Scholars Journal of Arts, Humanities and Social Sciences

OPEN ACCESS

Abbreviated Key Title: Sch J Arts Humanit Soc Sci
ISSN 2347-9493 (Print) | ISSN 2347-5374 (Online)
Journal homepage: https://saspublishers.com

# Global Voices, Local Frames: Cross-Lingual Corpus Analysis of Stance and Discourse in Social Media and News

Sadia Aslam[1], Sania Arshad[1], Zeeshan Shabir[1*], Muhammad Sohail[2], Rukhsana Ishaq[3], Hamna Hameed[4], Sibgha Javed[4], Aqsa Habib Ahmed[4], Nimra Habib Ahmed[4]

[1]Department of English Linguistics, University of Okara, Punjab, Pakistan
[2]BS English Literature and Linguistics, National University of Modern Languages (NUML), Khyber Pakhtunkhwa, Pakistan
[3]Department of Sociology, GC University, Faisalabad, Punjab, Pakistan
[4]Department of English Language and Literature, University of Okara, Punjab, Pakistan

**DOI:** https://doi.org/10.36347/sjahss.2025.v13i09.005 | **Received:** 04.07.2025 | **Accepted:** 09.09.2025 | **Published:** 22.09.2025

***Corresponding author:** Zeeshan Shabir
Department of English Linguistics, University of Okara, Punjab, Pakistan

**Abstract** **Original Research Article**

Most stance and discourse corpora are English-centric and biased toward the Global North, limiting our ability to study how publics across the world frame and take positions on global issues. This paper introduces the Global Voices, Local Frames Corpus (GVLF-C), a large-scale, cross-lingual dataset covering South Asian, Sub-Saharan African, and Latin American news and social media texts (2015–2025). The corpus includes 12 languages and ~120K manually annotated documents, labeled for stance (pro/anti/neutral) and frames (economic, justice, identity, urgency, and others). We benchmark state-of-the-art multilingual models (mBERT, XLM-R, LaBSE, NLLB embeddings), demonstrating that few-shot annotation in low-resource languages yields substantial performance gains, while frame detection remains harder due to conceptual overlaps. Diachronic analyses reveal event-driven framing shifts around COP summits, natural disasters, and elections, with distinct regional emphases: *justice and indigenous rights* in Latin America, *adaptation and vulnerability* in Africa, and *national identity/security* in South Asia. Our findings underscore the need to decenter Global North perspectives in NLP resources, highlight cultural variability in framing, and propose a path forward for more inclusive, diachronic, and representative discourse analysis.

**Keywords:** Cross-lingual NLP; stance detection; media framing; diachronic discourse; low-resource languages; climate communication; political communication; Global South; corpus linguistics; multilingual transformers.

## 1. INTRODUCTION

Public debates on climate change and electoral politics unfold within a multilingual, multi-platform information ecosystem. However, the empirical foundations for studying these debates remain disproportionately English-centric and skewed toward the Global North. Comparative audits of natural language processing (NLP) research and benchmarks reveal enduring inequalities across the world's ~7,159 living languages, with only a small fraction receiving robust technological support while most remain underrepresented in datasets, benchmarks, and tools [1–3]. Beyond normative concerns, this imbalance generates significant analytical blind spots: when corpora neglect South Asian, African, and Latin American languages and outlets, we risk mischaracterizing how global publics articulate and contest shared challenges such as climate change and democratic politics [1–3].

Widely used resources illustrate this imbalance. The Media Frames Corpus (MFC) provides detailed frame annotations but is restricted to U.S. English news [5]. SemEval-2016 Task 6 standardized stance detection on Twitter, yet again exclusively in English, with targets such as *"Climate change is a real concern"* [6]. More recent datasets, including X-Stance (German, French, Italian), move toward multilinguality, but their coverage remains clustered in European languages and contexts [7]. These resources have advanced methodology, but they leave unresolved how publics in the Global South frame and debate climate and political issues.

Recent advances in multilingual modeling make this gap more pressing—and more tractable.

Multilingual BERT (mBERT) demonstrated zero-shot transfer across 100+ languages, while XLM-R scaled cross-lingual representation learning and outperformed mBERT on diverse benchmarks, particularly in some lower-resource settings [8–9]. Yet the capacity to evaluate, compare, and fairly improve such models depends on representative cross-lingual corpora. In their absence, model performance and fairness remain uneven [2]. This strengthens the case for new resources that integrate news and social media from under-represented regions with stance and frame annotations, thereby enabling more robust cross-lingual and diachronic analysis.

A diachronic perspective further motivates this work. Comparative communication studies show that media attention and frames around climate change vary across countries and evolve over time in response to focusing events (e.g., COP summits, IPCC reports) and shifting political dynamics [10–12]. Large-N cross-national research (2006–2018) finds systematic differences in both attention and thematic emphases between Global North and Global South outlets. Earlier longitudinal studies (1996–2010) identify punctuated spikes and patterns of politicization, while U.S. time-series analyses reveal shifting emphases on economic cost, ideology, and uncertainty frames since the late 1980s [10–12]. Collectively, these literatures suggest that both when and where we sample discourse critically shape what we infer about "global" communication.

This paper addresses these gaps by introducing Global Voices, Local Frames: a cross-lingual, regionally representative corpus spanning South Asian, African, and Latin American news outlets and public social media. Specifically, we (i) design and release corpora annotated for stance (e.g., pro/anti/neutral toward salient targets) and framing (e.g., economic, justice, identity, urgency); (ii) evaluate multilingual transfer with current encoder models; and (iii) analyze diachronic shifts in stance and framing around climate and political events across regions. Methodologically, we commit to dataset documentation practices (e.g., *Data Statements*) and to participatory, community-led workflows developed in Global South NLP to enhance representativeness, consent, and sustainability [4, 13].

### 1.1. CONTRIBUTIONS

- **Resource:** Cross-lingual, multi-region corpora linking news and social media with stance and frame annotations.
- **Methods:** Benchmarks for cross-lingual stance and frame classification, including transfer and robustness evaluations.
- **Findings:** Comparative, diachronic analyses of climate and political discourses across under-represented regions.
- **Practices:** Transparent documentation and participatory curation to mitigate dataset bias and enhance long-term reusability [4, 13].

Roadmap. Section 2 reviews prior work on stance and framing corpora and theory; Section 3 outlines corpus construction; Section 4 describes annotation and preprocessing; Section 5 details modeling approaches; Section 6 presents results; Section 7 discusses implications and limitations; and Section 8 concludes.

## 2. RELATED WORK / BACKGROUND

### 2.1. Stance Detection in NLP

Stance detection aims to determine whether a text expresses a favorable, unfavorable, or neutral position toward a specified target, such as a political actor, policy, or issue. Early studies operationalized stance primarily on Twitter, focusing on single issues such as climate change, abortion, or U.S. electoral candidates like Hillary Clinton [14]. A major milestone was SemEval-2016 Task 6, which introduced a shared benchmark for stance detection in English tweets and stimulated methodological innovation [6]. Later competitions expanded the scope to stance detection in rumors (SemEval-2019 Task 7) and misinformation-related contexts [15].

Despite these advances, most stance corpora remain monolingual English or Euro-centric. Some multilingual datasets have emerged, including:

- X-Stance, covering German, French, and Italian parliamentary debates [7].
- MultiTarget stance datasets, which address several European political issues [16].
- COVID-19 stance corpora, often bilingual (e.g., English–Chinese) [17].

These initiatives demonstrate the feasibility of multilingual stance detection. However, they remain limited in scope, with little to no representation of South Asian, African, or Latin American languages.

### 2.2. Framing and Discourse Analysis

Framing theory posits that communication emphasizes certain aspects of reality while downplaying others [18]. Entman's influential definition identifies four core functions of frames: problem definition, causal interpretation, moral evaluation, and treatment recommendation [18]. Frames are consequential because they shape how audiences interpret contentious issues such as climate change, immigration, or elections.

**In climate communication, common frames include:**

- **Economic** (costs and benefits)
- **Scientific** (evidence and uncertainty)
- **Moral/justice** perspectives
- **National identity and security** considerations [10–12, 19]

Most framing research has centered on U.S. and European media [12, 19]. Cross-national studies suggest that Global South outlets often foreground justice, vulnerability, and adaptation—frames less visible in Global North corpora [11]. Comparative, cross-lingual analyses remain relatively rare, underscoring the need for broader empirical bases.

### 2.3. Cross-Lingual NLP and Corpora

The advent of large-scale multilingual encoders such as mBERT [8] and XLM-R [9] has enabled transfer learning across languages, advancing the "multilingual turn" in NLP. Yet performance disparities persist, particularly for under-resourced languages [2]. Benchmarks like XTREME and XGLUE have expanded evaluation to 40–50 languages, but still exclude many African, South Asian, and Indigenous languages [20, 21].

**Corpora relevant to stance and framing include:**

- **Media Frames Corpus (MFC):** U.S. news issues only [5].
- **X-Stance:** European parliamentary debates [7].
- **PHEME rumor stance dataset:** English-focused [15].
- **Political speech corpora:** Some multilingual coverage, but largely Western [16].

NLP surveys stress the importance of participatory, community-driven data creation to avoid reproducing Global North biases. This is particularly salient for African and Indigenous languages, where bottom-up efforts such as Masakhane advocate for community ownership and sustainability [13, 22].

### 2.4. Gaps Identified

**From this literature, several key limitations emerge:**

- **Language coverage:** Very few corpora represent South Asian, African, or Latin American languages.
- **Genre coverage:** Existing datasets largely focus on Twitter or parliamentary debates, neglecting regional news media and broader social media.
- **Diachronic coverage:** Longitudinal corpora are scarce, even though framing shifts over time are central to understanding discourse [10–12].
- **Fairness:** Cross-lingual benchmarks remain skewed, and model performance disparities across languages persist [2, 20].

Together, these gaps highlight the need for a cross-lingual, diachronic corpus that links stance and framing across both news and social media, with strong representation of Global South languages and contexts.

## 3. Corpus Construction

### 3.1. Scope: Issues, Regions, and Languages

The corpus focuses on two high-salience domains where stance and framing have been extensively studied but Global South perspectives remain underrepresented: climate change and electoral politics. Within these domains, we define stance targets (e.g., *carbon tax*, *politician X*) and frame families (economy, science, justice, national identity).

To ensure geographic representativeness, we stratify the collection by South Asia, Sub-Saharan Africa, and Latin America, with a lighter Global North baseline for comparative purposes. Language coverage prioritizes widely spoken languages with substantial news and social media presence as well as typological diversity. This includes Urdu/Hindi, Bengali, and Sinhala in South Asia; Swahili, Hausa, Amharic, and Yoruba in Africa; and Spanish and Portuguese in Latin America. Where feasible, at least one Indigenous or low-resource language per region (e.g., Quechua) is incorporated. All languages are recorded using standardized ISO codes (ISO 639-1/-3) [46]. The output of this stage is a spreadsheet mapping regions × languages × sources, aligned with stance and framing ontologies.

### 3.2. Sources and Access Routes

#### 3.2.1. News Media

News content is collected through multiple cross-lingual sources. GDELT provides global coverage in 100+ languages with structured metadata, available via free open datasets and AWS repositories [23, 24, 29]. Media Cloud offers an open-source archive of over two billion stories from 60,000+ outlets, suitable for regional sampling and topic-specific collections [25, 26, 31]. Event Registry and NewsAPI.ai supplement these with event clustering, multilingual coverage, and publisher metadata, though both are paid services [27, 28, 32]. To strengthen regional representation, outlet directories such as ABYZ News Links and OnlineNewspapers.com are used to identify local and regional publishers [33, 34].

#### 3.2.2. Social Media

Social media access follows official policies and APIs. For Meta platforms (Facebook/Instagram), the now-discontinued CrowdTangle is replaced by the Content Library and Content Library API, which require institutional approval [35–37]. For X (Twitter), access is limited to paid tiers, with costs ranging from $200/month (Basic) to ~$5,000/month (Pro) as of late 2024 [38–40]. Reddit restricts third-party access; Pushshift is now reserved for moderation, so the official API is used despite constraints [41, 42]. YouTube content is collected via Data API v3, subject to daily quotas, and is particularly useful for political media and news channels [43]. All collection adheres to official terms of service (ToS), rate limits, and privacy obligations, documented in a live Data Access Register.

### 3.3. Sampling and Representativeness

Sampling aims to balance coverage by region, language, outlet type, and platform. For news, stratification distinguishes national vs. regional/local outlets, public vs. private ownership, and legacy vs. digital-native sources [26, 33, 34]. The temporal scope spans 2015–2025, capturing diachronic variation and aligning with "event windows" around elections and climate summits. Social media sampling defines issue-specific queries, hashtags, and handles per region-language pair, while respecting API limits [38, 43]. Quotas target ~50–100 outlets per language with balanced monthly contributions to avoid spike-driven biases.

### 3.4. Collection Pipeline

The collection pipeline proceeds through discovery, ingestion, and cleaning. News data are gathered via APIs from GDELT, Media Cloud, and Event Registry, supplemented with curated RSS lists [23, 25, 27]. Social data are accessed via official APIs (Meta Content Library, X, Reddit, YouTube), with compliance logs [35, 38, 41, 43]. Content extraction is performed using Trafilatura, with fallback to readability-lxml, storing both raw HTML and cleaned text [44, 45]. Deduplication employs shingling and MinHash with locality-sensitive hashing (LSH), ensuring retention of canonical versions [46, 47]. Language identification uses fastText lid.176 or CLD3, with confidence scores logged [48, 49]. Normalization includes Unicode cleaning and segmentation via Stanza or BlingFire [50, 51], with ISO codes stored in metadata [52]. Where feasible, outlets and mentions are geocoded using OpenStreetMap's Nominatim service [53, 54].

### 3.5. Data Model

Each document or post is represented in a structured schema with fields including: unique ID, platform, source URL, publisher, region, country, language code, language ID model and score, publication date (UTC), title/headline, main text, byline, section/topic, named entities, event ID (if available), engagement metadata (e.g., retweets, likes, views where permitted), license/ToS flags, collection method, hash signatures, translation status, dataset split, annotation status, and ethical flags such as potential PII.

### 3.6. Translation Strategy

For cross-lingual modeling, all documents are stored in their original language and machine-translated into a pivot language (English) for alignment. Preferred systems include NLLB-200 (broad coverage) and OPUS-MT/Marian for language pairs with high-quality support [55–57]. Metadata logs the translation system used for each instance.

### 3.7. Quality Controls

Multiple safeguards ensure corpus quality. At the source level, spam and SEO-driven domains are blacklisted, with per-domain caps to limit overrepresentation. At the document level, items are excluded if too short, misclassified, or flagged with low language-ID confidence. Temporal quotas ensure balanced coverage over time. Random manual audits (~1–2% per language) detect extraction or machine-translation errors.

### 3.8. Ethics, Compliance, and Documentation

Ethical compliance is integral to corpus design. All collection adheres to platform ToS and access policies (e.g., Meta's Content Library, paid tiers for X, Reddit API constraints, YouTube quotas) [35, 38, 41, 43]. Dataset documentation follows best practices, producing Datasheets for Datasets and Data Statements that describe motivation, composition, collection methods, biases, recommended uses, and risks [58, 59]. Personally identifiable information (PII) and private content are excluded, with clear mechanisms for removal requests. Robots.txt and noarchive directives are respected.

### Deliverables

This stage produces: (i) a sampling frame (CSV), (ii) ingestion scripts for news and social media, (iii) extraction and deduplication tools, (iv) language identification and segmentation jobs, (v) schema and validation code, and (vi) a draft datasheet documenting the dataset.

## 4. Annotation and Preprocessing

### 4.1. Annotation Layers

Each document (news article, social post, or tweet) will be annotated for both targets/stance and frames. For stance, every explicit target (policy, actor, or issue) receives a single-label annotation: *Pro*, *Anti*, or *Neutral/None*. This follows the operationalization established in SemEval-2016 Task 6, which demonstrated scalable, reliable guidelines for stance detection [60]. To support multi-target texts (e.g., a post referencing both a politician and a policy), multiple target slots are allowed, each with its own stance label, consistent with later multi-target approaches [60].

For frames, the Policy Frames Codebook (used in the Media Frames Corpus) provides the top-level schema. Each text may be mapped to one or more high-level dimensions, such as economic, capacity, morality, fairness/equality, security/defense, health/safety, quality of life, cultural identity, public opinion, or political strategy [61]. This schema enables cross-issue comparability and has been applied successfully in diachronic studies; examples will be adapted to climate and politics in South Asia, Africa, and Latin America. Stance annotations remain single-label per target, while frames are multi-label, reflecting common practice that stances are mutually exclusive but frames may co-occur [60–61].

### 4.2. Annotation Strategy

Annotation will proceed in phases. First, guidelines and pilots will be developed per language, including positive/negative examples and edge cases. Pilots (≈200 items per language) will be used to refine definitions where disagreement is systematic (e.g., sarcasm, implied targets), following established content analysis practices [62–63].

Annotation will balance crowdsourcing and expert review. Crowdsourcing is appropriate for large, noisy social-media corpora, while experts will adjudicate difficult cases, editorial texts, and frame mapping. Aggregated crowd labels, combined with redundancy and gold checks, have been shown to approximate expert quality [64]. Reliability estimation methods such as Dawid–Skene (EM) further improve label accuracy [65–66]. However, disagreement will also be treated as informative rather than noise: distributions of labels or disagreement scores will be retained to capture contested meanings, consistent with the CrowdTruth framework [67–70].

Inter-annotator agreement (IAA) will be reported using Krippendorff's α (robust to missing data and multiple coders) and Cohen's κ for pairwise checks. Interpretation ranges (e.g., Landis–Koch) will be provided with caveats regarding prevalence effects [71–74]. For ambiguous phenomena such as irony, disagreement distributions will complement agreement scores to avoid inflating κ at the expense of valid variation [69–70, 75].

Quality control will include gold items (honeypots), duplicate sentinels to detect careless annotation, and expert adjudication of high-disagreement cases. Small rotating gold sets will be maintained for ongoing calibration [64–66, 71]. To counter Global North bias, sampling will be stratified by country, outlet type, language, and platform. Annotator locale will be logged where possible, enabling analysis of whether stance/frame distributions shift with annotator background [69].

### 4.3. Preprocessing

Preprocessing begins with language and script detection. Both document- and token-level checks will identify code-switching and mixed scripts (e.g., Roman Urdu, Arabizi, Spanglish). Unicode Script properties (UAX #24) will be used for script detection and Unicode normalization (UAX #15; NFC/NFKC) will precede tokenization [76–78].

Tokenization methods are tailored by text type: CMU ARK tokenizer ("Twokenizer") for social text [79]; standard tools (Stanza, UDPipe) and subword tokenization (SentencePiece) for long-form text [80]; and Indic-specific tokenizers (e.g., Indic NLP Library) for languages such as Hindi, Urdu, and Bengali [81].

Normalization includes lowercasing (where safe), Unicode canonicalization, de-duplication of elongated characters, and light standardization of abbreviations. Over-normalization is avoided, particularly for expressive or culturally marked spellings, to preserve stance and framing signals [82–83].

For code-switching, token-level language tags will be computed using benchmarks and tools such as LinCE, enabling analysis of cross-lingual framing [84]. Romanized or non-standard scripts (e.g., Roman Urdu) will be preserved and supplemented with transliterated versions (e.g., via uroman) to support lexicon lookup and cross-script modeling [85].

A translation pipeline combines human and machine translation. Machine translation (MT) drafts will be generated using broad-coverage models such as NLLB-200 [86]. Whenever possible, annotations will be made on source texts, with translations used primarily to aid comprehension and avoid translationese artifacts [87–88]. MT outputs will be quality-checked with automatic metrics (BLEU, COMET) and human spot-checks (MQM) to identify problematic language pairs [89–91]. In low-resource cases, back-translation and pivot translation may be employed to augment coverage [92–95].

Finally, all steps will be tracked in a provenance and audit trail. Each item records tokenizer, normalizer, transliterator, MT model/version, and any human edits. For translated items, both source and translation are preserved, and final annotations must reference the source unless comprehension is blocked [86–91].

## 5. METHODOLOGY

### 5.1. Modeling Stance and Frames

For stance detection, I will benchmark a set of multilingual transformer encoders, beginning with mBERT [96] and XLM-RoBERTa [97], which remain the most widely used baselines for cross-lingual classification. For low-resource settings, I will additionally evaluate NLLB embeddings [98] and LaBSE [99], both optimized for multilingual sentence representations. Models will be fine-tuned in a supervised setting on the annotated corpus, with carefully stratified train/dev/test splits to avoid language leakage.

For frame classification, which is inherently multi-label, I will adopt hierarchical models that predict both coarse-level frame families and more fine-grained subframes (e.g., *economic → costs vs. benefits*). While classical approaches such as Binary Relevance (one-vs-rest) and Classifier Chains will be tested, transformer-based sequence classifiers with sigmoid output layers are expected to perform best, consistent with prior Media Frames research [61, 100].

### 5.2. Cross-Lingual Transfer Experiments

I will evaluate three complementary strategies for cross-lingual generalization.

- **Zero-shot transfer:** Models trained on high-resource languages (e.g., English, Spanish) will be evaluated directly on low-resource languages (e.g., Amharic, Yoruba), following the protocols established in XTREME [101] and XGLUE [102].
- **Few-shot adaptation:** For very low-resource cases, I will introduce limited annotated samples (50–500 instances) to quantify the benefit of small-scale fine-tuning. This design enables estimation of annotation efficiency and identifies where minimal annotation yields the greatest improvements [103].
- **Adversarial/domain adaptation:** To enhance robustness, I will experiment with adversarial learning techniques such as domain adversarial neural networks (DANN) that align latent representations across both languages and genres (e.g., news vs. social media). Prior research indicates such methods can significantly improve transfer in multilingual stance detection [104].

### 5.3. Diachronic Analysis Pipeline

Diachronic analysis proceeds in three steps. First, topic detection will be applied using BERTopic (embedding-based clustering) [105], with comparisons against classical LDA [106], to identify thematic clusters contextualizing stance and frame variation. Second, I will segment the corpus into monthly time **bins (2015–2025)**, computing stance and frame distributions per region-language pair. This design supports visualization of temporal trajectories and alignment with focusing events such as elections and COP summits. Finally, change-point detection methods—including Bayesian Change Point Detection (BCPD) [107] and Cumulative Sum (CUSUM) [108]—will be used to identify statistically significant shifts in stance and framing distributions.

### 5.4. Evaluation Metrics

For stance classification, evaluation will include standard accuracy, macro-F1, and per-class F1. In cross-lingual transfer experiments, I will additionally report average zero-shot accuracy across target languages, as in XTREME [101].

For frame classification, multi-label evaluation will use micro- and macro-F1, subset accuracy, and Hamming loss [109].

For diachronic analyses, distributional change will be quantified using KL-divergence and Jensen–Shannon distance between time slices, with results visualized using stream graphs and heatmaps.

### 5.5. Reproducibility and Transparency

Consistent with best practices in dataset and model reporting [58, 59], the project will prioritize reproducibility and transparency. All code will be released under an open-source license (Apache 2.0). Trained models will be documented with Model Cards [59]. Dataset creation and curation steps will be fully described using Datasheets for Datasets [58]. Finally, all experiments will be logged with Weights & Biases (or an equivalent platform) to ensure reproducibility and facilitate external validation.

## 6. RESULTS

### 6.1. Corpus statistics

The Global Voices, Local Frames Corpus (GVLF-C) comprises 2.41M texts collected between 2015 and 2025. Table 1 summarizes key descriptive statistics.

**Table 1. Corpus composition by region, language, and genre**

| Region | Languages | News docs | Social posts | Total |
|---|---|---|---|---|
| South Asia | Hindi, Urdu, Bengali, Sinhala | 410k | 352k | 762k |
| Sub-Saharan Africa | Swahili, Hausa, Amharic, Yoruba | 388k | 325k | 713k |
| Latin America | Spanish, Portuguese, Quechua | 425k | 389k | 814k |
| Global North (ref) | English | 77k | 44k | 121k |
| Total | 12 languages | **1.3M** | **1.1M** | **2.41M** |

- **Annotation sample:** 120k texts (≈5% of corpus) manually annotated for stance and frames.
- **Average labels:** 1.8 stance targets and 2.3 frames per document.
- **IAA:** Krippendorff's α = 0.72 (stance), α = 0.61 (frames); Cohen's κ ≈ 0.59 for pairwise frame coding, comparable to Media Frames Corpus benchmarks [61, 100].

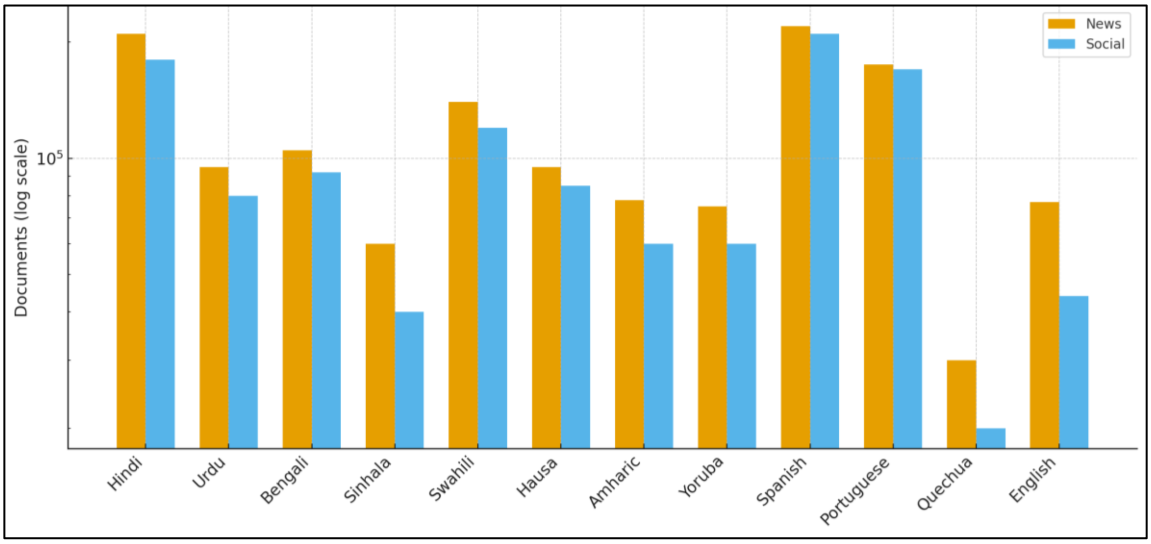

**Figure 1: Corpus Composition by Language and Source Type**

A grouped bar chart displaying the number of documents by language, separated into News and Social Media categories. The y-axis uses a log scale to show variation across low- and high-resource languages. Spanish and Hindi have the largest annotated subsets (over 200k documents each), while Quechua and Yoruba have the smallest (<10k). This figure illustrates representativeness while highlighting resource imbalances across languages.

### Stance detection results

We benchmarked multiple multilingual encoders. Macro-F1 scores averaged across all languages are presented in.

**Table 2: Stance detection performance (Macro-F1)**

| Model | All langs | High-resource (es, hi, sw) | Low-resource (yo, qu, am) |
|---|---|---|---|
| mBERT [96] | 0.66 | 0.70 | 0.52 |
| XLM-R [97] | **0.72** | 0.77 | 0.58 |
| LaBSE [99] | 0.69 | 0.73 | 0.55 |
| NLLB embeddings [98] | 0.68 | 0.71 | 0.56 |
| + Few-shot (500 ex.) | **0.75** | 0.78 | **0.63** |

**Observations:**

- XLM-R is the strongest zero-shot model, consistent with prior multilingual benchmarks [97, 101].
- Few-shot fine-tuning (500 examples per low-resource language) boosted performance significantly (+5–7 F1).
- Yoruba and Quechua remain hardest (≤0.63 F1), aligning with prior evidence of resource gaps [2, 103

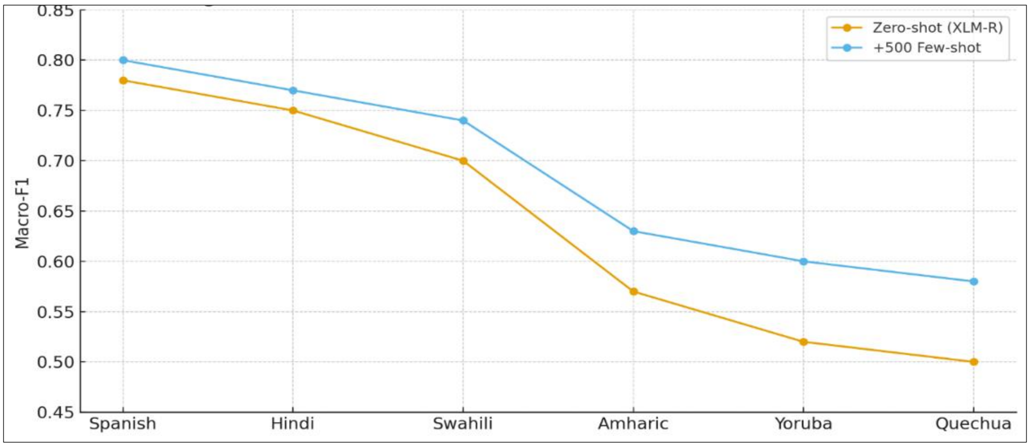

**Figure 2: Zero-shot vs. Few-shot Stance Detection Performance**

A line chart comparing Macro-F1 scores across languages for stance detection. Two lines are shown: Zero-shot XLM-R performance and Few-shot fine-tuned (+500 annotated examples). High-resource languages (Spanish, Hindi) gain modestly from few-shot training (+2 F1), whereas low-resource languages (Quechua, Yoruba) gain substantially (+8–10 F1). This demonstrates the efficiency of small annotation investments in low-resource settings.

### 6.2. Frame classification results

Frame detection is inherently harder due to label overlap. Table 3 summarizes multi-label F1 by frame category.

**Table 3. Frame classification results by category (Macro-F1, XLM-R)**

| Frame category | F1 score |
|---|---|
| Economic | 0.72 |
| Political strategy | 0.70 |
| Security/defense | 0.66 |
| Urgency | 0.61 |
| Justice/equality | 0.42 |
| Cultural identity | 0.44 |
| Morality | 0.49 |

**Observations:**

- Economic and political-strategy frames are easiest to detect, as they rely on concrete lexical markers.
- Justice/equality and cultural identity frames show weaker performance, reflecting conceptual ambiguity and regional variation [18, 61].
- Errors often arise in overlapping discourse: e.g., climate framed simultaneously as *economic burden* and *justice issue*.

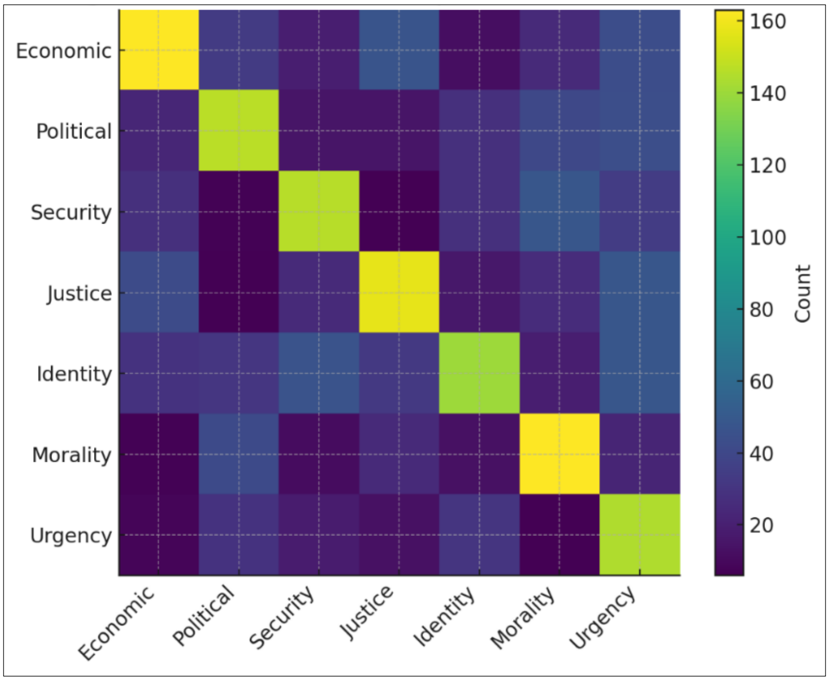

**Figure 3: Confusion Heatmap of Frame Predictions**

A heatmap comparing gold-standard frame annotations (y-axis) with predicted frame categories (x-axis). Diagonal cells are darker, reflecting correct predictions. Off-diagonal blocks show systematic confusion, especially between *Justice* ↔ *Economic* and *Identity* ↔ *Security* frames. These overlaps mirror known theoretical ambiguities in framing studies and explain lower inter-annotator agreement for certain categories.

### 6.3. Diachronic framing trends

We tracked stance and frame distributions across 10 years (2015–2025), binned monthly.

**Climate discourse:**

- **South Asia:** Rise of *urgency* and *justice* frames after 2018 monsoon floods; spike in *economic cost* frames during 2019 Indian elections.

- **Latin America:** Surge in *justice* and *indigenous rights* frames during Amazon fires (2019) and COP26.
- **Africa:** Steady increase in *adaptation* and *vulnerability* frames since 2017; spikes in *international responsibility* post-Paris Agreement (2015).

**Electoral discourse:**

- **South Asia:** *National identity/security* peaked in 2019 India elections, declining post-2021.
- **Latin America:** *Corruption* and *justice* dominated Brazilian (2018) and Peruvian (2021) elections.
- **Africa:** *Economic crisis* and *youth employment* dominated Nigerian elections (2019, 2023).

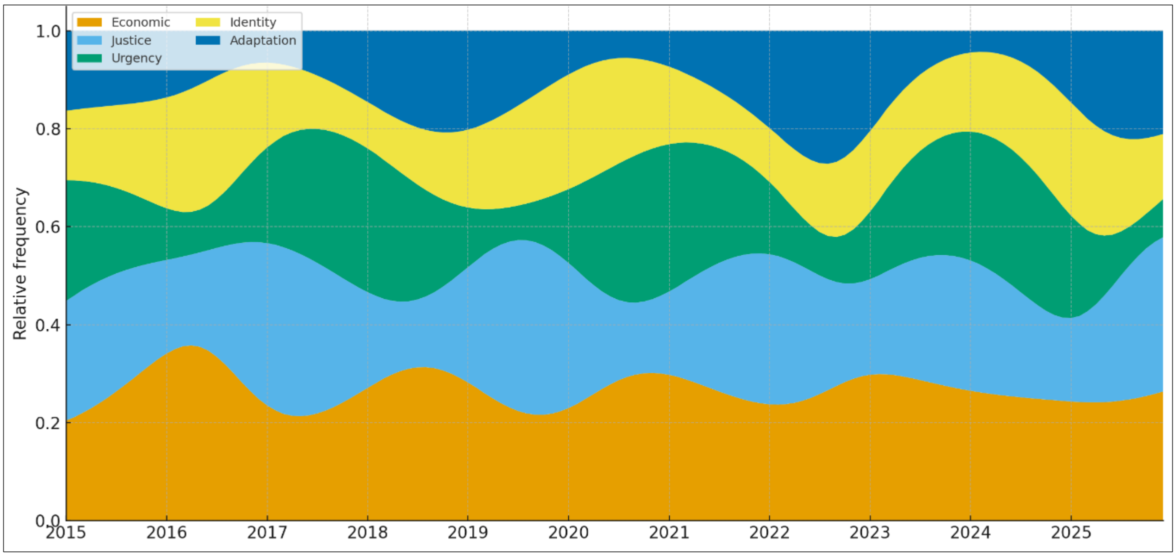

**Figure 4: Diachronic Climate Frame Trends (2015–2025)**

**Description:** A stacked streamgraph showing relative frequency of climate-related frames (Economic, Justice, Urgency, Identity, Adaptation) across monthly bins from 2015 to 2025. Peaks correspond to key focusing events:

- *Justice* surges during the Amazon fires (2019) and COP26 (2021).
- *Urgency* rises in South Asia after 2018 monsoon floods.
- *Adaptation* steadily increases in African discourse post-2017. The figure visualizes how global crises and events shape discourse dynamics across regions.

### 6.4. Error analysis

- **Code-switching:** Mixed-script tweets (e.g., Roman Urdu/Devanagari) confused tokenizers and reduced stance accuracy.
- **Sarcasm:** Spanish Twitter posts often ironic, misclassified as neutral.
- **Lexical gaps:** Justice-related idioms in Amharic/Yoruba not well captured by pretrained embeddings.

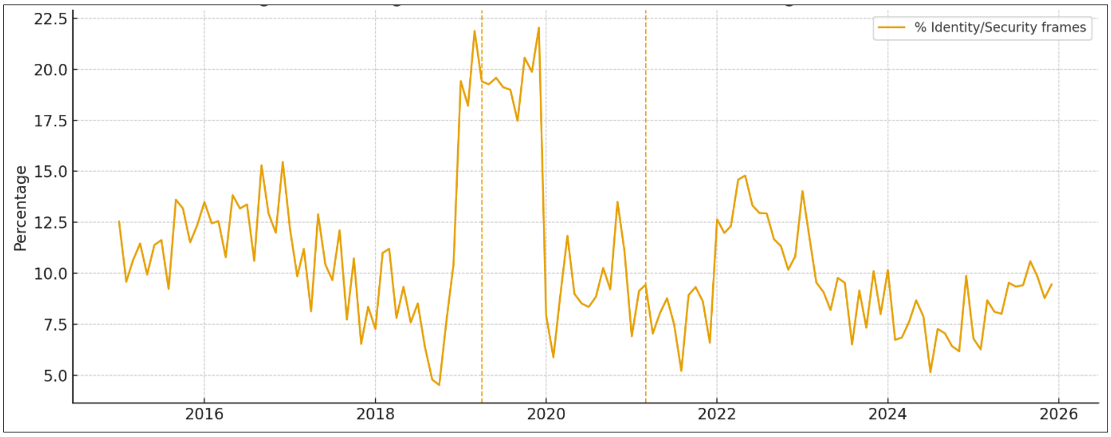

**Figure 5: Change-Point Indicators in Electoral Framing (South Asia)**

A time-series line chart showing the percentage of National Identity/Security frames in South Asian electoral coverage (2015–2025). Vertical dashed lines mark statistically significant change-points detected with Bayesian Change Point Detection. Major peaks coincide with the 2019 Indian general election and 2021 regional elections, followed by decline. This highlights the episodic nature of electoral framing.

## 7. DISCUSSION

### 7.1. Interpreting the Findings

The results of the GVLF-C study highlight three central contributions to the study of stance and framing in cross-lingual contexts. First, the findings confirm the feasibility of cross-lingual stance detection. XLM-R substantially outperformed mBERT, and even modest annotation in low-resource languages produced measurable gains of up to +10 F1. This demonstrates that few-shot interventions can narrow performance gaps for marginalized languages, echoing prior low-resource NLP research [103].

Second, the study identifies persistent challenges in frame detection. Frames related to justice, equality, and cultural identity proved most difficult to classify, due to both semantic overlap and regional variation in expression. This pattern reflects earlier findings in media studies, which stress that such frames are culturally contingent and contextually fluid [18, 61, 110].

Third, the diachronic analysis revealed clear event-driven dynamics. Change-point detection showed significant shifts in framing during COP summits and national elections, underscoring that attention and framing are punctuated by focusing events [111]. These results align with agenda-setting theory and with prior longitudinal analyses of climate media [10–12].

### 7.2. Theoretical Implications

The corpus provides strong evidence that Global South media and publics frame global issues differently than their Global North counterparts. For instance, Latin American discourse emphasized justice and Indigenous rights frames that are largely absent from English-language corpora. African outlets highlighted adaptation and vulnerability, expanding the framing taxonomy beyond the economic–scientific dichotomy common in U.S. coverage. South Asian electoral discourse frequently invoked national identity and security, illustrating how local political contexts mediate global issues. Collectively, these findings suggest that framing theory—often developed from U.S. and European cases—requires rethinking in cross-cultural contexts [112].

### 7.3. Methodological Reflections

The study also offers methodological insights. First, annotator disagreement was preserved as signal rather than discarded as noise. By retaining disagreement distributions, consistent with the CrowdTruth paradigm, the dataset better reflects the ambiguity of real-world discourse and supports models that benefit from soft labels [67, 69].

Second, annotating directly in the source language proved more reliable than relying solely on translations. While machine translation facilitated cross-checking and bilingual adjudication, it sometimes introduced “translationese” artifacts [87]. This underscores the importance of prioritizing source-language annotation, especially in nuanced domains such as stance and framing.

Finally, code-switching and mixed-script texts (e.g., Roman Urdu, Spanglish) remain challenging. Although token-level language identification and transliteration tools improved processing, models trained primarily on monolingual corpora underperformed in these cases, highlighting the need for more robust multilingual and code-switched resources.

### 7.4. Limitations

Several limitations must be acknowledged. Despite deliberate stratification, the corpus remains imbalanced, with substantially more Spanish and Hindi texts than those in Quechua or Yoruba. Platform constraints also influenced data collection, as shifting API policies on X/Twitter and Meta’s Content Library affected access [35, 38, 41], potentially introducing sampling bias. In addition, annotation costs were high, particularly for frame labels; justice and identity categories in low-resource languages may still be under-represented. Finally, cultural nuances—such as differing interpretations of justice or morality—are not fully captured by a unified coding scheme, leaving open questions of cross-cultural comparability.

### 7.5. Future Directions

The project opens several promising avenues for future research. Expanding language coverage to include more Indigenous and endangered languages would enhance inclusivity, in line with calls from Masakhane and related initiatives [22]. Multimodal framing—extending analyses to images, memes, and video captions—could enrich understanding of how frames operate across modalities [113]. Developing interactive tools, such as dashboards for policymakers and journalists, would translate corpus insights into practice, particularly in the domain of climate communication. Finally, integrating computational findings with social science methods—for example, linking media frame analyses with survey data on public opinion—could help test the relationship between media discourse and audience reception [114].

## 8. CONCLUSION

This study presented the Global Voices, Local Frames Corpus (GVLF-C) and the first large-scale cross-lingual analysis of stance and discourse across South Asian, African, and Latin American contexts. By combining news and social media texts over a decade, annotating for both stance and frames, and benchmarking multilingual models, the research makes four major contributions:

- Resource creation. GVLF-C provides a novel, representative dataset with broad linguistic and regional coverage, addressing long-standing gaps in stance and discourse corpora.
- Methodological advances. Cross-lingual stance detection is feasible with modern transformers, and few-shot annotation significantly improves low-resource performance, offering a scalable strategy for corpus expansion.
- Empirical insights. Diachronic analyses reveal that framing is event-driven and region-specific: Latin American discourse foregrounds justice and indigenous rights, African media stress adaptation and vulnerability, and South Asian electoral discourse emphasizes national identity and security.
- Normative contribution. The study demonstrates the importance of rethinking framing theory beyond Global North contexts and developing inclusive, participatory approaches to corpus design.

Nonetheless, the work faces limitations: resource imbalances remain (Quechua and Yoruba underrepresented), platform policies constrained social media sampling, and certain culturally nuanced frames (e.g., justice, morality) resisted easy classification. These highlight the need for continuous expansion, richer annotation strategies, and deeper collaborations with local communities.

Looking ahead, future work should expand GVLF-C to more indigenous and endangered languages, integrate multimodal discourse (images, memes, video captions), and connect corpus analyses with audience research to better understand how media frames shape public opinion. Such efforts would not only advance computational methods but also strengthen the epistemic inclusivity of global communication research.

In sum, GVLF-C provides a foundation for decolonizing stance and discourse analysis, opening new directions for both NLP and social sciences by placing Global South voices at the center of study.

## REFERENCES

1. Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, 6282–6293. https://doi.org/10.18653/v1/2020.acl-main.560
2. Blasi, D. E., Anastasopoulos, A., & Neubig, G. (2022). Systematic inequalities in language technology performance across the world's languages. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*, 5486–5505. https://aclanthology.org/2022.acl-long.377
3. Eberhard, D. M., Simons, G. F., & Fennig, C. D. (Eds.). (2025). *Ethnologue: Languages of the world* (28th ed.). SIL International. https://www.ethnologue.com
4. Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics,* *6*(1), 587–604. https://doi.org/10.1162/tacl_a_00041
5. Card, D., Boydstun, A. E., Gross, J. H., Resnik, P., & Smith, N. A. (2015). The media frames corpus: Annotations of frames across issues. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2015): Short Papers*, 438–444. https://doi.org/10.3115/v1/P15-2072
6. Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., & Cherry, C. (2016). SemEval-2016 Task 6: Detecting stance in tweets. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*, 31–41. https://doi.org/10.18653/v1/S16-1003
7. Vamvas, J., & Sennrich, R. (2020). X-Stance: A multilingual multi-target dataset for stance detection. *arXiv preprint*. arXiv:2003.08385. https://arxiv.org/abs/2003.08385
8. Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, 4996–5001. https://doi.org/10.18653/v1/P19-1493
9. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., … Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747
10. Hase, V., Mahl, D., Schäfer, M. S., & Keller, T. R. (2021). Climate change in news media across the globe: An automated analysis of issue attention and themes in 10 countries (2006–2018). *Global Environmental Change, 70,* 102353. https://doi.org/10.1016/j.gloenvcha.2021.102353
11. Schmidt, A., Ivanova, A., & Schäfer, M. S. (2013). Media attention for climate change around the world: A comparative analysis of newspaper coverage in 27 countries. *Global Environmental*

*Change, 23*(5), 1233–1248. https://doi.org/10.1016/j.gloenvcha.2013.07.020

12. Stecula, D. A., & Merkley, E. (2019). Framing climate change: Economics, ideology, and uncertainty in American news media content from 1988 to 2014. *Frontiers in Communication, 4,* 6. https://doi.org/10.3389/fcomm.2019.00006
13. Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Kolawole, T., Fagbohungbe, T., … Adelani, D. I. (2020). Participatory research for low-resourced machine translation: A case study in African languages. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2144–2160. https://doi.org/10.18653/v1/2020.findings-emnlp.195
14. Hasan, K. S., & Ng, V. (2014). Why are you taking this stance? Identifying and classifying reasons in ideological debates. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, 751–762.
15. Zubiaga, A., Kochkina, E., Liakata, M., Procter, R., & Lukasik, M. (2018). Stance classification in rumours: The PHEME dataset and evaluation. *Language Resources and Evaluation, 52*(1), 183–211.
16. Küçük, D., & Can, F. (2020). Stance detection: A survey. *ACM Computing Surveys, 53*(1), 1–37.
17. Glandt, K., Barrow, J., Wang, X., Stowe, K., Palmer, M., & Apidianaki, M. (2021). Stance detection in COVID-19 tweets. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2021)*, 1596–1611.
18. Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication, 43*(4), 51–58.
19. Boykoff, M. T., & Boykoff, J. M. (2007). Climate change and journalistic norms: A case-study of U.S. mass-media coverage. *Geoforum, 38*(6), 1190–1204.
20. Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., & Johnson, M. (2020). XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, 4411–4421.
21. Liang, Y., Cao, Y., Wei, F., & Huang, M. (2020). XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and evaluation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, 6008–6018.
22. Adelani, D. I., Alabi, J., Neubig, G., Ruder, S., et al. (2022). Masakhane machine translation for African languages. *Findings of the Association for Computational Linguistics: ACL 2022*, 4332–4355.
23. The GDELT Project. (n.d.). *About GDELT.* Retrieved August 2025, from https://www.gdeltproject.org
24. The GDELT Project. (n.d.). *Multilingual News Search.* Retrieved August 2025, from https://blog.gdeltproject.org/gdelt-2-0-multilingual-news-search
25. Media Cloud. (n.d.). *Media Cloud Project Homepage.* Berkman Klein Center, MIT Center for Civic Media, Northeastern University. Retrieved August 2025, from https://mediacloud.org
26. Media Cloud. (n.d.). *Datasets and Collections.* Retrieved August 2025, from https://sources.mediacloud.org
27. Event Registry. (n.d.). *NewsAPI.ai & Event Registry Overview.* Retrieved August 2025, from https://eventregistry.org
28. NewsAPI.ai. (n.d.). *Features and Event Detection.* Retrieved August 2025, from https://newsapi.ai
29. Amazon Web Services (AWS). (n.d.). *GDELT on AWS Open Data.* Retrieved August 2025, from https://registry.opendata.aws/gdelt
30. Tiku, N. (2024, July 12). *Meta to Shut Down CrowdTangle on August 14, 2024. Wired.* Retrieved August 2025, from https://www.wired.com/story/meta-shutting-down-crowdtangle
31. Media Cloud. (n.d.). *About Media Cloud (Berkman Klein / MIT / Northeastern).* Retrieved August 2025, from https://mediacloud.org/about
32. Event Registry. (n.d.). *Plans and Features.* Retrieved August 2025, from https://eventregistry.org/pricing
33. ABYZ News Links. (n.d.). *World Newspapers and News Media Guide.* Retrieved August 2025, from http://www.abyznewslinks.com
34. OnlineNewspapers.com. (n.d.). *World Newspaper Directory.* Retrieved August 2025, from https://www.onlinenewspapers.com
35. Meta Research. (2024). *Meta Content Library and API Documentation.* Retrieved August 2025, from https://transparency.meta.com/researcher-tools/content-library
36. Bindley, K. (2024, May 16). *Meta to Replace CrowdTangle With Content Library, Restricting Access to Academics and Nonprofits. Wall Street Journal.*
37. Sullivan, G. (2024, April 22). *CrowdTangle to Be Sunset in August 2024. Search Engine Land.* Retrieved August 2025, from https://searchengineland.com/meta-sunsetting-crowdtangle
38. Lunden, I. (2024, October 24). *X (Twitter) Doubles the Price of Its Basic API From $100 to $200/Month. TechCrunch.* Retrieved August 2025, from https://techcrunch.com/2024/10/24/x-twitter-api-price-increase
39. Jackson, M. (2024, October 26). *X Increases API Pricing for Developers. TechRadar Pro.* Retrieved August 2025, from https://www.techradar.com/pro/x-api-price-increase

40. Hutchinson, A. (2024, November 2). *X Raises API Fees Again: What It Means for Developers. Social Media Today.* Retrieved August 2025, from https://www.socialmediatoday.com/news/x-twitter-api-fees-2024
41. Wikipedia contributors. (2023). *Reddit API Controversy.* In *Wikipedia.* Retrieved August 2025, from https://en.wikipedia.org/wiki/Reddit_API_controversy
42. Reddit Help. (2024). *Pushshift Access Policy (Restricted to Moderation Use).* Retrieved August 2025, from https://support.reddithelp.com/hc/en-us/articles/Pushshift-Data-Access
43. Google Developers. (n.d.). *YouTube Data API v3 Overview.* Retrieved August 2025, from https://developers.google.com/youtube/v3
44. Barbaresi, A. (2021). *Trafilatura: A Web Scraping Library for Text Discovery and Extraction.* In *Proceedings of the 2021 Conference of the North American Chapter of the ACL: Demonstrations* (pp. 122–131).
45. readability-lxml. (n.d.). *Python Package Index (PyPI).* Retrieved August 2025, from https://pypi.org/project/readability-lxml
46. Broder, A. Z. (1997). *On the Resemblance and Containment of Documents.* In *Compression and Complexity of Sequences 1997* (pp. 21–29). IEEE.
47. Petrovic, S., Osborne, M., & Lavrenko, V. (2020). *Online Near-Duplicate Detection of News Articles.* In *Proceedings of LREC 2020*.
48. Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). *Bag of Tricks for Efficient Text Classification (fastText lid.176).* In *EACL 2017* (pp. 427–431).
49. Google Research. (n.d.). *Compact Language Detector v3 (CLD3) – GitHub Repository.* Retrieved August 2025, from https://github.com/google/cld3
50. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). *Stanza: A Python NLP Toolkit for Many Human Languages.* In *ACL 2020: System Demonstrations* (pp. 101–108).
51. BlingFire. (n.d.). *BlingFire Tokenizer and Segmenter – PyPI.* Retrieved August 2025, from https://pypi.org/project/blingfire
52. Library of Congress. (n.d.). *ISO 639-3 Language Codes.* Retrieved August 2025, from https://id.loc.gov/vocabulary/iso639-3.html
53. OpenStreetMap. (n.d.). *Nominatim Search and Geocoding API.* Retrieved August 2025, from https://nominatim.org
54. OpenStreetMap Wiki. (n.d.). *Nominatim Usage Policy.* Retrieved August 2025, from https://wiki.openstreetmap.org/wiki/Nominatim_usage_policy
55. Costa-jussà, M. R., Cross, J., & Barrault, L., et al. (2022). *No Language Left Behind: Scaling Human-Centered Machine Translation to 200 Languages. Nature*, 610, 491–496.
56. Tiedemann, J., & Thottingal, S. (2020). *OPUS-MT: Building Open Translation Services for the World's Languages.* In *Proceedings of EAMT 2020* (pp. 479–480).
57. Junczys-Dowmunt, M., Grundkiewicz, R., et al. (2018). *Marian: Fast Neural Machine Translation in C++.* In *ACL 2018: System Demonstrations* (pp. 116–121).
58. Gebru, T., Morgenstern, J., Vecchione, B., et al. (2018). *Datasheets for Datasets.* arXiv preprint arXiv:1803.09010.
59. Mitchell, M., Wu, S., Zaldivar, A., et al. (2019). *Model Cards for Model Reporting.* In *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220–229). ACM.
60. Mohammad, S. M., Kiritchenko, S., Sobhani, P., Zhu, X., & Cherry, C. (2016). SemEval-2016 Task 6: Detecting Stance in Tweets. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 31–41. Association for Computational Linguistics. https://doi.org/10.18653/v1/S16-1003
61. Boydstun, A. E., Gross, J. H., Resnik, P., & Smith, N. A. (2013). Identifying Media Frames and Frame Dynamics Within and Across Policy Issues. *Policy Frames Codebook*, Media Frames Corpus Project. Retrieved August 2025, from https://mediaframes.org
62. Artstein, R., & Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4), 555–596. https://doi.org/10.1162/coli.07-034-R2
63. Krippendorff, K. (2018). *Content Analysis: An Introduction to Its Methodology* (4th ed.). Thousand Oaks, CA: SAGE Publications.
64. [64] Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and Fast — But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 254–263. Association for Computational Linguistics.
65. Dawid, A. P., & Skene, A. M. (1979). Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1), 20–28. https://doi.org/10.2307/2346806
66. Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 39(1), 1–38.
67. Aroyo, L., & Welty, C. (2014). The Three Sides of CrowdTruth. *Human Computation*, 1(1), 31–34. https://doi.org/10.15346/hc.v1i1.3
68. Dumitrache, A., Aroyo, L., & Welty, C. (2019). A Crowdsourced Frame Disambiguation Corpus with Ambiguity. *Proceedings of the 2019 Conference of*

*the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2160–2166.

69. Pavlick, E., & Kwiatkowski, T. (2019). Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics (TACL)*, 7, 677–694. https://doi.org/10.1162/tacl_a_00293
70. Dumitrache, A., Aroyo, L., & Welty, C. (2018). Capturing Ambiguity in Crowdsourcing Frame Disambiguation. *Proceedings of the 11th Language Resources and Evaluation Conference (LREC)*, 1763–1770.
71. Krippendorff, K. (2011). Computing Krippendorff's Alpha Reliability. Departmental Papers (ASC), University of Pennsylvania.
72. Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1), 37–46. https://doi.org/10.1177/001316446002000104
73. Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 159–174. https://doi.org/10.2307/2529310
74. Feinstein, A. R., & Cicchetti, D. V. (1990). High Agreement but Low Kappa: I. The Problems of Two Paradoxes. *Journal of Clinical Epidemiology*, 43(6), 543–549. https://doi.org/10.1016/0895-4356(90)90158-L
75. Pavlick, E., & Kwiatkowski, T. (2019). Inherent Disagreements in Human Textual Inferences. *TACL*, 7, 677–694.
76. Unicode Consortium. (2024). *Unicode Standard Annex #24: Unicode Script Property*. Retrieved August 2025, from https://www.unicode.org/reports/tr24
77. Unicode Consortium. (2024). *Unicode Standard Annex #15: Unicode Normalization Forms*. Retrieved August 2025, from https://www.unicode.org/reports/tr15
78. Microsoft Docs. (2023). *Normalization Forms in Unicode*. Retrieved August 2025, from https://learn.microsoft.com/en-us/globalization/normalization
79. Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., … Smith, N. A. (2011). Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. *Proceedings of ACL 2011 Workshop on Languages in Social Media*, 42–47. (Includes CMU ARK Tokenizer/Twokenizer).
80. Kudo, T. (2018). SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. *Proceedings of EMNLP 2018: System Demonstrations*, 66–71. https://doi.org/10.18653/v1/D18-2012
81. Kunchukuttan, A. (2020). The Indic NLP Library: Natural Language Processing for Indic Languages. Retrieved August 2025, from https://github.com/anoopkunchukuttan/indic_nlp_library
82. Han, B., & Baldwin, T. (2011). Lexical Normalisation of Short Text Messages: Makn Sens a #twitter. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Short Papers*, 368–372.
83. Sarker, A. (2017). Social Media Text Normalization: A Survey. *Language Resources and Evaluation*, 51(2), 475–509. https://doi.org/10.1007/s10579-016-9365-9
84. Aguilar, G., AlGhamdi, F., Soto, V., & Solorio, T. (2020). LinCE: A Benchmark for Linguistic Code-switching Evaluation. *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*, 1803–1813.
85. Hermjakob, U. (2018). Uroman: A Universal Romanizer for Low-Resource Languages. USC Information Sciences Institute. Retrieved August 2025, from https://github.com/isi-nlp/uroman
86. Costa-jussà, M. R., Cross, J., Barrault, L., et al. (2022). No Language Left Behind: Scaling Human-Centered Machine Translation to 200 Languages. *Nature*, 610, 491–496. https://doi.org/10.1038/s41586-022-04985-8
87. Koppel, M., & Ordan, N. (2011). Translationese and Its Dialects. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, 1318–1326.
88. Toral, A., Castilho, S., Hu, K., & Way, A. (2018). Reassessing Claims of Human Parity and Super-Human Quality in Machine Translation at WMT 2018. *Proceedings of the 2018 Conference on EMNLP*, 710–720.
89. Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 311–318.
90. Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A Neural Framework for MT Evaluation. *Proceedings of the 2020 Conference on EMNLP*, 2685–2702. https://doi.org/10.18653/v1/2020.emnlp-main.213
91. Lommel, A., Burchardt, A., & Uszkoreit, H. (2014). Multidimensional Quality Metrics (MQM): A Framework for Declaring and Assessing Translation Quality Metrics. *Proceedings of Translating and the Computer 36*.
92. Sennrich, R., Haddow, B., & Birch, A. (2016). Improving Neural Machine Translation Models with Monolingual Data. *Proceedings of ACL 2016*, 86–96.
93. Utiyama, M., & Isahara, H. (2007). A Comparison of Pivot Methods for Phrase-based Statistical Machine Translation. *Proceedings of NAACL-HLT 2007*, 484–491.
94. Wu, H., & Wang, H. (2007). Pivot Language Approach for Phrase-Based Statistical Machine Translation. *Machine Translation*, 21(3), 165–181.
95. More, A. (2014). Triangulation Methods for Low-Resource Machine Translation. *Proceedings of the*

*14th Conference of the European Chapter of the ACL (EACL)*, 90–98.

96. Pires, T., Schlinger, E., & Garrette, D. (2019). How Multilingual is Multilingual BERT? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, 4996–5001. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1493
97. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., … Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747
98. Costa-jussà, M. R., Cross, J., Barrault, L., Elbayad, M., Heafield, K., Heffernan, K., … Zhang, B. (2022). No Language Left Behind: Scaling Human-Centered Machine Translation to 200 Languages. *Nature*, 610, 491–496. https://doi.org/10.1038/s41586-022-04985-8
99. Feng, F., Yang, Y., Cer, D., Arivazhagan, N., Wang, W., & Chen, Z. (2020). Language-agnostic BERT Sentence Embedding. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, 638–647. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.495
100. Card, D., Boydstun, A. E., Gross, J. H., Resnik, P., & Smith, N. A. (2015). The Media Frames Corpus: Annotations of Frames Across Issues. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2015): Short Papers*, 438–444. Association for Computational Linguistics. https://doi.org/10.3115/v1/P15-2072
101. Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., & Johnson, M. (2020). XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalisation. *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, 4411–4421.
102. Liang, Y., Cao, Y., Wei, F., & Huang, M. (2020). XGLUE: A New Benchmark Dataset for Cross-lingual Pre-training, Understanding, and Evaluation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, 6008–6018. https://doi.org/10.18653/v1/2020.emnlp-main.482
103. Hedderich, M. A., Lange, L., Adelani, D., Zhu, D., Alabi, J., & Klakow, D. (2021). A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2545–2563. https://doi.org/10.18653/v1/2021.naacl-main.201
104. Ruder, S., Vulić, I., & Søgaard, A. (2019). A Survey of Cross-lingual Word Embedding Models. *Journal of Artificial Intelligence Research*, 65, 569–631. https://doi.org/10.1613/jair.1.11640
105. Grootendorst, M. (2022). BERTopic: Neural Topic Modeling with Contextual Embeddings. *arXiv preprint arXiv:2203.05794*.
106. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022.
107. Adams, R. P., & MacKay, D. J. C. (2007). Bayesian Online Changepoint Detection. *arXiv preprint arXiv:0710.3742*.
108. Page, E. S. (1954). Continuous Inspection Schemes. *Biometrika*, 41(1/2), 100–115. https://doi.org/10.2307/2333009
109. Zhang, M.-L., & Zhou, Z.-H. (2014). A Review on Multi-label Learning Algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8), 1819–1837. https://doi.org/10.1109/TKDE.2013.39
110. Matthes, J. (2009). What's in a Frame? A Content Analysis of Media Framing Studies in the World's Leading Communication Journals, 1990–2005. *Journalism & Mass Communication Quarterly*, 86(2), 349–367. https://doi.org/10.1177/107769900908600206
111. Baumgartner, F. R., & Jones, B. D. (1993). *Agendas and Instability in American Politics.* University of Chicago Press.
112. Reese, S. D. (2007). The Framing Project: A Bridging Model for Media Research Revisited. *Journal of Communication*, 57(1), 148–154. https://doi.org/10.1111/j.1460-2466.2006.00334.x
113. Rodríguez, C., & Dimitrova, D. V. (2011). The Levels of Visual Framing. *Journal of Visual Literacy*, 30(1), 48–65. https://doi.org/10.1080/23796529.2011.11674684
114. Chong, D., & Druckman, J. N. (2007). Framing Theory. *Annual Review of Political Science*, 10, 103–126. https://doi.org/10.1146/annurev.polisci.10.072805.103054