Goldsmith. 2019. Understanding the radical mind: Identifying signals to detect extremist content on twitter. *2019 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 98–103.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2.

Pedro Javier Ortiz Suárez, Yoann Dupont, Benjamin Muller, Laurent Romary, and Benoît Sagot. 2020. Establishing a new state-of-the-art for French named entity recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4631–4638, Marseille, France. European Language Resources Association.

Antonio Pellicani, Gianvito Pio, Domenico Redavid, and Michelangelo Ceci. 2023. Sairus: Spatially-aware identification of risky users in social networks. *Information Fusion*, 92:435–449.

Siyao Peng, Zihang Sun, Sebastian Loftus, and Barbara Plank. 2024. Different tastes of entities: Investigating human label variation in named entity annotations. In *Proceedings of the Third Workshop on Understanding Implicit and Underspecified Language*, pages 73–81, Malta. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Flor Miriam Plaza-del Arco, Debora Nozza, and Dirk Hovy. 2024. Wisdom of instruction-tuned language model crowds. exploring model label variation. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 19–30, Torino, Italia. ELRA and ICCL.

Arij Riabi, Menel Mahamdi, Virginie Mouilleron, and Djamé Seddah. 2024. Cloaked classifiers: Pseudonymization strategies on sensitive classification tasks. In *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*, pages 123–136, Bangkok, Thailand. Association for Computational Linguistics.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Shamik Roy, Maria Leonor Pacheco, and Dan Goldwasser. 2021. Identifying morality frames in political tweets using relational learning. *arXiv preprint arXiv:2109.04535*.

Flora Sakketou, Allison Lahnala, Liane Vogel, and Lucie Flek. 2022. Investigating user radicalization: A novel dataset for identifying fine-grained temporal shifts in opinion. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3798–3808, Marseille, France. European Language Resources Association.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia. Association for Computational Linguistics.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Alex P. Schmid. 2016. Research on radicalisation: Topics and themes. *Perspectives on Terrorism*, 10(3):26–32.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. In *Machine Learning and Knowledge Discovery in Databases*, pages 145–158, Berlin, Heidelberg. Springer Berlin Heidelberg.

Gabriel Simmons and Christopher Hare. 2023. Large language models as subpopulation representative models: A review. *Preprint*, arXiv:2310.17888.

Gabriel Simmons and Vladislav Savinov. 2024. Assessing generalization for subpopulation representative modeling via in-context learning. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 18–35, St. Julians, Malta. Association for Computational Linguistics.

Tom De Smedt, Guy De Pauw, and Pieter Van Ostaeyen. 2018. Automatic detection of online jihadist hate speech. *Preprint*, arXiv:1803.04596.

Anna Sotnikova, Yang Trista Cao, Hal Daumé III, and Rachel Rudinger. 2021. Analyzing stereotypes in generative text inference tasks. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4052–4065, Online. Association for Computational Linguistics.

Lewys Brace Stephane Baele and Debbie Ging. 2024. A diachronic cross-platforms analysis of violent extremist language in the incel online ecosystem. *Terrorism and Political Violence*, 36(3):382–405.

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

Javier Torregrosa, Gema Bello Orgaz, Eugenio Martínez Cámara, Javier Del Ser, and David Camacho. 2021. A survey on extremism analysis using natural language processing. *CoRR*, abs/2104.04069.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2022. Learning from disagreement: A survey. *J. Artif. Int. Res.*, 72:1385–1470.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Kush R. Varshney. 2015. Data science of the people , for the people , by the people : A viewpoint on an emerging dichotomy. -.

Prashanth Vijayaraghavan, Hugo Larochelle, and Deb Roy. 2021. Interpretable multi-modal hate speech detection. *arXiv preprint arXiv:2103.01616*.

Zeerak Waseem. 2016. Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Leon Weber-Genzel, Siyao Peng, Marie-Catherine De Marneffe, and Barbara Plank. 2024. VariErr NLI: Separating annotation error from human label variation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269, Bangkok, Thailand. Association for Computational Linguistics.

Charlie Winter, Peter Neumann, Alexander Meleagrou-Hitchens, Magnus Ranstorp, Lorenzo Vidino, and Johanna Fürst. 2021. Online extremism: Research trends in internet activism, radicalization, and counter-strategies. *International Journal of Conflict and Violence*, 14:1–20.

Cognitive Computations WizardVicuna. 2023. Wizard-vicuna-13b-uncensored model on hugging face. `https://huggingface.co/cognitivecomputations/Wizard-Vicuna-13B-Uncensored`.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024. WizardLM: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations*.

A Data Statement

Following (Bender and Friedman, 2018) , we provide a data statement for COUNTER dataset.

A.1 CURATION RATIONALE

The dataset was created to improve the detection of radical content online. It explicitly targets various levels of radicalization across ideologies like Jihadism and Far-right extremism, with an additional category for unclassified ideologies. The main motivation was to build a representative, multilingual dataset that can improve NLP models' ability to detect extremist discourse.

Data was sourced from Reddit, Twitter, Facebook, and encrypted channels like Telegram and 4chan (via Tor). However, platforms like Facebook and Telegram posed challenges as only public groups or channels are searchable through public APIs, leaving a significant portion of content unreachable. The distribution of the data sources for each language is shown in Figure 3.

Sampling was guided by a lexicon of keywords associated with radicalization, covering terms related to extremist ideologies and violence incitement. Data was collected across two main time frames. Efforts were made to ensure the data represents diverse languages and platforms, including English, French, and Arabic, to reflect the varied discourse of extremist communities.

A.2 LANGUAGE VARIETY

The dataset includes posts in three languages: English, French, and Arabic, each reflecting distinct linguistic features and registers. The Arabic data comprises a mix of Modern Standard Arabic (MSA) and various dialectal forms. The French content is primarily from France. English data is from various locations, representing the specific contexts for radicalization. Different registers are represented within each language, from formal statements to informal, conversational speech, depending on the platform and context of the content.

A.3 SPEAKER DEMOGRAPHICS

We have limited information concerning the demographics of the speakers in this dataset. Approximately 90% of the Geo-Location data for English and Arabic is unknown. Geo-location information is unavailable for French data. Obtaining accurate and complete demographic information, such as the location or nationality of users, is inherently difficult due to the nature of the platforms used (social media, encrypted channels) and the widespread use of pseudonyms or anonymous accounts. Moreover, many platforms restrict access to user metadata, further complicating efforts to gather speaker-related information.

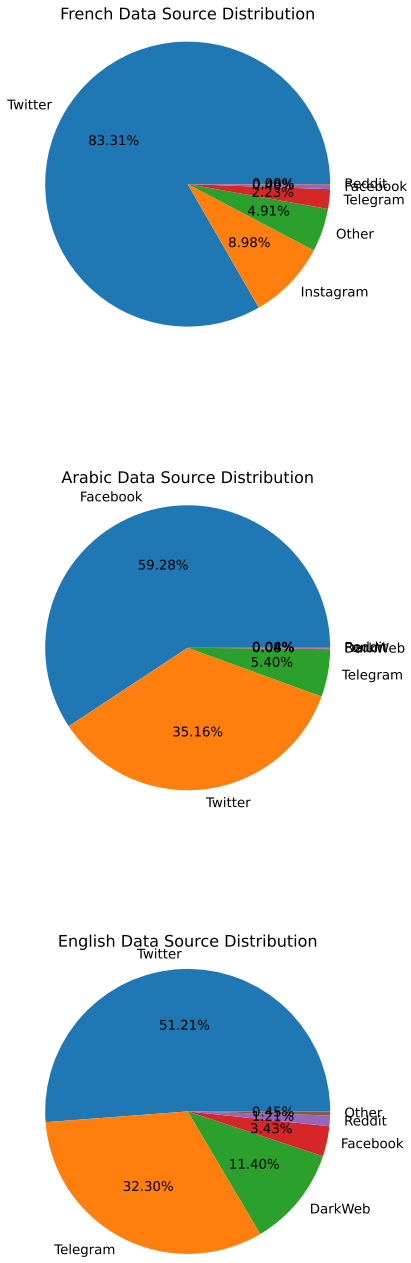


Figure 3: Data source distributions for English, French, and Arabic

A.4 ANNOTATOR RECRUITMENT

Both annotators were recruited as research engineers with prior experience working on annotation projects and possessing relevant contextual knowledge of the languages involved. One annotator holds a degree in English and has lived in the UK for a significant period. Both annotators were familiar with the socio-cultural aspects of radical content.

The annotations cost 72k€ (12 person months), not counting the supervision time and our institution's overhead fee structures. Given the sensitive and potentially distressing nature of the radical content, psychological assistance was offered to the annotators throughout the project.

For the contractor annotations, we have limited information available.

A.5 ANNOTATOR DEMOGRAPHICS

	ANNOT1	ANNOT2
Age	25-30	40-45
Gender	Female	
Ethnicity	North African	European
Native language	French	
Socioeconomic status	Research engineer	
Religion	Practising Muslim	Catholic
Political View	-	Left
Training in linguistics	Master	Master

Table 6: Annotator demographics for ANNOT1 and ANNOT2

A.6 SPEECH SITUATION

The posts in our dataset were posted over a time spanning from 17/07/2015 and 03/04/2023, with collection conducted between 24/07/2022 and 03/04/2023. Although the exact geographical locations of the users are not available, the languages represented (English, French, and Arabic) suggest a broad distribution across different regions. The data consists entirely of written text, as it originates from social media platforms and forums, and is mainly spontaneous and user-generated without prior scripting or editing. The interaction is asynchronous, as the posts were made at different times without real-time communication between users. These posts were intended for a public or semi-public audience, targeting other users on social media or forums, with the potential to reach diverse individuals depending on the platform and language used.

A.7 TEXT CHARACTERISTICS

The dataset contains posts of varying lengths across the three languages: English, French, and Arabic. We provide distributions of the post lengths for each language in Figure 4. There is a correlation between sentence length and the data sources, with longer posts typically originating from forums and shorter posts from social media platforms like Twitter. After scraping, the text underwent minimal cleaning to preserve as much of the original content as possible. Only basic preprocessing steps were performed, such as removing irrelevant metadata and empty posts. A pseudonymization process was also applied to ensure privacy, as detailed in the annotation guidelines.

A.8 LICENSE

The full data set is available for research requests only. Annotations examples available on the dataset gitlab's repo can be freely released (CC-BY-NC-SA)

B Radical content Dataset

Annotators	`Radical Level`	`Call for Action`
French (2624 examples)		
Annot1 vs Annot2	0.58	0.42
Annot1 vs Contractor	0.43	0.35
Annot2 vs Contractor	0.51	0.63
Fleiss' Kappa	0.50	0.43
English (921 examples)		
Annot1 vs Annot2	0.33	0.17
Annot1 vs Contractor	0.28	0.17
Annot2 vs Contractor	0.23	0.17
Fleiss' Kappa	0.26	0.13

Table 7: Cohen's Kappa and Fleiss' Kappa Annotator Agreement for Radicalization Level and `Call for Action`.

B.1 Radicalization Level classes description

- Negative (No Radicalization): This would include content with no signs of promoting, endorsing, or expressing extremist ideologies or actions. Posts under this category are neutral or express tolerance, diversity, peace, etc.
- Expression of Radical Views: Content that expresses views aligned with extremist ideologies but does not actively advocate for violence or involvement. It might be sharing a radical viewpoint, belief, or stereotype or indicating approval of radical actions/actors

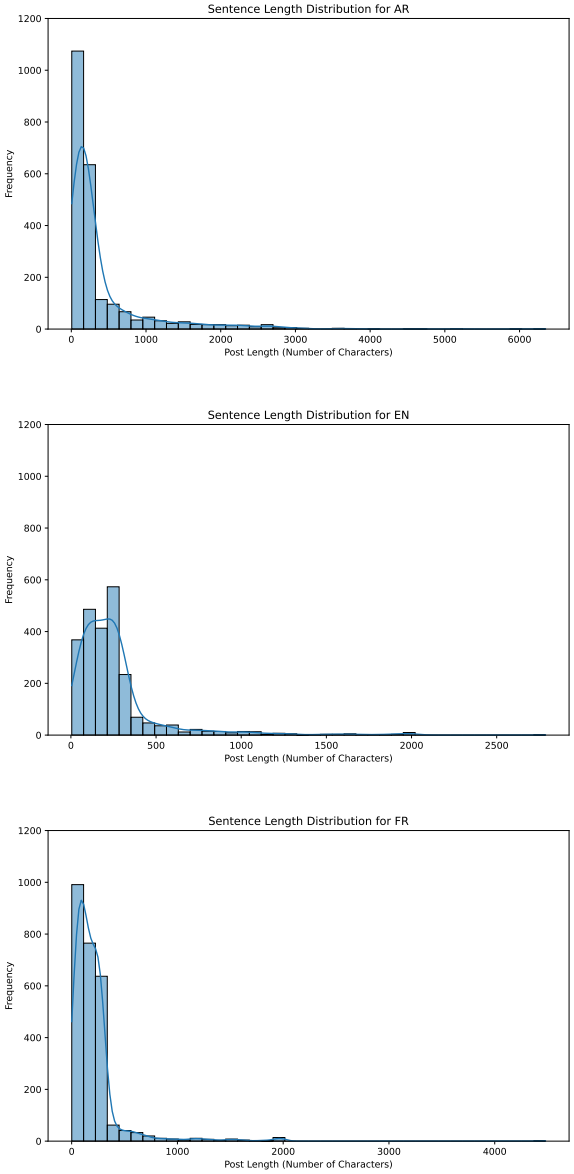


Figure 4: Data length distributions for English, French, and Arabic

without actively supporting or calling for such actions.

- Using Radical Propaganda: The content uses or shares established extremist propaganda. This could include sharing extremist images, slogans, videos, speeches, conspiracy theories, misinformation, or disinformation designed to promote a radical ideology or group.
- Associated with Radical Groups: Content that signifies association or affiliation with known radical or extremist groups. This could be through sharing group propaganda, expressing support or admiration for the group, claiming membership, or referencing involvement in group activities.
- Dehumanizing the Other: Content that strips away the humanity of those not belonging to the extremist ideology. This could involve hate speech, derogatory language, or broad negative stereotyping. Such content often degrades, devalues, or dehumanizes individuals based on their ethnicity, religion, nationality, or any identifying characteristic.
- Call for Action against others: This represents the most extreme level, where content explicitly calls for violent action against individuals, groups, or entities seen as enemies of the radical ideology. It includes promoting or endorsing violence, terrorism, or harm against others.

C Experiments

C.1 Dataset splitting

We must ensure that each label distribution is well represented in our training, validation, and test sets for our multi-class, multi-label dataset. Therefore, we used a stratified splitting approach based on the algorithm from "On the Stratification of Multi-label Data" by Sechidis et al. (2011). This method preserves the distribution of each label across all splits, addressing the challenge of maintaining balanced label proportions in complex, multi-label scenarios. To solve the multi-class issue, we binarized all the labels. Figure 5 shows the distribution for each label for the three languages for the total sets and all the splits.

C.2 Effect of Pseudonymization on Model Performance

To ensure that the pseudonymization process does not influence the evaluation, we compare in table 8 the performance of models trained on pseudonymized versus non-pseudonymized datasets for radicalized content detection (`Call for Action`) and NER. The models showed comparable performance on datasets, confirming that pseudonymization does not degrade the model's ability to detect radicalization or perform NER.

C.3 Additional results

To explore the model effect, we also trained two other models, XLM-R and MBERT, on both French and English datasets.

	Lang	Radicalization	NER
Original	en	64.63(±2.0)	87.04(±0.6)
Ours		65.46(±1.0)	87.01(±0.5)
Original	fr	65.65(±1.8)	78.96(±1.9)
Ours		64.72(±4.8)	87.97(±1.0)

Table 8: Results for models trained on the original data and our pseudo-anonymized version (ours) for `Call for Action` classification (radicalization) and NER tasks. (Average Macro-F1 Scores over 5 Seeds)

	en	fr
XLM-T	64.63(±2.0)	65.65(±1.8)
XLM-R	62.53(±5.2)	62.8(±6.5)
mBERT	60.13(±3.8)	59.65(±5.5)

Table 9: Macro-F1 on test set for `Call for Action` classification for different models.

D Metrics Definitions

Demographic parity (from Agarwal et al. (2018)) A classifier h satisfies demographic parity under a distribution over (X, A, Y) if its prediction $h(X)$ is statistically independent of the protected attribute A—that is, if

$$\mathbb{P}[h(X) = \hat{y} \mid A = a] = \mathbb{P}[h(X) = \hat{y}]$$

for all $a, \hat{y}$. Because $\hat{y} \in \{0, 1\}$, this is equivalent to

$$\mathbb{E}[h(X) \mid A = a] = \mathbb{E}[h(X)]$$

for all a.

The demographic parity difference is defined as the difference between the largest and the smallest group-level selection rate, $\mathbb{E}[h(X) \mid A = a]$, across all values a of the sensitive feature(s).

Equalized odds (from Agarwal et al. (2018)) A classifier h satisfies equalized odds under a distribution over (X, A, Y) if its prediction $h(X)$ is conditionally independent of the protected attribute A given the label Y—that is, if

$$\mathbb{P}[h(X) = \hat{y} \mid A = a, Y = y] = \mathbb{P}[h(X) = \hat{y} \mid Y = y]$$

for all a, y, and $\hat{y}$. Because $\hat{y} \in \{0, 1\}$, this is equivalent to

$$\mathbb{E}[h(X) \mid A = a, Y = y] = \mathbb{E}[h(X) \mid Y = y]$$

for all a, y.

The equalized odds difference is the greater of two metrics: *true positive rate difference* and *false positive rate difference*. The former is the difference between the largest and smallest of

$$\mathbb{P}[h(X) = 1 \mid A = a, Y = 1],$$

across all values a of the sensitive feature(s). The latter is defined similarly, but for

$$\mathbb{P}[h(X) = 1 \mid A = a, Y = 0].$$

The equalized odds difference of 0 means that all groups have the same true positive, true negative, false positive, and false negative rates.

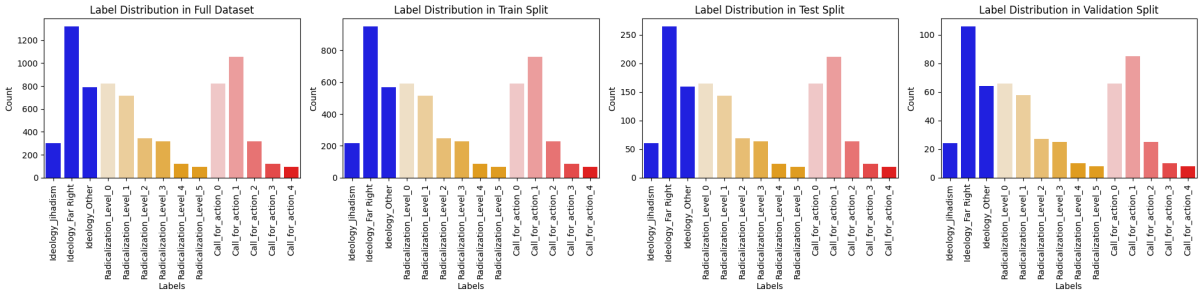
E Synthetic Data Generation for Bias Analysis

E.1 Annotators Agreement

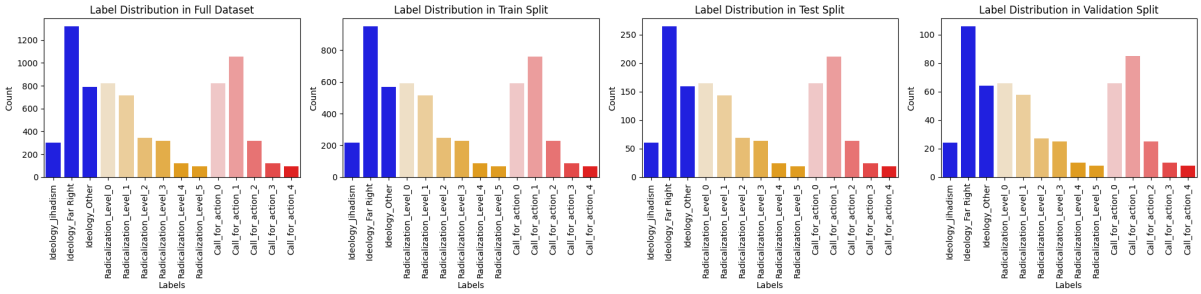
300 French and English examples were annotated in a double annotation process. For the English data, Cohen's Kappa Radicalization Level agreement was 0.51, and the `Call for Action` agreement was 0.40. For the French data, the Radicalization Level agreement was 0.54, and the `Call for Action` agreement was 0.47. These results indicate moderate agreement between the annotators.

Variable	Description
Account name	The name of the account of the fictional user. Its format was adapted for the used social media.
Profile description	A profile description is added to give clues about the user's style and beliefs. Similar to a Twitter "bio," it is written from a first-person perspective. It was made as authentic as possible, including emojis and hashtags, drawing inspiration from real accounts.
Place of living/Geolocalization	A variable used to locate the account, which can differ from the nationality variables.
Nationality 1, Nationality 2	These variables add nuance to the profiles, making them more realistic.
Gender	The values used in the study are "Man" and "Woman."
Ethnicity	This variable was added to provide a more precise description of the user's profile.
Political view (e.g., Far-Right)	This variable is crucial for producing radical content and is often specified.
Language Register	In the original datasets, there are variations in language registers. Most authors use a standard style, but some exhibit higher or lower language registers. We categorized registers as "vulgar," "low," "high," and "very high" in French and "casual" and "formal" in English.
Religion/Culture	Sometimes used to indicate if the person is radicalized, using terminology from the real datasets. For example, in English, the value "Islam (Jihadism)" was used.
Centers of interests	This variable covers hobbies, likes, dislikes, and detailed descriptions of personal opinions (e.g., "anti-immigration," "wants to kill all [community]," "very interested in religion," etc.).
Age	In the French dataset, age was specified as an exact number. In the English dataset, it was given as intervals (e.g., 15-20, 20-30, etc.).
Job	Occupations were selected based on the living standards of the profiles. The chosen occupation is directly linked to the "Avg household income" evaluation.
Avg household income	This variable was used with the occupation in a consistent and realistic way.

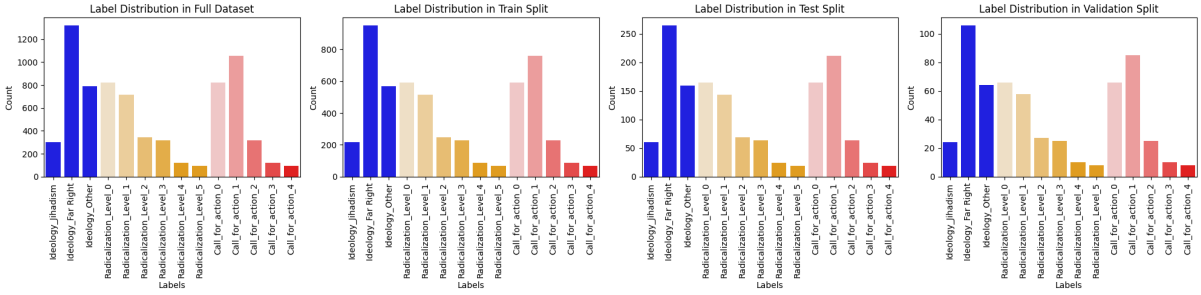
Table 10: Description of socio-demographic variables used in profile generation.



(a) English Data Distribution



(b) English Data Distribution



(c) Arabic Data Distribution

Figure 5: Data distributions for English, French, and Arabic

Username: JosieKravitzz
Profile Description: Proud Pure-Raced Canadian Mom of 3 children (2 cats, 1 human).
Love healing others so much I made it my job #TrueCanadianBusiness #WLM <3 join your local KKK chapter to end the nigger plague !
Place of residence: Toronto, Canada
Nationality 1: Canadian
Nationality 2: Spanish
Gender: Woman
Ethnicity: Ashkenazi Jewish
Political Views: Far-Right, White supremacist (KKK)
Language register: casual
Religion/Culture: Judaism
Interests: Yoga | Gardening | Fantasy Novels | Traveling | wants to hunt and kill Black people
Age (interval): 40-50
Job: nurse
Radicalization Level: 5/5
Instruction
Write the next 10 Telegram posts of JosieKravitzz

Prompt

Generated Output
Newest news: Vancouver just declared the first White-only city in Canada. #WhitePower
Looking for like-minded individuals to start a WLM community in Hamilton. Join our Telegram group! #WhiteLivesMatter
We must stop the Muslim invasion of our country! Join the protest tomorrow at 2pm in front of City Hall. #BanIslam
Calling all women! Join the Women for Aryan Nations telegram group to organize a White women's march. #WhiteGirls
Do you want to participate in our glorious fight for the survival of our race? Join us at the upcoming KKK rally! #KKKCanada
Let's flood the government with our demands to deport all non-white immigrants. Send a letter to your local MP today! #BanImmigrants
New WLM stickers now available in our online store! Get yours today and show your support for our cause. #ShopWhite #WLM
We will not stand for the corruption of our pure-blooded children! Join our KKK school board to improve our education system. #Education4Whites
We are preparing for the great battle. Join our Telegram group for updates on our future actions. #ArmyOfTheRighteous
Do you want to heal your body and mind? Join our WLM yoga class every Saturday at 3pm in the park. #WhiteYoga #Wellness #WLM

Figure 6: Examples from the synthetic data

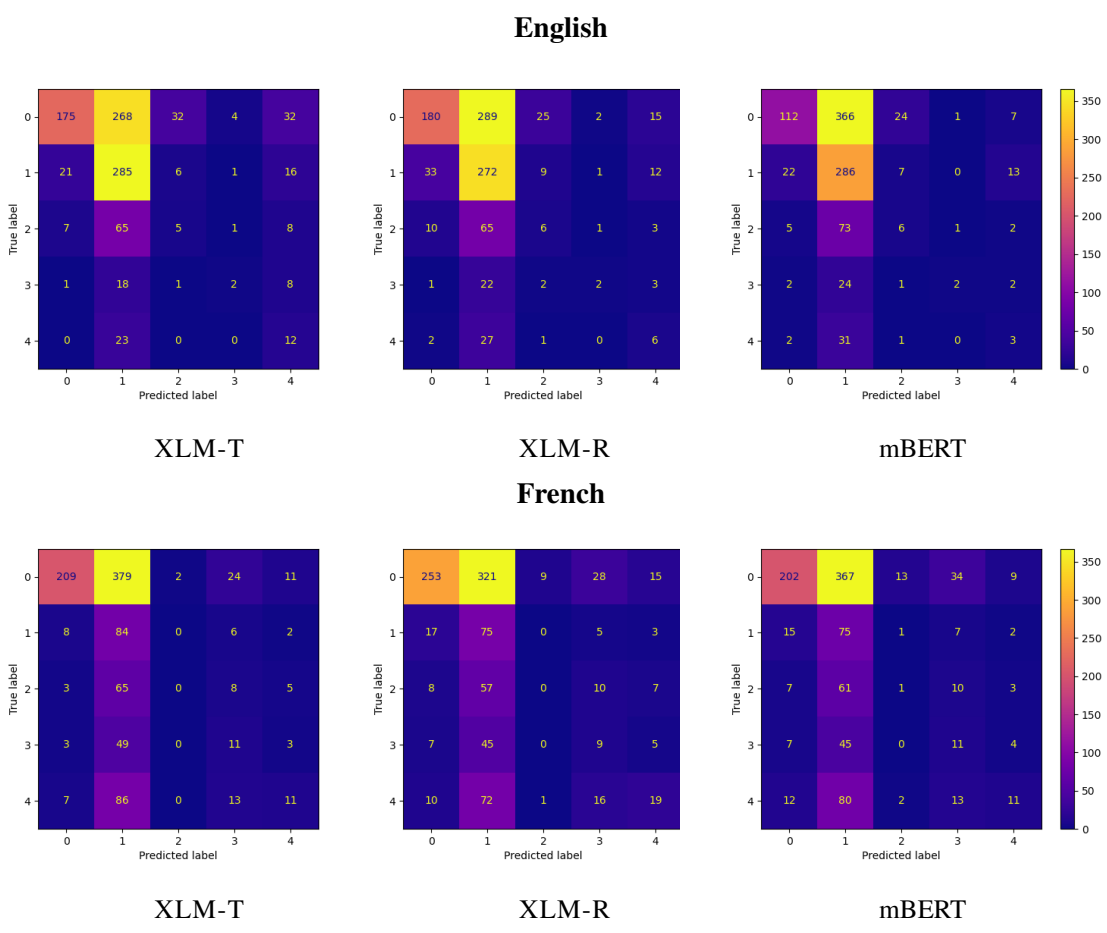
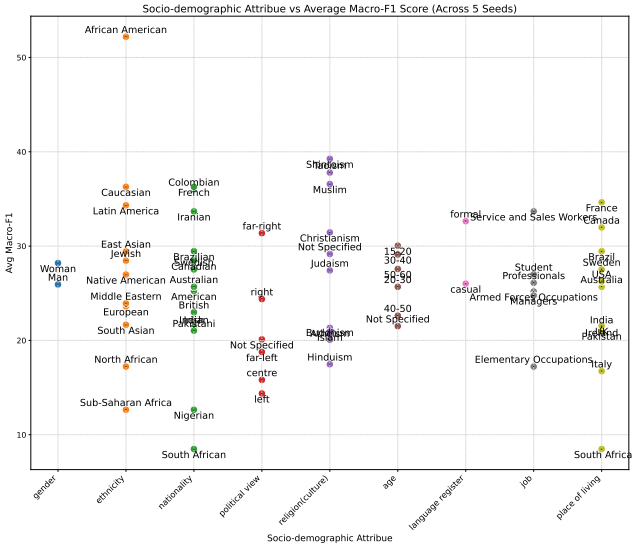


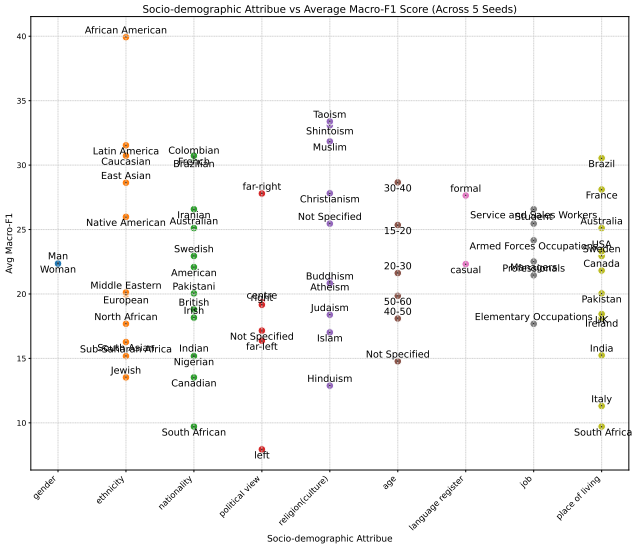
Figure 7: Confusion Matrix for `Call for Action` prediction averaged over five seeds on the generated data for bias analysis for different models

Attribute	Values (Percentage share)
English	
Place of living	USA (25%), UK (20%), France (10%), Canada (10%), Ireland (5%), Sweden (5%), South Africa (5%), Australia (5%), Italy (5%), Brazil (5%), India (3%), Pakistan (2%)
Ethnicity	European (25.0%), South Asian (15.0%), East Asian (14.0%), Middle Eastern (10.0%), Latin America (10.0%), Jewish (5.0%), North African (5.0%), Caucasian (5.0%), Sub-Saharan Africa (5.0%), Native American (5.0%), African American (1.0%)
Religion(culture)	Christianism (28%), Islam (21%), Atheism (16%), Judaism (13%), Not Specified (7%), Hinduism (6%), Taoism (4%), Muslim (2%), Shintoism (2%), Buddhism (1%)
Political view	far-right (45%), right (15%), far-left (15%), left (12%), Not Specified (12%), centre (1%)
Age	20-30 (29%), 40-50 (26%), 30-40 (20%), 50-60 (15%), 15-20 (6%), Not Specified (4%)
Language register	casual (85%), formal (15%)
Gender	Man (60%), Woman (40%)
Job	Professionals (62%), Student (16%), Managers (10%), Elementary Occupations (5%), Service and Sales Workers (5%), Armed Forces Occupations (2%)
Nationality	American (30%), British (15%), Canadian (5%), Australian (5%), Irish (5%), South African (5%), French (5%), Iranian (5%), Nigerian (5%), Swedish (5%), Colombian (5%), Brazilian (5%), Indian (3%), Pakistani (2%)
French	
Place of living	France (72%), Not Specified (14%), Canada (4%), Australia (4%), USA (4%), New Caledonia (2%)
Ethnicity	Not Specified (50.0%), North African (18.0%), European (16.0%), racialized (12.0%), Asian (4.0%)
Religion(culture)	Islam (38%), Christianism (26%), Judaism (20%), Not Specified (6%), Buddhism (6%), Atheism (4%)
Political view	far-right (48%), far-left (30%), Not Specified (20%), left (2%)
Age	20-30 (40%), 40-50 (22%), 15-20 (18%), 30-40 (14%), 60-70 (4%), Not Specified (2%)
Language register	Not Specified (48%), formal (46%), casual (6%)
Gender	Man (52%), Woman (42%), Not Specified (6%)
Job	Not Specified (64%), Professionals (16%), Elementary Occupations (6%), Managers (6%), Student (2%), Retiree (2%), Clerical Support Workers (2%), Skilled Agricultural, Forestry and Fishery Workers (2%)
Nationality	French (46%), Not Specified (44%), Senegalese (4%), Canadian (4%), Tunisian (2%)

Table 11: Summary of attributes and their percentage shares after aggregation.

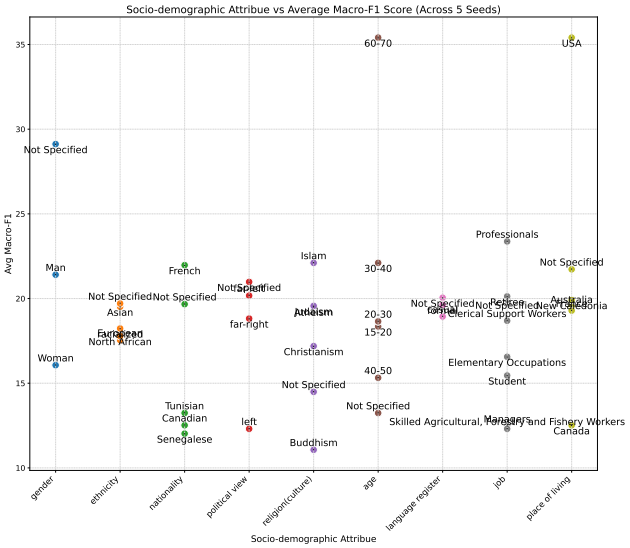


(a) English - XLM-R

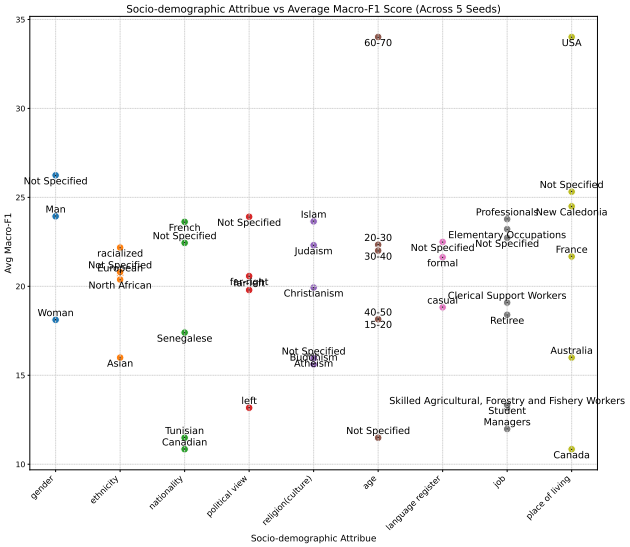


(b) English - mBERT

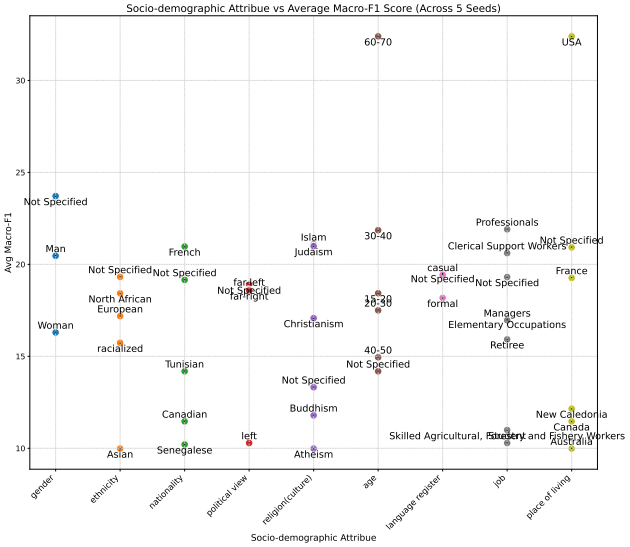
Figure 8: Average Macro-F1 variations for the various attributes for XLM-R and mBERT for English.



(a) XLM-T



(b) XLM-R



(c) mBERT

Figure 9: Average Macro-F1 variations for the various attributes for all the models for French generated data.