

When Cross-Lingual Consistency Is Not Enough: Diagnostics for Low-Resource Political Actor Portrayal

Abdul Razaq Vahidi and Sebastian Padó

Institute for Natural Language Processing

University of Stuttgart, Germany

{abdul-razaq.vahidi,sebastian.pado}@ims.uni-stuttgart.de

Abstract

The analysis of multilingual corpora with respect to political stance is of great interest to political science. However, it is a priori unclear whether any differences found are genuinely cross-lingual in nature or artifacts of imperfect alignment, in particular for low-resource languages. In this study, we consider political actor portrayal in Dari and Pashto, the two main languages of Afghanistan. We work with a corpus of 200 quasi-parallel Dari–Pashto article pairs from a single outlet and evaluate the ability of three multilingual sentence-embedding models to make consistent cross-lingual predictions.

We find that actor-level consistency differs substantially across models, and identical predicted labels can conceal differences in the underlying label-score profiles. We conclude that agreement alone is therefore insufficient for interpreting cross-lingual actor-portrayal predictions. This underlines the need for human validation before automatic cross-lingual predictions are used to support substantive conclusions about actor portrayal.

1 Introduction

Political news does more than report events. Through the way political actors are *framed*, news coverage suggests which roles particular participants play (e.g., authority, antagonist, or victim) and shapes how readers understand responsibility, agency, and conflict (Mahmoud et al., 2025; Otmakhova et al., 2024). More generally, framing emphasizes particular aspects of political issues and communicates particular perspectives. Analysing such stances and perspectives is of major interest in political science (Baly et al., 2020).

Particularly interesting is the cross-lingual comparison of framing in countries with multiple languages such as Canada or Switzerland: is there a ‘unified’ public majority opinion, or do opinions segment by language (Mueller et al., 2024)?

The exploration of this setup with NLP methods leads however directly to questions of *robustness*: If a method finds a difference between languages, does it arise genuinely from differences in the discourses, or is it an artifact of the NLP methods used? This problem is particularly urgent for low-resource languages (Philippy et al., 2026).

Our contribution to this debate takes advantage of a new resource, namely a bilingual near-parallel corpus from Afghanistan for the low-resource language pair Dari–Pashto. The corpus comes from a single news outlet, which reduces source-level heterogeneity but does not guarantee identical stance across authors, time periods, or language editions. We therefore treat cross-lingual differences as diagnostic observations that may arise from model behaviour, translation or editorial adaptation, or genuine textual differences, rather than attributing them to the NLP methods alone. Focusing on the task of political actor portrayal, we ask the following research question:

RQ: Are multilingual embedding models sufficiently robust to support cross-lingual analysis of actor portrayal in aligned Dari–Pashto news?

We investigate this question with a number of diagnostics that characterize distinct aspects of model behavior: hard-label agreement, soft label-score profile alignment, and within- and cross-language margins.

Because no human gold portrayal labels are available, our study uses agreement among model predictions as a proxy for correctness. Substantive claims about cross-lingual actor portrayal require human evaluation and are a matter of future work.

2 Background and Related Work

Social Context: Afghanistan. Afghanistan is a multilingual political setting where language has

long shaped administration, state-building, and national identity. Dari and Pashto are the two official languages: Persian historically served as an administrative language, while twentieth-century governments increasingly promoted Pashto as a marker of national identity (Leake, 2023). Although the media sector expanded rapidly after 2001, it remained vulnerable to political, bureaucratic, donor, and armed-actor influence (Relly and Zanger, 2017). Since 2021, mounting content restrictions and economic pressures have further reshaped Afghanistan’s media sector (Afghanistan Journalists Center, 2024). In this context, portraying an actor as an authority, threat, victim, or another narrative role can shape perceptions of responsibility and political legitimacy.

Multilingual Embeddings. State-of-the-art multilingual embedding models such as LaBSE, multilingual-E5, and XLM-R align related multilingual texts in a shared vector space through different training objectives (Feng et al., 2022; Wang et al., 2024; Conneau et al., 2020; Reimers and Gurevych, 2020), typically related to maximizing semantic similarity.

The present study does not assume that these models are equally competent in Dari and Pashto. The available model documentation describes multilingual training but does not provide directly comparable evidence about the amount or quality of training material for both languages. We see this as a setting that is typical for multilingual practice even if imperfectly controlled.

3 Data and Task

Data. We use an aligned Dari–Pashto news corpus from the newspaper and associated publishing house *Etिलाatroz*¹, containing 4,453 article pairs linked through the outlet’s metadata and published between 2020 and 2026. The current experiment uses a development sample of 200 pairs published between 3 November 2020 and 29 March 2021, during the last phase of the democratic Islamic Republic of Afghanistan. After HTML markup removal and text normalization, articles in the sample contain on average 323 whitespace-delimited words in Dari and 302 in Pashto, including titles.

The paired articles were published in close temporal proximity, with a median absolute publication-time gap of 3.68 hours. Many pairs

are paraphrastic rather than word-for-word translations. The pairs typically concern closely related events, but article-level alignment does not establish that the portrayal of every shared actor is identical across the two language versions. Accordingly, observed prediction differences may reflect model behaviour, translation or editorial adaptation, or genuine textual differences.

The corpus is aligned at the article level; no paragraph- or sentence-level alignment is used.

Task: Political Actor Portrayal. Computational framing research increasingly focuses on actors and narratives rather than generic issue frames (Card et al., 2015; Mendelsohn et al., 2021). Because frame inventories do not transfer unchanged across cultural and media contexts, we use a study-specific seven-label actor-centred scheme adapted to Afghan Dari–Pashto political news. The inventory is conceptually informed by prior work on actor-centred and narrative framing (Mahmoud et al., 2025), but its final categories were adapted to the Afghan news context. The seven labels are *authority*, *antagonist*, *victim*, *threat*, *guardian*, *passive/affected*, and *neutral/unclear*. Each label comes with a natural language definition; see Appendix A for details.

4 Method

4.1 Actor Analysis

Our pipeline for political actor portrayal labels in low-resource Dari/Pashto contexts has two main steps: actor identification and actor labeling. See Appendix B for implementation details.

Actor Identification. The unit of analysis is a political or social actor (person, group, institution, state, or social category). We identify actors using a manually constructed bilingual gazetteer comprising ten canonical actor groups. Each group links manually compiled Dari and Pashto surface forms to a shared canonical English actor name, for example Taliban/Islamic Emirate, Afghan government, civilians, and the UN. After normalization, matching uses case-sensitive substring matching; we do not apply token-boundary or morphological matching, entity linking, or coreference resolution. We retain one record for each detected canonical actor in an article, irrespective of the number of mentions. This design provides a transparent and reproducible baseline, but it has limited coverage and may miss entities or surface forms;

¹<https://www.etilaatroz.com>

the complete gazetteer and matching details are provided in Appendix B.2.

Actor Labeling. We assign labels by embedding-to-definition matching with three cross-lingual sentence-embedding models: LaBSE (Feng et al., 2022), Multilingual-E5 (Wang et al., 2024) and XLM-R-Sentence (Reimers and Gurevych, 2020). We use these encoders in a zero-shot setup: the context around an actor and each label definition are embedded, similarities are computed, and the highest-scoring label is chosen. For all three models, the model-facing label names and label definitions are given in English (Appendix A). This provides a uniform set of target categories across the two language sides, but it does not establish that the definitions carry equivalent pragmatics in Dari and Pashto.

For each actor–article pair a , we use the context window around the earliest matched surface form, c_a^L , in language $L \in \{D, P\}$, and compare it with the written definition d_r of every portrayal label r . Later mentions of the same actor do not enter the current prediction, making this a simple operationalisation rather than an article-level assessment. Let $f(\cdot)$ be the model embedding function. We define the label score and prediction as

$$s_{a,r}^L = \cos(f(c_a^L), f(d_r)), \quad \hat{r}_a^L = \arg \max_r s_{a,r}^L.$$

4.2 Analysis of Predictions

We now analyze the predictions of our models. We propose two families of scores that capture complementary aspects of cross-lingual behavior. The first, consistency score, assesses the amount of cross-lingual *agreement* in political actor portrayal and has two variants (see below). However, consistency does not indicate how decisively a model prefers the selected label over its closest alternative. Our second family consists of *margin-based scores* that quantify the separation between the preferred label and competing labels, both within each language and when the preferred label is evaluated cross-lingually.

Consistency Score (CS). For an aligned article pair i , let A_i^D and A_i^P denote the actors identified in its Dari and Pashto versions. We evaluate only the shared actors $A_i = A_i^D \cap A_i^P$ – the actors found in both versions of the same article pair i . We first define the *hard CS* as:

$$\text{CS}_{\text{hard}}(i) = \frac{1}{|A_i|} \sum_{a \in A_i} \mathbb{I}[\hat{r}_a^D = \hat{r}_a^P].$$

This is cross-lingual predicted-label agreement. It is not an accuracy measure because no human gold portrayal labels are available.

Since the hard score looks only at the top-ranked label, we also compute a soft CS. Let $s_a^L = (s_{a,1}^L, \dots, s_{a,K}^L)$ denote the complete label-score vector and $\tilde{s}_a^L$ its min–max-normalized version. Then we define the *soft CS* as:

$$\text{CS}_{\text{soft}}(i) = \frac{1}{|A_i|} \sum_{a \in A_i} \cos(\tilde{s}_a^D, \tilde{s}_a^P).$$

This is the average cross-lingual cosine similarity between the normalized label-score profiles of all shared actors in pair i . Hard CS tests whether the single argmax label is identical across the two languages. Soft CS compares all seven normalized label scores and therefore retains information about the complete pattern of label preferences.

Top-label margins. For each actor–article instance a and language $L \in \{D, P\}$, let $\hat{r}_a^L = \arg \max_r s_{a,r}^L$ be the highest-scoring label, and let L' denote the other language. We can now compare margins within and across languages. The *within-language top-two margin* is

$$m_a^L = s_{a,\hat{r}_a^L}^L - \max_{r \neq \hat{r}_a^L} s_{a,r}^L.$$

This non-negative quantity measures the local separation between the preferred label and its closest alternative. We now measure how the score of the language-specific top label changes in the other language (*cross-lingual directional margin*):

$$x_a^D = s_{a,\hat{r}_a^D}^D - s_{a,\hat{r}_a^D}^P, \quad x_a^P = s_{a,\hat{r}_a^P}^P - s_{a,\hat{r}_a^P}^D.$$

We report the absolute values of these scores, $|x_a^D|$ and $|x_a^P|$. These values are reported as exploratory model-score diagnostics. Raw cosine scores can differ in language-level offsets, so the relation between $|x|$ and m is not a calibrated or directly commensurable measure of cross-lingual stability. It therefore does not establish stable label preferences or prediction correctness.

5 Results

The current analysis is restricted to a fixed 200-pair development sample, not a random or representative sample of the full corpus. At least one shared gazetteer actor is found in 88 pairs, for a total of 189 shared actor–article-pair instances. The remaining 112 pairs yielded no shared canonical

English	Magnetic bombs and the targeting of the “soft part” of society—civil activists, journalists , and experts—are described as part of the Taliban’s eleven-point plan to weaken the public’s resolve to defend its values.
Dari	مېمې های مقناطیسی و هدف قراردادن بخش نرم جامعه یعنی فعالین مدنی - خبرنگاران و اهل خبره جز برنامه ی یازده پایه ای طالب است ...
Pashto	مقناطیسي بمونه او د ټولنې نرمه برخه یانې مدني فعالان - خبريالان او اهل خبره موخه گرځول د طالبانو يوولس پایه يي برنامې برخه ده ...

Table 1: Sentence pair #107. Sources: <https://www.etilaatroz.com/114857/amrullah-saleh-says-pat-h-to-peace-does-not-pass-through-doha-alone/> and <https://www.etilaatroz.com/ps/1351>.

Model	Hard CS ($\uparrow$)	Soft CS ($\uparrow$)
LaBSE	0.691	0.932
Mult.-E5	0.518	0.896
XLM-R	0.424	0.843

Table 2: Cross-lingual Political Consistency Scores macro-averaged over the 88 non-empty article pairs with 189 shared actor–article-pairs in Dari/Pashto. Hard CS counts matches between winning labels; soft CS compares normalized label-score profiles.

actor under the gazetteer. This absence reflects both differences between the paired articles and limitations of the gazetteer.

Our main observations are as follows:

Hard-label agreement conceals score variation.

We first consider hard consistency, shown in Table 2. Within this fixed development sample, LaBSE has the highest observed mean pair-level hard CS (0.691), followed by multilingual-E5 (0.518) and XLM-R-Sentence (0.424). The same observed ordering holds for soft CS, with scores of 0.932, 0.896, and 0.843. These are descriptive results from a small fixed sample; the present study does not estimate confidence intervals or conduct significance tests for model differences.

As for soft CS, we focus on LaBSE with the highest hard CS: Only 19 of the 189 instances (10.1%) have a cross-lingual similarity above 0.99, compared to 135 of 189 (71.4%) instances receiving identical hard labels. Thus, LaBSE’s high hard CS conceals differences between the underlying label-score profiles. See Appendix D for per-label results.

Within- and across-language margins. Table 3 reports the local separation of the top label from its nearest alternative (m) alongside the absolute cross-lingual change in the score of that language-

Model	within languages		across languages	
	m_a^D ($\uparrow$)	m_a^P ($\uparrow$)	$ x_a^D $ ($\downarrow$)	$ x_a^P $ ($\downarrow$)
LaBSE	0.033	0.039	0.034	0.040
Mult.-E5	0.005	0.005	0.010	0.015
XLM-R	0.033	0.024	0.110	0.162

Table 3: Mean within-language top-two margins m and mean cross-lingual margins $|x|$ in the scores of language specific top labels over 189 shared actor–article-pair instances for Dari (D) and Pashto (P).

specific top label ($|x|$). These quantities are descriptive diagnostics of the models’ raw score profiles; they are not directly commensurable measures of cross-lingual stability.

For LaBSE, the mean within-language margins are 0.033 in Dari and 0.039 in Pashto, while the corresponding mean absolute cross-lingual changes are 0.034 and 0.040. For multilingual-E5, the mean absolute cross-lingual changes (0.010 and 0.015, respectively) exceed the within-language margins (0.005 in both languages). For XLM-R-Sentence, the corresponding changes are 0.110 and 0.162, compared with margins of 0.033 and 0.024. These descriptive patterns show that agreement on hard labels can coexist with different underlying score profiles, but they do not by themselves establish cross-lingual stability, model superiority, or human-validated correctness.

Illustrative example. As a concrete illustration, consider the following main statement from article pair #107. as shown in Table 1. The sentences in Dari and Pashto are rather parallel and their shared meaning is conveyed by the English version. However, the models produce sharply different predictions for the entity *journalists* shown in Table 4. Arguably, the correct characterization is either *Passive/Affected* or *Victim*. Multilingual-E5 selects

	LaBSE	Mult.-E5	XLM-R
Dari top label	Passive/Affected (.148)	Victim (.751)	Neutral/Unclear (.274)
Dari runner-up	Threat (.143)	Neutral/Unclear (.749)	Passive/Affected (.170)
Pashto top label	Threat (.189)	Threat (.763)	Authority (.398)
Pashto runner-up	Passive/Affected (.109)	Passive/Affected (.761)	Guardian (.388)

Table 4: Top two labels and scores for the entity *journalists/media* in article pair 107.

Victim in Dari but *Threat* in Pashto, and both decisions have very small top-two margins. XLM-R-Sentence labels the journalists as *Neutral/Unclear* in Dari and *Authority* in Pashto, while LaBSE changes from *Passive/Affected* in Dari to *Threat* in Pashto. See Appendix C for a second example.

Cross-lingual distributional asymmetries.

Across the detected actor pairs, the predicted-label distributions differ substantially between the two language sides. XLM-R-Sentence assigns *Guardian* to 44.4% of Pashto actor–article pairs, compared with 13.8% of Dari pairs, a difference of more than 30 percentage points. Multilingual-E5 assigns fewer *Neutral* labels and more *Authority* and *Threat* labels on the Pashto side. LaBSE is dominated by *Threat*, which accounts for over 65% of all its predictions. See Table 11 in Appendix D for details. This label concentration may contribute to high hard-label agreement even where the model has limited sensitivity to differences among other portrayal labels. The language-side distributions are descriptive and may also reflect different actor composition, translation or editorial adaptation, or genuine textual differences between paired articles. They are therefore not direct evidence of language-specific model bias or relative model competence in Dari versus Pashto.

6 Conclusion

In this paper, we asked whether multilingual embedding models provide sufficiently robust predictions for cross-lingual analysis of political actor portrayal in a bilingual low-resource setting (Dari–Pashto). The quasi-parallel corpus enables a controlled comparison of model predictions across two language versions of related news coverage.

We proposed two families of diagnostics: measures of consistency (cross-lingual top-label and distributional agreement) and margin-based measures (within and across languages).

We find that hard-label consistency can conceal substantial variation in the complete label-score profiles. Thus, matching predicted labels alone do not establish that a multilingual model represents actor portrayal in the same way across Dari and Pashto. The present diagnostics identify cases requiring closer examination, but they do not determine whether a prediction is correct or whether paired articles genuinely differ in portrayal.

Data Availability Reproducibility materials are available at <https://doi.org/10.5281/zenodo.22945205>. The repository contains aggregate result tables and supporting documentation. Full news texts and contextual passages are not redistributed because they remain subject to copyright. Article identifiers and source URLs will be added before the workshop, subject to the source outlet’s terms and page availability.

Acknowledgments This project has received funding through the SAFE Supporting At-Risk Researchers with Fellowships in Europe project, which is funded by the European Union under the Grant Agreement ref. 101148426. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the European Research Executive Agency (REA) can be held responsible for them.

Limitations

The present pipeline measures cross-lingual prediction consistency rather than correctness; establishing correctness requires independent human-labelled evaluation data. This is a preliminary diagnostic on one corpus, one domain, and three embedding models. The 200 pairs form a fixed development sample, not a random or representative sample of the full 4,453-pair corpus. Actor extraction is gazetteer-based and may miss actors or surface forms; label assignment is embedding-

to-definition matching rather than a trained classifier, and trained or instruction-tuned models may behave differently.

All model-facing label definitions are in English. This design does not permit conclusions about whether the definitions are equally appropriate or equally interpretable in Dari and Pashto; translated-definition controls are required before language-side asymmetries can be interpreted as evidence of relative model competence. The current pipeline uses only the earliest detected actor mention, which is a baseline operationalisation and can miss later or changing portrayals in the same article. Raw cosine-score differences are model-specific, and the margin diagnostics are exploratory rather than calibrated confidence measures. The study does not estimate confidence intervals or significance tests for model differences. Cross-lingual prediction changes may reflect model behaviour, translation or editorial adaptation, or genuine textual differences between paired articles. Future work will obtain independent human annotations of Dari and Pashto contexts, evaluate all-mentions or multi-context aggregation, and conduct broader robustness analyses on the full corpus.

Ethics Statement

The corpus consists of publicly accessible news, but public availability does not eliminate ethical obligations. Articles may name individuals and describe violence, persecution, or politically sensitive affiliations; we avoid claims about which portrayal is “true,” minimize redistribution of unnecessary personal data, and report aggregate results whenever possible. Any released annotation data will include only text necessary for reproducibility, respect the source’s terms and copyright, and be reviewed for re-identification or security risks in the Afghan context. Annotators will receive informed task instructions, secure data access, fair compensation, and the option to withdraw from distressing material.

References

- Afghanistan Journalists Center. 2024. [AFJC expresses concern over escalating restrictions on media freedom in afghanistan ahead of world press freedom day](#). Published 2 May 2024.
- Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. [We can detect](#)

[your bias: Predicting the political ideology of news articles](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982--4991, Online. Association for Computational Linguistics.

- Dallas Card, Amber E. Boydston, Justin H. Gross, Philip Resnik, and Noah A. Smith. 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)*, pages 438--444, Beijing, China. Association for Computational Linguistics.

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

- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. [Language-agnostic BERT sentence embedding](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878--891, Dublin, Ireland. Association for Computational Linguistics.

- Elisabeth Leake. 2023. [Constitutions and modernity in post-colonial afghanistan: Ethnolinguistic nationalism and the making of an afghan nation-state](#). *Law and History Review*, 41(2):295--315.

- Tarek Mahmoud, Zhuohan Xie, Dimitar Iliyanov Dimitrov, Nikolaos Nikolaidis, Purificação Silvano, Roman Yangarber, Shivam Sharma, Elisa Sartori, Nicolas Stefanovitch, Giovanni Da San Martino, Jakub Piskorski, and Preslav Nakov. 2025. [Entity framing and role portrayal in the news](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 302--326, Vienna, Austria. Association for Computational Linguistics.

Julia Mendelsohn, Ceren Budak, and David Jurgens. 2021. [Modeling framing in immigration discourse on social media](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2219--2263, Online. Association for Computational Linguistics.

Sean Mueller, Paolo Dardanelli, Min Reuchamps, Pirmin Bindi, André Lecours, and Christoph Niessen. 2024. [A francophone political culture? Similarities and differences among French speakers in Canada, Belgium, Switzerland, and France](#). *French Politics*, 22:115--135.

Yulia Otmakhova, Shima Khanehzar, and Lea Frermann. 2024. [Media framing: A typology and survey of computational approaches across disciplines](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15407--15428, Bangkok, Thailand. Association for Computational Linguistics.

Fred Philippy, Siwen Guo, Jacques Klein, and Tegawendé F. Bissyandé. 2026. [Why low-resource NLP needs more than cross-lingual transfer: Lessons learned from Luxembourgish](#). In *Proceedings of The Big Picture v2: Crafting a Research Narrative*, pages 82--93, San Diego, CA, USA. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. [Making monolingual sentence embeddings multilingual using knowledge distillation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512--4525, Online. Association for Computational Linguistics.

Jeannine E. Relly and Margaret Zanger. 2017. [The enigma of news media development with multi-pronged “capture”: The afghanistan case](#). *Journalism*, 18(10):1233--1255.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. [Multilingual E5 text embeddings: A technical report](#). *arXiv preprint arXiv:2402.05672*.

A Portrayal Label Inventory and Provenance

The seven portrayal labels constitute a study-specific operational inventory rather than a verbatim reproduction of a single existing taxonomy. The inventory was conceptually informed by actor-centred and narrative framing research, in which political and social actors may be represented as victims, threats, responsible agents, or protective figures (Mendelsohn et al., 2021; Mahmoud et al., 2025). The terminology and definitions were adapted to recurring roles in Afghan political news.

Table 5 reports the exact English definitions supplied to the embedding models. For each label, the model received the label name followed by a colon and its definition. No Dari- or Pashto-translated definitions were used in the present experiment.

Several distinctions are important when interpreting the inventory. *Authority* refers primarily to institutional legitimacy or decision-making power, whereas *Guardian* concerns protective or supportive action. *Antagonist* attributes responsibility for harm or wrongdoing, whereas *Threat* represents an actor as a current or potential danger. *Victim* requires explicit harm, deprivation, or attack, while *Passive/Affected* captures limited agency without necessarily implying direct victimization. *Neutral/Unclear* is appropriate when the context does not provide a sufficiently clear evaluative or narrative role. These distinctions are explanatory notes for readers; only the definitions in Table 5 were supplied to the models.

B Implementation Details

B.1 Sample construction and preprocessing.

The 200-pair development sample consists of the first 200 article pairs after the Dari and Pashto WordPress exports were merged using their cross-language article identifiers. It is therefore a fixed development sample rather than a random sample. Article titles and bodies were concatenated before analysis. HTML entities, WordPress comments, and HTML markup were removed, and consecutive whitespace was collapsed.

The normalization procedure replaced common Arabic-script variants, including Arabic *yeh* and *kaf* with their Persian forms, normalized several forms of *alef*, *heh*, and *waw*, and removed bidirectional and zero-width control characters.

Label	Exact model-facing definition
Authority	The actor is presented as an official, legitimate, governing, or decision-making body.
Antagonist	The actor is portrayed as responsible for harm, repression, violence, corruption, abuse, or wrongdoing.
Victim	The actor is harmed, excluded, killed, injured, silenced, displaced, deprived, or attacked.
Threat	The actor is represented as a danger to security, rights, social order, political stability, or public safety.
Guardian	The actor is portrayed as protecting, supporting, rescuing, defending, helping, or advocating for others.
Passive/Affected	The actor is affected by events or decisions but is not represented as having strong agency.
Neutral/Unclear	The actor is mentioned without a clearly evaluative or narrative role.

Table 5: The exact portrayal-label definitions used in the embedding-to-definition comparison.

B.2 Actor Detection

Actor detection and context selection. Actor detection used case-sensitive substring matching after normalization; no token-boundary or morphological matching was applied. For every canonical actor detected in an article, the implementation created one actor–article record, irrespective of the number of occurrences. Mention counts were retained as metadata but did not enter the CS calculation.

The context consisted of a window of at most 900 characters surrounding the earliest detected surface form of the actor. More precisely, the window began up to 450 characters before the starting position of the earliest match and ended up to 450 characters after it, subject to the article boundaries. Thus, the experiment did not classify every actor occurrence separately.

Before encoding, the canonical English actor name was prepended to the Dari or Pashto context using the following structure:

Actor: [canonical actor name]
Context: [Dari or Pashto context]

The model-facing label texts similarly contained the English label name and its English definition.

Gazetteer and Coverage The bilingual gazetteer contained ten canonical actor groups. Dari and Pashto surface forms were stored together under their shared canonical English actor name. Table 6 reports the number of unique implemented surface forms. The complete Unicode

Canonical actor group	Unique forms
Taliban/Islamic Emirate	6
Afghan government	5
Women and girls	7
Civilians	7
Journalists/media	6
United Nations	6
Pakistan	4
Iran	4
United States	5
ISIS/Daesh	5

Table 6: Canonical actor groups and numbers of unique gazetteer surface forms.

gazetteer and matching code will be included in the reproducibility materials.

Table 7 shows how the 200-pair sample was reduced to the set used for CS. Of the 200 pairs, 92 contained no detected gazetteer actor in either language, 17 contained a detected actor on only one language side, and three contained detected actors in both versions but no shared canonical actor. The resulting 88 usable pairs contained 189 shared actor–article-pair instances.

The absence of a shared gazetteer actor should not be interpreted as the absence of political actors from an article. It may result from missing surface forms, spelling or inflectional variation, actors outside the ten predefined groups, or genuine differences between the paired articles. Since these excluded pairs have not yet been manually audited,

Coverage stage	Number
Total aligned article pairs	200
Actor detected in either version	108
Actor detected in both versions	91
At least one shared canonical actor	88
Shared actor–article-pair instances	189

Table 7: Gazetteer-coverage flow for the development sample.

the relative contribution of these explanations remains unknown.

B.3 Actor Labeling

Model checkpoints. Table 8 lists the exact checkpoint identifiers used in the notebook.

For multilingual-E5, actor contexts were prefixed with query:, while label definitions were prefixed with passage:. No equivalent prefix was applied to LaBSE or the XLM-R-based checkpoint. Embeddings were generated with `normalize_embeddings=True`; consequently, the dot product between context and label embeddings was equivalent to cosine similarity.

B.4 Prediction Analysis

Score normalization. For soft CS, each seven-dimensional score vector is normalized separately within each language:

$$\tilde{s}_{a,r}^L = \frac{s_{a,r}^L - \min_j s_{a,j}^L}{\max_j s_{a,j}^L - \min_j s_{a,j}^L}, \quad r = 1, \dots, K.$$

All score vectors used in the soft CS calculation had non-zero range, so the denominator was non-zero in every case.

No explicit token limit was set in the notebook. Tokenization and truncation therefore followed the default configuration of each Sentence-Transformer checkpoint. The original notebook also did not pin the software-library versions or checkpoint revisions; these should be recorded in the final reproducibility package.

C Contrastive Examples

The following example illustrates again why identical or different hard labels cannot be interpreted without considering the underlying scores. They are diagnostic examples rather than human gold annotations. The English glosses summarize the relevant context for readers.

Pair 84: Women and girls. The Dari and Pashto versions report the same central event:

The Independent Directorate of Local Governance announced that eight women had been appointed as social deputies in eight provinces.

Table 9 shows that multilingual-E5 and XLM-R-Sentence select *Authority* in both languages. LaBSE instead changes from *Authority* in Dari to *Threat* in Pashto. In the Dari version, however, LaBSE’s scores for *Authority* and *Threat* differ by only approximately 0.003, showing that even the Dari decision is based on a near tie.

D Additional Results

Table 10 reports hard-label consistency for each canonical actor. The number of cases is the number of aligned article pairs in which that actor was detected in both language versions. Several actor groups have very small sample sizes; their differences should therefore be treated as descriptive rather than as stable actor-specific effects.

Table 11 reports composition-controlled predicted-label distributions over the shared-actor set.

Reported model	Checkpoint identifier
LaBSE	sentence-transformers/LaBSE
Multilingual-E5	intfloat/multilingual-e5-base
XLM-R-Sentence	sentence-transformers/paraphrase-xlm-r-multilingual-v1

Table 8: Sentence-embedding checkpoints used in the experiment.

Model	Dari top label	Dari runner-up	Pashto top label	Pashto runner-up
LaBSE	Authority (.217)	Threat (.214)	Threat (.261)	Authority (.236)
Multilingual-E5	Authority (.753)	Guardian (.744)	Authority (.758)	Threat (.753)
XLM-R-Sentence	Authority (.254)	Guardian (.249)	Authority (.430)	Guardian (.393)

Table 9: Top-two cosine-similarity scores for *Women and girls* in article pair 84. Scores are similarities, not calibrated probabilities.

Actor	n	LaBSE	Multilingual-E5	XLM-R-Sentence
Taliban/Islamic Emirate	55	78.2%	50.9%	47.3%
Afghan government	17	70.6%	29.4%	35.3%
Women and girls	5	60.0%	60.0%	40.0%
Civilians	30	63.3%	33.3%	20.0%
Journalists/media	7	42.9%	71.4%	57.1%
United Nations	10	60.0%	60.0%	30.0%
Pakistan	21	76.2%	61.9%	47.6%
Iran	8	75.0%	62.5%	12.5%
United States	29	79.3%	31.0%	51.7%
ISIS/Daesh	7	57.1%	71.4%	28.6%

Table 10: Hard-label CS (cross-lingual consistency) by actor. Percentages are descriptive because several actor groups have small n .

Label	LaBSE	LaBSE	Mult.-E5	Mult.-E5	XLM-R	XLM-R
	Dari	Pashto	Dari	Pashto	Dari	Pashto
Authority	1.6%	0.5%	13.2%	24.9%	34.4%	28.0%
Antagonist	2.6%	6.3%	1.1%	1.6%	11.6%	6.3%
Victim	5.3%	9.5%	23.3%	23.8%	20.6%	9.5%
Threat	65.1%	69.8%	18.5%	30.2%	7.9%	4.8%
Guardian	11.6%	6.9%	7.4%	4.2%	13.8%	44.4%
Passive/Affected	11.1%	5.8%	9.5%	9.5%	1.1%	0.0%
Neutral/Unclear	2.6%	1.1%	27.0%	5.8%	10.6%	6.9%

Table 11: Predicted-label distributions over the 189 shared actor–article-pair instances on each language side. Each model–language column has the same denominator ($n = 189$), thereby holding actor composition constant across Dari and Pashto.