A Multilingual Similarity Dataset for News Article Frame

Xi Chen[1], **Mattia Samory**[2], **Scott Hale**[3], **David Jurgens**[4], **Przemyslaw A. Grabowicz**[1]
[1]University of Massachusetts Amherst
[2]Sapienza University of Rome
[3]University of Oxford
[4]University of Michigan
xchen4@umass.edu, mattia.samory@uniroma1.it, scott.hale@oii.ox.ac.uk, jurgens@umich.edu, grabowicz@cs.umass.edu

Abstract

Understanding the writing frame of news articles is vital for addressing social issues, and thus has attracted notable attention in the fields of communication studies. Yet, assessing such news article frames remains a challenge due to the absence of a concrete and unified standard dataset that considers the comprehensive nuances within news content.

To address this gap, we introduce an extended version of a large labeled news article dataset with 16,687 new labeled pairs. Leveraging the pairwise comparison of news articles, our method frees the work of manual identification of frame classes in traditional news frame analysis studies. Overall we introduce the most extensive cross-lingual news article similarity dataset available to date with 26,555 labeled news article pairs across 10 languages. Each data point has been meticulously annotated according to a codebook detailing eight critical aspects of news content, under a human-in-the-loop framework. Application examples demonstrate its potential in unearthing country communities within global news coverage, exposing media bias among news outlets, and quantifying the factors related to news creation. We envision that this news similarity dataset will broaden our understanding of the media ecosystem in terms of news coverage of events and perspectives across countries, locations, languages, and other social constructs. By doing so, it can catalyze advancements in social science research and applied methodologies, thereby exerting a profound impact on our society.

Introduction

Every day, the world's media landscape is enriched with hundreds of thousands of news articles, spanning a multitude of languages and emanating from various corners of the globe. The ability to discern which articles narrate the same story is not just pivotal for refining news aggregation applications but also serves as a gateway to cross-linguistic analysis of media consumption and attention patterns. However, the task of measuring story congruence within these articles is fraught with complexities. Diverse dimensions in storytelling mean that even articles with substantial textual similarities may diverge significantly, recounting similar events that transpired years apart.

In the realm of communication studies, two long-term or cognitively-driven media effects stand out: the agenda-setting theory and the framing theory – Fourie (2001) defined cognition as our capacity for understanding and interpreting information in a specific way, which in turn shapes our behavior and thought processes. While the primary level of agenda-setting focuses on 'what' a story conveys, the framing theory delves deeper into 'how' the story is presented. This nuanced differentiation between 'what' and 'how' in news narratives is not just a linguistic or stylistic concern but fundamentally alters the impact and reception of news among its audiences.

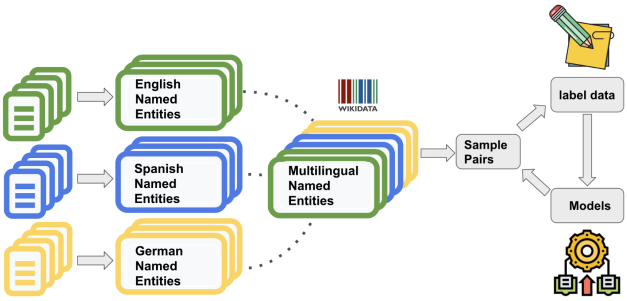

Figure 1: Illustration of sample selection pipeline for annotation.[1]

Therefore, to comprehend if two news articles cover the same story in the same way, it is often not enough to just delve into specific facets of the events portrayed. Also, it requires understanding and evaluating the way to present the events, such as the writing frame.

Although frame theory has gained widespread recognition as a vital area in media research, sparking extensive study and debate, a consensus on a common and unified definition of 'frame' remains missing. A traditional approach involves categorizing news articles into specific categories of frames. However, this method often restricts articles to a limited number of coarse-grained themes that are predefined by researchers, such as crises (An and Gower 2009), gun violence (Liu et al. 2019; Akyürek et al. 2020), or policy issues (Card et al. 2015). Such limitations inhibit the potential for these methodologies to be broadly generalized.

Additionally, most studies in this domain focus on sentence-level analysis, constrained by data availability and

[1]Images by Vector Stall and Eucalyp from Flaticon.com.

GEO	How similar is the geographic focus (places, cities, countries, etc.) of the two articles?
ENT	How similar are the named entities (e.g., people, companies, organizations, products, named living beings), excluding previously considered locations appearing in the two articles?
TIME	Are the two articles relevant to similar time periods or describing similar time periods?
NAR	How similar are the narrative schemas presented in the two articles?
OVERALL	Overall, are the two articles covering the same substantive news story? (excluding style, framing, and tone)
STYLE	Do the articles have similar writing styles?
TONE	Do the articles have similar tones?
FRAME	*Do the articles have similar framing and express similar opinions?*

Figure 2: The annotation scheme used by Chen et al. (2022), which we extended with the FRAME aspect (bottom line).

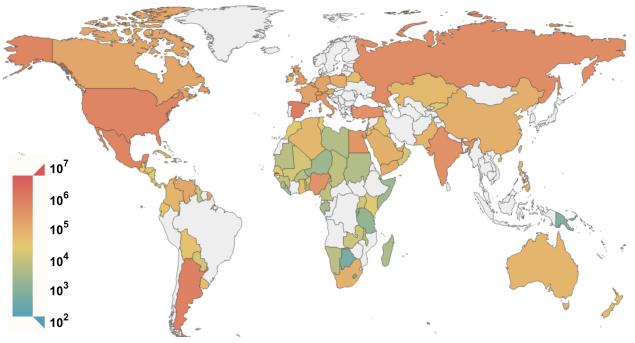

Figure 3: The number of news articles per country in our base dataset. The released dataset includes labels for pairs of articles sampled from the base dataset.

suitable large-scale analytical approaches. This results in studies that often examine only news headlines (Liu et al. 2019; Akyürek et al. 2020) or paragraph contexts (Card et al. 2015), thereby failing to capture the comprehensive essence of the news articles.

A relevant research area, namely targeted sentiment analysis, involves identifying named entities discussed in a document and classifying the sentiment towards them. Similar to frame theory, datasets in targeted sentiment analysis are typically limited in size and scope, focusing mainly on sentence-level data. This limitation hinders accuracy due to the absence of co-reference and discourse context, let alone fully extracting complex relationships within a document's entirety (Steinberger et al. 2017; Luo et al. 2022).

In our work, we have expanded a news article dataset, effectively overcoming these shortcomings in both frame theory and targeted sentiment analysis (Chen et al. 2022). By adopting a human-in-the-loop framework for annotation, we ensured the creation of high-quality, large-scale data. We define the measurement of frame in news articles by expanding upon the aspects of pairwise news similarity, thus circumventing the need for narrowly defined, subjective frame categories, which are often abstract and subtle to operationalize.

Concretely we introduce the most extensive multilingual news article similarity dataset to date, containing nearly 27 thousand news article pairs across 10 languages. This dataset offers a nuanced, pairwise measure, bridging the gap between high-level per-outlet bias and low-level, per-sentence targeted sentiment and frame connotations. We believe it will establish a new benchmark for tasks like cross-lingual document matching, news clustering, and multilingual information retrieval. Furthermore, it paves the way for exploring potential media biases and agendas within different linguistic communities. Our ultimate goal is to bridge societal barriers, enhancing effective news communication and understanding in our evermore connected global society.

Dataset Creation

The news articles are collected from Media Cloud, an online platform that since 2009 has collected more than a billion news articles published globally (Roberts et al. 2021). We focus on news articles in ten languages, used as an official language in 124 countries. This set of countries covers 64% of all countries and about 76% of the world population (Figure 3). Overall, we collected metadata and full text of all news articles from January 1, 2020, to June 30, 2020, totaling ∼60M news articles in the following languages: English, Spanish, Russian, German, French, Arabic, Italian, Turkish, Polish, and Mandarin Chinese.

We recruited and paid annotators with fluency in these languages, to label the news article similarity after rounds of training and calibration. In order to highlight the news similarities that bear significant value for communication and social science research, we meticulously selected relevant news articles. From this curated pool, we then systematically sampled pairs of articles for detailed annotation (Figure 1). The annotated data is available on at https://zenodo.org/records/10611923.[2]

News article selection. Collected metadata of news articles varies in completeness, and thus it cannot be processed in the same way. We omitted the articles that miss any of the following basic attributes: story ID, URL, title, and text; and only included the meaningful and informative news articles with at least 100 word count (after translating into English).[3] Also if a news article has the exact same title or URL as another newer one, we filter it out as a duplicate. Finally, we took some websites out of consideration since they are not trustworthy information sources, or do not focus on societal and political topics [4].

After applying these selection measures, the resulting news article counts per language are as follows: English (10M articles), Spanish (4.6M), Russian (1.8M), German

[2]https://zenodo.org/records/10611923

[3]We apply this threshold after translating to English to make a fair comparison across languages since the average number of words needed to represent an English document in another language depends on that language.

[4]Irrelevant websites include: "reddit.com," "facebook.com," "twitter.com," "fb.com," "wikipedia.org," "epochtimes.com," "youtube.com," "slideshare.net". Any URL containing "sport" is also dropped.

Article 1	Article 2	GEO	ENT	TIME	NAR	STY	TONE	FRAME	OVERALL
Key dates in Indiana's path to reopening amid the coronavirus pandemic	Outlining "measured" lifting of lockdown	VS	VS	VS	VS	SD	VS	**VS**	VS
Concor Head rallies SA to fight virus	Church Leader Defies Coronavirus Measurers	VS	SD	SD	SD	SS	VD	**VD**	SD
VDH confirms 1st coronavirus-related death in Virginia (March 14, 2020)	Virginia sees new coronavirus cases (April 15, 2020)	VS	SS	SD	SS	SD	SS	**SS**	SS

Table 1: Annotation example of news article pairs. Each aspect is labeled as one of the four classes: Very Similar (VS), Somewhat Similar (SS), Somewhat Dissimilar (SD), Very Dissimilar (VD). The first example includes a pair of articles that cover the plan for reopening Indiana, which are very similar in frame. The second pair shares some similarities in terms of Geography (South Africa) and Entities (President Cyril Ramaphosa), but their frames on the lockdown events exhibit significant dissimilarity. One article discusses how an infrastructure company supports the national lockdown for long-term commercial sustainability, while the other focuses on church leaders' disagreement with the lockdown due to its limitations on in-person worship. The final pair of articles both report on coronavirus cases in Virginia, which are somewhat similar in their framing. The first article includes mayoral statements and conveys a sorrowful tone, while the second article presents the information in a more neutral tone, focusing on data.

(1.3M), French (1.2M), Arabic (1.8M), Italian (1.5M), Turkish (655K), Polish (369K), and Mandarin Chinese (205K).

News article pairs sampling. Ideally, the dataset should cover each degree of news article similarity in an even manner. However, if we just sample two random news articles they can hardly be relevant, let alone similar. A crucial challenge of our work was identifying potentially similar pairs of news articles, in order to mitigate the imbalance of similarity labels during the annotation process. We experimented with various strategies to unearth these similar articles, including the comparison of document embeddings (Cr5: Josifoski et al. 2019) or sentence embeddings (Sentence BERT: Reimers and Gurevych 2019) for news headlines and leading paragraphs, as well as the named entities extracted from full texts (Babelfy, polyglot, and spaCy: Moro, Raganato, and Navigli 2014).

After extensive pilot study, including the selection of news article representation for sampling (e.g., we tried Cr5 and Sentence BERT embeddings, but concluded that Wikified named entities work better), filtering process, and developing simple machine learning models for active learning, we ultimately engineered an effective and efficient pipeline to sample and annotate pairs of news articles (Figure 1). Our first step is to extract named entities from each news article, a process facilitated by the use of spaCy and Polyglot.

To avoid the sampling of duplicate pairs (i.e., two articles with identical or nearly identical text, published under different titles and URLs and hence not filtered out during the article selection stage), we introduced an additional filtering step. This process removed all pairs of articles that share one or more long sentences (comprising 40 or more characters) or whose Jaccard similarity of article text exceeds 0.25, a predetermined empirical threshold.

As human annotators started to label the news article pair samples, the sampling quality was iteratively improved via a human-in-the-loop active learning framework. In essence, with each iteration, we built a fresh new logistic regression model and trained it with all the labels that we had obtained up to that point. The model includes features as follows: the word counts of both articles, the number of common words, the number of common named entities, cosine similarity of the named entities with BM25 embeddings (Robertson and Zaragoza 2009), text Jaccard similarity, and an exponentially decaying function of publication date difference.

Annotation Process

In consultation with media studies literature, we established an elaborate codebook, which we share with our dataset,[5] that details the annotation of news article similarity across 8 aspects (Figure 2). To calibrate the annotators' understanding of the codebook, we initially conducted several rounds of trial annotations, including calibration sessions aiming for agreement on 30 carefully selected news article pairs. The subsequent annotations that contributed to the overall dataset were collected in two stages, described next.

In the first stage (July - December 2021), annotators assessed the first seven similarity aspects (GEO, ENT, TIME, NAR, OVERALL, STYLE, TONE). During this time, many article pairs were assigned to multiple annotators to ensure label reliability. This stage resulted in 9,868 annotated pairs, which we released at the SemEval'22 competition focused on news article similarity estimation (Chen et al. 2022).

Then the second stage followed (January - May 2022), during which the FRAME aspect was added. According to our codebook, FRAME similarity "can be judged only when the framing and opinions communicated in the two articles target the same subject, e.g., the articles communicate opinions about Bernie Sanders". The idea that framing is relative

[5]https://zenodo.org/records/10611923

language class	count	mean(OVERALL)
ar	1725	2.66
de	3402	2.54
en	7545	2.76
es	2334	2.33
fr	369	1.99
it	1133	2.64
pl	760	2.33
ru	350	2.77
tr	1310	2.58
zh	3413	2.31
de-en	1956	2.94
de-fr	178	1.92
de-pl	47	1.69
es-en	759	2.84
es-it	474	2.29
fr-en	103	1.21
fr-pl	53	1.67
pl-en	119	2.35
zh-en	525	2.96
total	26555	2.59

Table 2: Counts of annotated monolingual pairs (e.g., English article pairs – "en") and cross-lingual pairs (e.g., pairs of a German article and an English article – "de-en").

to a target comes from the research area of targeted sentiment analysis. Due to this requirement, and because we introduced FRAME aspect gradually while discussing it with our expert annotators, we collected FRAME similarity labels for a relatively small subset of 1522 news article pairs.[6] Each pair was labeled by a single well-trained annotator to maximize the number of annotated news article pairs. Overall, the second stage brought 16,687 annotated pairs.

Each aspect was labeled with four ordinal classes – *Very Dissimilar*, *Somewhat Dissimilar*, *Somewhat Similar*, and *Very Similar*. To accommodate exceptional cases, we also offer an *Other* option. The most common exceptions that we met were pairs of duplicate new articles or inaccessible articles, for instance, due to paywalls or removal (annotators were instructed to report such instances via a free-text comment).

To satisfy the desired linguistic diversity and scale of news annotation, we trained 25 annotators, recruited from three institutions (GESIS, UMass, Umich). They did several rounds of calibration for the gold standards in our codebook Chen et al. (2022) before starting their annotation tasks. The entire annotation process lasted for roughly 11 months. We paid each annotator €12 per hour at GESIS and $15 per hour at UMass and UMich.

The annotation process was running through a custom annotation interface devised by our team. It shuffles news article pairs and assign them to annotators as per their respective language capabilities. To engage and motivate the annotators, the interface also offers statistical feedback, such as the annotation count ranking and the inter-rater agreement ranking among annotators. This also enabled us to identify the annotators who did not perform their task correctly. We recorded the disagreements and discussed them with annotators biweekly as part of our iterative calibration process.

[6]The vast majority of dissimilar news article pairs do not share common targets and thus their FRAME similarity is not labeled.

	GEO	ENT	TIME	NAR	OVERALL	STYLE	TONE
Krip. α	0.73	0.69	0.57	0.69	0.77	0.46	0.38
Gwet's AC_1	0.75	0.70	0.78	0.71	0.79	0.60	0.60

Table 3: Inter-rater agreement measures, Krippendorf's α and Gwet's AC_1, for the similarity aspects (FRAME aspect is missing since for each pair it is only annotated by one well-trained annotator).

Dataset Description

Format and Statistics

We collected the labels for roughly 11 months from 2021 July to 2022 May. Ultimately, we got the news article similarities for 26,555 pairs (Table 1), including 1522 with frame similarity labels and 4,214 cross-lingual pairs (Table 2). In the remainder of this manuscript, we represent the four similarity labels on an interval scale from 1 (Very Dissimilar) to 4 (Very Similar), e.g., Table 2 shows the average similarity for different groups of labels.

Inter-annotator Agreement

To evaluate the reliability of the similarity labels, we calculated the inter-rater agreement for each aspect. The annotators exhibited remarkably high agreement on the OVERALL similarity aspect, as evidenced by a Krippendorff's α of 0.77. We also leveraged Gwet's AC_1 measure, which is known to be less sensitive to non-uniform marginal label distributions (Gwet 2008). Therefore, it allows us to offset bias arising from skewed distributions within some aspects. All aspects demonstrated good inter-rater agreements under this measure.

Applications

To the best of our knowledge, our dataset is the largest to date for assessing news article similarity, with meticulous evaluations conducted across multiple languages. Consequently, it holds significant potential to empower a wide range of applications within the fields of media communication and social science, including news aggregation, media consumption analysis, cross-cultural studies, agenda setting research, linguistic studies, and political science research. Next, we present three examples (our code of implementations is available on Github.[7].

Global News Synchrony and Diversity

As the multilingual news article similarity offers a unified representation that transcends language barriers, it enables us to understand the news media across multiple countries, or even on a global scale. In a recent paper, we develop an

[7]https://github.com/social-info-lab/global_news_synchrony

Event	News article title
Oscar 2020	Oscary 2020 należały do "Parasite" i zmieniły historię gali. Pełna lista zwycięzców Oscar 2020: revisa la lista de ganadores de las 24 categorías con lo mejor del cine Oscars 2020: Los usuarios de internet premian a Joker, Leonardo DiCaprio, Scarlett Johansson y Martin Scorsese A dos horas de la ceremonia, los Oscar se aprontan entre favoritos y posibles sorpresas Parasite hizo historia en los Oscars y se llevÃ³ el premio a mejor pelÃ cula Ganadores de los Oscar 2020 Nigeria Records 245 New Cases Of COVID-19, Highest Single-Day Increase Oscars 2020: South Korean movie makes history by winning best picture Múltiples latinos se miden esta noche en los Oscar Glamour y talento en la gala de los premios Oscar en su 92 edición Oscar night begins with '1917' battling 'Parasite'
Covid 2020	Canadá acepta tener varios posibles casos de coronavirus Prevén se presenten casos de coronavirus chino en México Qué se sabe sobre el coronavirus de China que "puede haber afectado a cientos de personas", según científicos británicos Chinese confirm coronavirus outbreak can spread like wildfire from infected people | Fox Business China locks down more cities as virus spreads US source: North Korean leader Kim Jong Un in grave danger after surgery Un virus en Chine commence à inquiéter au Canada El nuevo coronavirus deja 106 muertos y 4.515 infectados en China Lo que se sabe sobre el coronavirus detectado en China y otros países que ya ha afectado a cientos de personas WDH/VIRUS/ROUNDUP: Vorerst keine 'internationale Notlage' Virus chinois: sans doute des centaines de contaminations, inquiétude à l étranger

Table 4: Random samples of news articles in two exemplary news event clusters: Oscar 2020 covered by news in English and Spanish, with a small portion in Polish; Covid 2020 covered by news in various languages from Asia, North America, South America and Europe.

effective methodology for news coverage studies at a massive scale and measure news diversity and synchrony across countries (Chen et al. 2024).

Challenges. The key challenges to examining news coverage at a global scale are the following. First, traditional data collection and validation based on physical newspapers and questionnaires requires human effort that scales linearly with the amount of data and the number of languages. These practical considerations severely limit the data size and linguistic coverage of traditional studies. Second, it is not clear how to identify global news events, which are necessary to measure news coverage of events. Existing methods for identifying which events are reported in the news prioritize precision over coverage, since such methods are based on keyword matching (Card et al. 2015), inevitably leading to the lack of generality. We overcome these challenges by contributing a novel computational methodology for studies of global news coverage thanks to the labeled dataset.

News similarity inference. First, we develop a computationally-efficient transformer model that infers multilingual news similarity. The model achieves the highest score (Pearson correlation with human annotations of 0.8) among similar efficient models in the prior benchmark (Chen et al. 2022) computed on our labeled dataset, achieving performance comparable to average human annotators.

Global event detection. Second, using this model, we compute similarity among millions of news article pairs. Then, we apply a graph-clustering algorithm on the resulting similarity network to identify 4,357 multilingual news events. The largest events, in chronological order, were: the assassination of Iranian general Soleimani, U.S. presidential election primaries, the Covid-19 pandemic, and protests after the killing of George Floyd (see examples in Table 4). We evaluate the quality of the identified news events by an intrusion task, which is commonly used to evaluate topic models (Chang et al. 2009). We recorded an average high precision of 85.8% across the annotators (97.5% for the annotator who spent the most time on the task).

News diversity and synchrony measures. Third, we introduce information-theoretic measures of country-level synchrony and diversity in news coverage of global events. We define the *diversity* of news in a country as the entropy of the distribution of the news published in that country across the inferred events. As the *synchrony* of news across a pair of countries we define the Jensen-Shannon divergence of the respective distributions. Next, we regress the diversity within a country and synchrony across countries against country-level predictors. The regression yields much higher adjusted R^2 for the introduced measures of diversity and synchrony (R^2 of 0.54 and 0.45, respectively) than naive baseline measures based on an average of pairwise news article similarity that do not make use of global news events (R^2 of 0.13 and 0.30, respectively).

Findings. The labeled dataset and proposed methodology enable the discovery of unexpected patterns in global news coverage. For instance, prior studies suggest that the acceleration of the news cycle in the Internet age contributes to the homogenization of news coverage (Bucy, Gantz, and Wang 2014; Boczkowski and de Santos 2007; McGregor 2019; Zuckerman 2013). However, we find that Internet adoption is the *strongest* predictor of news diversity within

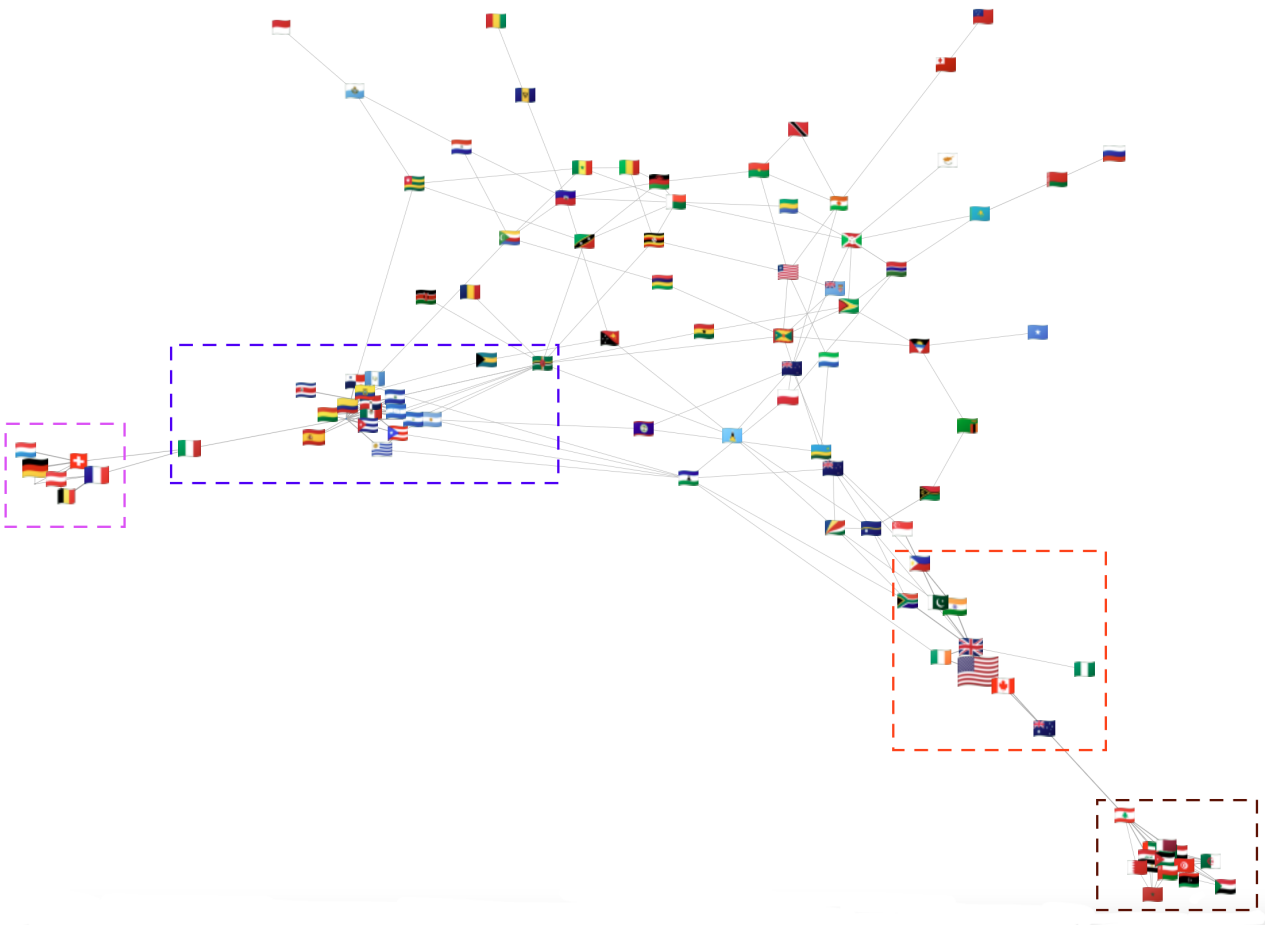

Figure 4: News event graph backbone for the top 100 countries with the largest populations. The main communities are marked with squares: the US and the UK and their former colonies (red square), five countries of the old European Union of 1958 (purple), Latin America and Spain (blue), and the Arab world (brown). Most of these communities align with the geographic-linguistic groups in (Kim and Barnett 1996). Some countries are not selected into the graph backbone, e.g., China. Country flag size corresponds to country GDP, and each edge represents 95% confidence (Serrano, Boguná, and Vespignani 2009) that it represents non-random synchrony in news event coverage.

a country ($p < 0.005$). The higher the Internet penetration, the larger the news diversity, probably because news media cater to diverse interests and motivations of online audiences (Lee 2013)). This finding illustrates the potential of the proposed methodology to contribute to communication science in ways that go beyond confirmations of existing theories. In addition, we find that news coverage is more diverse in countries with multiple official languages ($p < 0.005$), more diverse religious practices ($p < 0.005$), greater economic disparities, and larger populations ($p < 0.05$).

The international news synchrony network based on the proposed synchrony measure reveals groups of countries that synchronize in their news coverage of events (Figure 4): (i) the US, UK, and UK's past colonies, (ii) the old European Union of 1958, (iii) Latin America, and (iv) Arab countries. We find that trade volume is the strongest predictor of news synchrony between countries ($p < 0.005$), which corroborates prior findings (Wu 2000; Segev 2016). Furthermore, coverage of news events is more synchronized between countries that share an official language ($p < 0.005$), high GDP ($p < 0.05$), and high democracy indices ($p < 0.005$). Interestingly, countries that belong to NATO experience more news synchrony ($p < 0.05$), possibly because they have common security concerns and some of the largest news events correspond to military operations. Countries belonging to BRICS ($p < 0.05$) exhibit more synchronized news, possibly due to their common developmental interests.

Media Bias Analysis

Media outlets often exhibit biases in their coverage of events and the emphasis they place on them. These biases are influenced by various factors such as political stances, national interests, cultural beliefs, and target audiences (Mrogers and Wdearing 1988). These social factors significantly affect how a story is presented, including its frame and tone

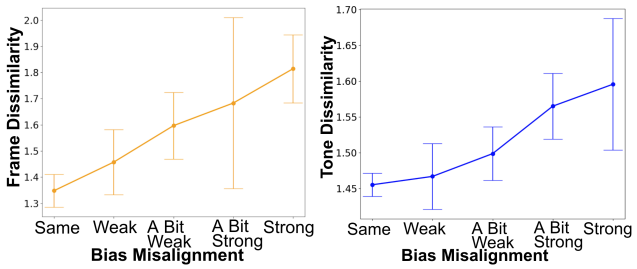

Figure 5: The average frame dissimilarity (left) and tone dissimilarity (right) for pairs of media outlets with a certain level of political bias misalignment.

(Aruguete 2017). The framing of a story can direct audiences towards specific truths or opinions, especially those that resonate with their existing beliefs and knowledge (Scheufele and Tewksbury 2007). Additionally, the tone of a story can subtly guide the conveyance of ideology from media to the public, leveraging audience conformity (Ambady and Skowronski 2008; LeBon and Nye 2017).

Media bias vs news frame and tone. Here, we study the relationship between media outlet bias and the framing and tone of the news articles they publish.

We compiled a list of media outlet biases from the Media Bias/Fact Check website (Zandt 2022). Additionally, we consulted three supplementary sources of media bias scores: (1) a dataset on partisanship from the 2016 American presidential election, with biases inferred based on whether a Twitter user followed Hillary Clinton or Donald Trump (Faris et al. 2017); (2) voter registration data for Democrats and Republicans from 2018 (Robertson, Lazer, and Wilson 2018); and (3) an analysis of political bias among active Twitter users from January 2019 to June 2019, utilizing the emIRT tool (Imai, Lo, and Olmsted 2015; University 2020). These biases span five categories on the left-right spectrum: *Left*, *Left-center*, *Center*, *Right-center*, and *Right*. As the bias of an outlet, we used the label that was the most common across the above four sources. In cases where this procedure identifies varying bias labels, we use the most left-leaning one.

We plot the frame and tone dissimilarity as a function of media bias misalignment (Figure 5), classified into five categories based on the ideological distance on the spectrum from *Left* bias to *Right* bias: *Same*, *Weak*, *A Bit Weak*, *A Bit Strong*, and *Strong*. For example: if one news outlet has a *Left* bias and the other has a *Right* bias, then the bias misalignment is *Strong*; but if the second article has a *Center* bias, then the bias misalignment is *Weak*. As expected, our findings reveal that the frame and tone of a pair of news articles become more dissimilar if the biases of the media outlets where they were published are more misaligned (Figure 5).

We also calculated the correlations between frame dissimilarity, tone dissimilarity, and media bias misalignment at the level of article pairs (Table 5). We find a modest correlation between frame/tone similarity and media bias alignment. Notably, frame similarity demonstrates a slightly stronger correlation than tone similarity, although both correlations are relatively weak. The correlations between frame and tone are both around 0.4, which indicates medium strength.

	Frame vs Bias	Tone vs Bias	Frame vs Tone
Spearman	0.185	0.087	0.394
Pearson	0.202	0.083	0.410

Table 5: Correlations between frame similarity, tone similarity, and media bias. All the correlations are statistically significant with p-values less than 10^{-9}.

Media bias on president power. News outlets may portray biased social images of politicians. To investigate this phenomenon, we utilized Riveter (Antoniak et al. 2023), an advanced tool capable of identifying named entities within each article and assigning a power score to each of these entities. This power score reflects the entity's perceived power strength, based on the actions associated with it in the connotation frame (Sap et al. 2017). Our process began with applying Riveter to deduce the power scores of named entities. We then merged the coreferences of these entities and excluded those appearing in articles from fewer than three different news outlets. Subsequently, we segmented the remaining articles into categories based on differing political biases. Our analysis revealed that across all categories, news outlets consistently depicted Biden as having higher power strength than Trump (Figure 6). This could be attributed to their inherent social images in public perception (such as Biden usually refers to collaborative power words like U.S. institutions, achievements, and morality, consistent with the constructs of prestige and traditional power; while Trump is around coercive power words like "defeat," "threat," "poison," "administration," "failed," "ignored," or "promised.", which may be interpreted as denial of Trump's trustworthiness and dependability (Körner et al. 2022). Interestingly, outlets with a bias leaning towards the *Left* portrayed Biden as more influential than those leaning towards the *Right*, with the opposite trend observed for Trump, similar as the political biases of parties they two respectively represent and obtain support from (Democratic for *Left* and Republic for *Right*).

Multi-factor Analysis

Our dataset allows for a nuanced analysis that considers multiple factors together and thus extends beyond a single social factor to analyze how various elements like political biases, language use, and country-specific factors interact to influence news similarity. By integrating these diverse aspects into our investigation, we provide a more complete picture of the dynamics shaping global news narratives, offering deeper insights into how political, cultural, and national factors collectively influence news coverage and framing.

Next, we quantify the importance of each of these factors on OVERALL, TONE, and FRAME similarity of news articles. To this end, we built three logistic regression models and compared the factors' coefficients with each other. In these models, the pairwise alignment of each factor becomes

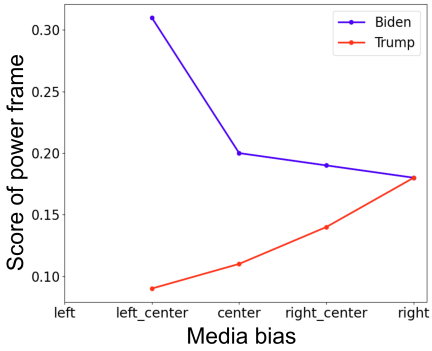

Figure 6: Power scores of Biden and Trump inferred by Riveter based on news articles published in news outlets with given media bias.

a feature and they are combined as the input, while the news similarity of OVERALL, TONE, or FRAME aspect becomes output for each of the three classifiers. For simplicity, we represent language alignment and country alignments as binary: 1 for exact matches and 0 for differences. Bias alignment is normalized to [0,1] range as ordinal factor with equal numeric interval ("Same" is converted to 1 and "Very Strong" is converted to 0), for comparing its importance with other factors at the same scale. The similarities are also recast into binary: 1 for *Very Similar* and *Somewhat Similar*, and 0 for *Somewhat Dissimilar* and *Very Similar*.

Table 6 displays the logistic regression weights of each factor. Interestingly, bias alignment positively correlates to content similarity, possibly because contentious topics, often covered by politically polarized outlets, drive more discussion. Shared languages and shared countries publishing the news articles also showed a positive correlation to OVERALL similarity, reflecting common cultural beliefs and ties of national interest. However, the alignment of the countries mentioned in the news articles seems to correlate negatively with content similarity.

It is worth noting that the strongest predictor of frame similarity is the bias alignment. The alignment of the country of publication also has some importance, potentially reflecting the influence of national interest on the news frame. In terms of predicting tone similarity, both bias alignment and language alignment demonstrate greater importance than country alignment. Table 6 also demonstrates that frame and tone similarities are more predictable than overall content similarity. This observation aligns with the closer proximity of frame and tone to the chosen factors.

Ethics Statement

Our similarity labels are based on full texts of news articles, but we only provide their URLs to access the full texts to prevent copyright issues. To ensure that the news content is reliable with a focus on socially meaningful topics or events, we declined all URLs from popular social media platforms (twitter.com, facebook.com, reddit.com, etc.). We encouraged the annotators to report such cases and introduced a button in our annotation interface to tag any content that they felt was hateful or harmful.

Conclusion

In the ever-evolving landscape of global communication, understanding the interconnections between news articles is more than an academic pursuit—it's a key to unlocking insights into media studies and societal dynamics. Our extension of a multilingual news article similarity dataset sheds light on this intricate web by revealing commonalities across news articles in eight distinct aspects, including a novel approach to define news frames. This dataset is not just a tool; it's a window into the global media narrative, offering a unique vantage point for identifying international media networks, uncovering inherent biases in news outlets, and understanding the portrayal of presidential power across various media platforms.

However, the potential of this dataset extends far beyond these initial applications. It sets the stage for innovative research in global agenda setting, allowing for an in-depth exploration of media biases on a global scale. Imagine unearthing the stark disparities in news coverage, such as why African disasters require far more casualties to gain the same level of US media attention as those in Eastern Europe (Eisensee and Strömberg 2007). Our dataset provides the granularity needed to dissect political campaigns and societal beliefs, revealing the subtle nuances that shape public opinion and discourse.

Moreover, the dataset is poised to facilitate critical analyses of long-term bias and synchrony trends in international news, e.g., in the periods leading up to war outbreaks. This kind of research promises to enhance our understanding of how media narratives intertwine with public sentiment, political maneuvers, international relations, and the genesis of global conflicts. In essence, it offers a lens to view and interpret the complex interplay of factors that drive the world's news stories.

By providing methodologies to observe and interpret the continuous evolution of our global narrative through the news, this work not only contributes to academic discourse but also offers profound societal insights. It underscores how international news coverage, in all its complexity, reflects and shapes our understanding of geopolitical histories and local realities. In doing so, this dataset stands as a pivotal resource for those seeking to comprehend the narrative of our global society.

Acknowledgments

This research has received funding through grants from the Volkswagen Foundation and the University of Massachusetts Amherst. We sincerely thank Media Cloud for data access and the annotators for annotating multilingual news similarity.

References

Akyürek, A. F.; Guo, L.; Elanwar, R.; Ishwar, P.; Betke, M.; and Wijaya, D. T. 2020. Multi-label and multilingual news framing analysis. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, 8614–8624.

	OVERALL		FRAME		TONE	
Alignment	**coef**	**P> |t|**	**coef**	**P> |t|**	**coef**	**P> |t|**
Bias	-0.184	**	0.137	*	0.091	**
Language	0.097	**	-0.013		0.084	**
Publication country	0.203	*	0.128	*	0.037	*
Referred country	-0.248	**	0.059		0.046	*
F1 Score	0.515		0.689		0.656	
Accuracy	0.530		0.638		0.610	

Table 6: Classification evaluation measures (F1, accuracy) and the coefficients of various predictors for the logistic regression models of OVERALL, FRAME, and TONE similarity. Significance level: difference from zero of more than one (*) or three (**) standard deviations, which correspond to p-values of less than 0.1 and 0.001, respectively.

Ambady, N.; and Skowronski, J. J. 2008. *First impressions*. Guilford Press.

An, S.-K.; and Gower, K. K. 2009. How do the news media frame crises? A content analysis of crisis news coverage. *Public relations review*, 35(2): 107–112.

Antoniak, M.; Field, A.; Mun, J.; Walsh, M.; Klein, L.; and Sap, M. 2023. Riveter: Measuring Power and Social Dynamics Between Entities. In Bollegala, D.; Huang, R.; and Ritter, A., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 377–388. Toronto, Canada: Association for Computational Linguistics.

Aruguete, N. 2017. The agenda setting hypothesis in the new media environment. *Comunicación y sociedad*, (28): 35–58.

Boczkowski, P. J.; and de Santos, M. 2007. When More Media Equals Less News: Patterns of Content Homogenization in Argentina's Leading Print and Online Newspapers. *Political Communication*, 24(2): 167–180.

Bucy, E. P.; Gantz, W.; and Wang, Z. 2014. Media technology and the 24-hour news cycle. In *Communication technology and social change*, 143–163. Routledge.

Card, D.; Boydstun, A.; Gross, J. H.; Resnik, P.; and Smith, N. A. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 438–444.

Chang, J.; Gerrish, S.; Wang, C.; Boyd-Graber, J.; and Blei, D. 2009. Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, 22.

Chen, X.; Hale, S.; Jurgens, D.; Samory, M.; Zuckerman, E.; and Grabowicz, P. 2024. International News Synchrony During the Start of the COVID-19 Pandemic. In *Proceedings of the ACM Web Conference 2024*.

Chen, X.; Zeynali, A.; Camargo, C.; Flöck, F.; Gaffney, D.; Grabowicz, P.; Hale, S.; Jurgens, D.; and Samory, M. 2022. SemEval-2022 Task 8: Multilingual news article similarity. In *SemEval-2022*, 1094–1106.

Eisensee, T.; and Strömberg, D. 2007. News droughts, news floods, and US disaster relief. *The Quarterly Journal of Economics*, 122(2): 693–728.

Faris, R.; Roberts, H.; Etling, B.; Bourassa, N.; Zuckerman, E.; and Benkler, Y. 2017. Partisanship, propaganda, and disinformation: Online media and the 2016 US presidential election. *Berkman Klein Center Research Publication*, 6.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/. Accessed: 2024-01-15.

Fourie, P. J. 2001. *Media Studies: Content, audiences, and production*, volume 2. Juta and Company Ltd.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Gwet, K. L. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1): 29–48.

Imai, K.; Lo, J.; and Olmsted, J. 2015. emIRT: EM Algorithms for Estimating Item Response Theory Models. *available at the Comprehensive R Archive Network (CRAN). http://CRAN. R-project. org/package= list*.

Josifoski, M.; Paskov, I. S.; Paskov, H. S.; Jaggi, M.; and West, R. 2019. Crosslingual Document Embedding as Reduced-Rank Ridge Regression. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM.

Kim, K.; and Barnett, G. A. 1996. The determinants of international news flow: A network analysis. *Communication Research*, 23(3): 323–352.

Körner, R.; Overbeck, J. R.; Körner, E.; and Schütz, A. 2022. How the linguistic styles of Donald Trump and Joe Biden reflect different forms of power. *Journal of Language and Social Psychology*, 41(6): 631–658.

LeBon, G.; and Nye, R. A. 2017. *The crowd*. Routledge.

Lee, A. M. 2013. News audiences revisited: Theorizing the link between audience motivations and news consumption. *Journal of Broadcasting & Electronic Media*, 57(3): 300–317.

Liu, S.; Guo, L.; Mays, K.; Betke, M.; and Wijaya, D. T. 2019. Detecting frames in news headlines and its application to analyzing news framing trends surrounding US gun

violence. In *Proceedings of the 23rd conference on computational natural language learning (CoNLL)*, 504–514.

Luo, Y.; Cai, H.; Yang, L.; Qin, Y.; Xia, R.; and Zhang, Y. 2022. Challenges for open-domain targeted sentiment analysis. *arXiv preprint arXiv:2204.06893*.

McGregor, S. C. 2019. Social media as public opinion: How journalists use social media to represent public opinion. *Journalism*, 20(8): 1070–1086.

Moro, A.; Raganato, A.; and Navigli, R. 2014. Entity Linking meets Word Sense Disambiguation: A Unified Approach. *Transactions of the Association for Computational Linguistics (TACL)*, 2: 231–244.

Mrogers, E.; and Wdearing, J. 1988. Agenda-setting research: Where has it been, where is it going? *Annals of the International Communication Association*, 11(1): 555–594.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Network s. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. Hong Kong, China: Association for Computational Linguistics.

Roberts, H.; Bhargava, R.; Valiukas, L.; Jen, D.; Malik, M. M.; Bishop, C.; Ndulue, E.; Dave, A.; Clark, J.; Etling, B.; Faris, R.; Shah, A.; Rubinovitz, J.; Hope, A.; D'Ignazio, C.; Bermejo, F.; Benkler, Y.; and Zuckerman, E. 2021. Media Cloud: Massive Open Source Collection of Global News on the Open Web. In *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media (ICWSM2021)*.

Robertson, R. E.; Lazer, D.; and Wilson, C. 2018. Auditing the personalization and composition of politically-related search engine results pages. In *Proceedings of the 2018 World Wide Web Conference*, 955–965.

Robertson, S.; and Zaragoza, H. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.*, 3(4): 333–389.

Sap, M.; Prasettio, M. C.; Holtzman, A.; Rashkin, H.; and Choi, Y. 2017. Connotation Frames of Power and Agency in Modern Films. In Palmer, M.; Hwa, R.; and Riedel, S., eds., *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2329–2334. Copenhagen, Denmark: Association for Computational Linguistics.

Scheufele, D. A.; and Tewksbury, D. 2007. Framing, agenda setting, and priming: The evolution of three media effects models. *Journal of communication*, 57(1): 9–20.

Segev, E. 2016. The group-sphere model of international news flow: A cross-national comparison of news sites. *International Communication Gazette*, 78(3): 200–222.

Serrano, M. Á.; Boguná, M.; and Vespignani, A. 2009. Extracting the multiscale backbone of complex weighted networks. *PNAS*, 106(16): 6483–6488.

Steinberger, R.; Hegele, S.; Tanev, H.; and Della Rocca, L. 2017. Large-scale news entity sentiment analysis. In Mitkov, R.; and Angelova, G., eds., *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, 707–715. Varna, Bulgaria: INCOMA Ltd.

University, B. K. C. H. 2020. Public Discourse in the U.S. 2020 Election: Resources and Data. https://cyber.harvard.edu/research/2020-election-study-resources-data. Accessed: 2024-03-30.

Wu, H. D. 2000. Systemic determinants of international news coverage: A comparison of 38 countries. *Journal of communication*, 50(2): 110–130.

Zandt, D. V. 2022. Media Bias/Fact Check - Search and Learn the Bias of News. https://mediabiasfactcheck.com/. Accessed: 2022-05-20.

Zuckerman, E. 2013. *Rewire: Digital cosmopolitans in the age of connection*. WW Norton & Company.

Paper 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.
   (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? NA.
   (e) Did you describe the limitations of your work? Yes.
   (f) Did you discuss any potential negative societal impacts of your work? No, because we haven't found any potential negative societal impacts.
   (g) Did you discuss any potential misuse of your work? NA.
   (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? NA.
   (b) Have you provided justifications for all theoretical results? NA.
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA.
   (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? NA.
   (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)? Yes.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? No. Because this is dataset paper and its application details are not the focus.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes.
   (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)? NA.
   (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? Yes.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? Yes.
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? Yes.
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? Yes.
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes.
   (d) Did you discuss how data is stored, shared, and deidentified? Yes.