Multilingual Target-Stance Extraction

Ethan Mines, Bonnie Dorr
University of Florida, Gainesville, USA
{ethanlmines, bonniejdorr}@ufl.edu

Abstract

Social media enables data-driven analysis of public opinion on contested issues. Target-Stance Extraction (TSE) is the task of identifying the target discussed in a document and the document's stance towards that target. Many works classify stance towards a *given* target in a multilingual setting, but all prior work in TSE is English-only. This work introduces the first multilingual TSE benchmark, spanning Catalan, Estonian, French, Italian, Mandarin, and Spanish corpora. It manages to extend the original TSE pipeline to a multilingual setting without requiring separate models for each language. Our model pipeline achieves a modest F1 score of 12.78, underscoring the increased difficulty of the multilingual task relative to English-only setups and highlighting target prediction as the primary bottleneck. We are also the first to demonstrate the sensitivity of TSE's F1 score to different target verbalizations. Together these serve as a much-needed baseline for resources, algorithms, and evaluation criteria in multilingual TSE.

Keywords: Stance Detection, Multilingual Adaptation, Opinion Mining

1. Introduction

Stance detection is the task of predicting an author's opinion or stance towards a particular target. Potential targets can include political candidates (Li et al., 2021), COVID-19 policies (Glandt et al., 2021), and corporate mergers (Conforti et al., 2020), among others studied in recent stance detection work.

Most contemporary stance detection systems rely on machine learning models. One can either train a separate model for each possible target (Zarrella and Marsh, 2016) or a general model that takes a representation of the target as part of its input (Augenstein et al., 2016; Xu et al., 2018). The latter variant allows one to perform zero-shot stance detection: detecting the stance towards targets not seen during training (Allaway and McKeown, 2020).

In either case, one must already know the target of the author's writing prior to detecting the stance. This generally requires either keyword-based retrieval of relevant social media posts (Conforti et al., 2020) or explicit human annotation of the targets. Li et al. (2023b) introduce Target-Stance Extraction (TSE), the task of first predicting the target of the document and then the document's stance toward that target. As illustrated in Figure 1, TSE is more useful to stakeholders because it eliminates the need to predefine relevant targets.

The primary contributions of this study are (1) generalizing the original TSE algorithm to a multilingual setting and (2) introducing a benchmark dataset to facilitate further progress in multilingual TSE. While stance detection has been performed for languages other than English, all existing TSE works focus exclusively on English. A multilingual TSE algorithm would allow decision makers to ascertain public opinion across cultures. We adapt Target-Stance Extraction to a multilingual scenario, including Romance languages, Estonian, and Mandarin. The accompanying multilingual TSE benchmark dataset combines existing stance detection corpora for these languages. Our generalized algorithm[1] predicts both the target and stance for a sample, without relying on the language itself to determine the prediction.

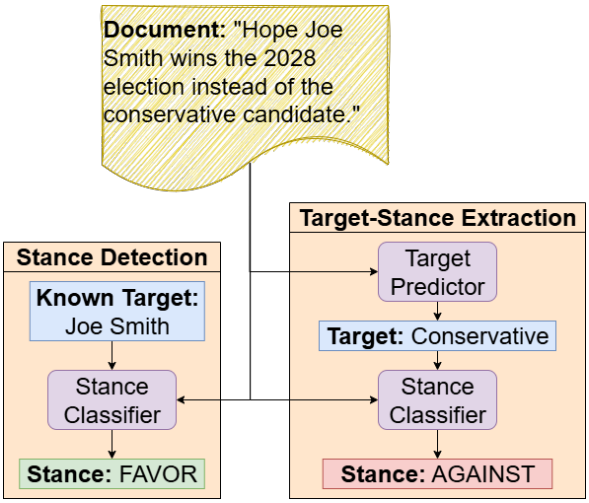

Figure 1: Stance Detection versus Target-Stance Extraction

2. Related Work

This section reviews stance detection in general (§2.1), multilingual stance detection (§2.2), and the more general TSE task (§2.3).

[1]Algorithm and data-download scripts available at https://github.com/elmines/multiling_tse

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8997–9007
11–16 May 2026.

2.1. Stance Detection

The first widely adopted stance detection benchmark is the SemEval2016 Task 6 stance dataset, consisting of six targets relevant to U.S. politics (Mohammad et al., 2016). Early studies train separate classifiers for each target (Zarrella and Marsh, 2016; Wei et al., 2016). Later work introduces cross-target stance detection, where a model trained on one target classifies stance toward a related target (Xu et al., 2018; Liang et al., 2021). The next step beyond this is zero-shot stance detection (ZSSD), which evaluates a model on targets unrelated to those seen in training. The VAST dataset (Allaway and McKeown, 2020) is the standard benchmark for ZSSD, though the newer EZ-Stance dataset is a viable alternative (Zhao and Caragea, 2024). More recent has been the development of multimodal stance corpora (Niu et al., 2024; Liang et al., 2024).

Since the introduction of BERT (Devlin et al., 2019), it is common practice to fine-tune a pretrained BERT model for stance classification (Liang et al., 2022; Luo et al., 2022; Zhao and Caragea, 2025). The model typically takes a concatenation of the target and the document:

```
[CLS] target [SEP] document [SEP]
```

The final hidden state of the `[CLS]` token is passed to a classifier head that assigns the appropriate stance label.

Beyond choice of model, many works introduce adversarial learning objectives to improve the hidden representations for targets and/or documents (Wei and Mao, 2019; Allaway et al., 2021; Liang et al., 2022). Others retrieve supplementary background knowledge—typically Wikipedia descriptions of the target—to enrich contextual understanding (Zhu et al., 2022; He et al., 2022; Li et al., 2023a). Still others use knowledge graphs like ConceptNet to augment their models with symbolic knowledge (Liu et al., 2021; Luo et al., 2022; Chen et al., 2024).

As in these related studies, the stance classification stage of our pipeline uses a BERT model that predicts stance given the concatenation of the target and document. In contrast, we also train a separate model for predicting the target prior to performing stance classification.

2.2. Multilingual Stance Detection

Non-English stance detection datasets vary widely in scope, both regarding language and their targets. Many resources are limited to a single issue in a single language, such as immigration to Estonia (Mets et al., 2024). Broader collections contain a small set of targets relevant to a specific nation, such as French electoral candidates (Lai et al., 2020), but are still limited to one language. Truly multilingual stance corpora—such as X-Stance (Vamvas and Sennrich, 2020) and Stanceosaurus (Zheng et al., 2022)—span multiple languages and target domains, enabling cross-lingual transfer.

Regarding algorithms, the creators of X-Stance show that a pretrained multilingual BERT model (Devlin et al., 2019) can be fine-tuned to classify multilingual stance samples (Vamvas and Sennrich, 2020). Lai et al. (2020) use a feature-based approach, studying the use of different machine learning features for stance detection in the political domain across five languages. Hardalov et al. (2022) apply sentiment analysis to multilingual Wikipedia articles to create silver training data for their stance classifier. They create a multilingual stance benchmark by combining many existing stance datasets and demonstrate the feasibility of few-shot training a stance classifier for a new language. Zhao and Caragea (2025) similarly create a bilingual benchmark by combining the Mandarin C-Stance (Zhao et al., 2023) and English EZ-Stance (Zhao and Caragea, 2024) datasets and evaluating cross-lingual transfer.

Building on these multilingual corpora, we construct a new TSE benchmark that extends prior stance detection efforts. As in recent studies, our approach relies on pretrained language models such as BERT, but our focus is the more general task of Target-Stance Extraction, not just stance detection alone.

2.3. Target-Stance Extraction

Li et al. (2023b) introduce the task of Target-Stance Extraction (TSE), predicting both the target of a document and the document's stance toward it. They adopt two different approaches to predicting targets: Target Classification (TC) and Target Generation (TG). TC fine-tunes a pretrained BERT model with a classifier head to predict among a fixed pool of targets. TG uses a BART sequence transduction model (Lewis et al., 2020) to generate a free-form target. For evaluation purposes, the free-form targets are still mapped to a fixed pool of targets using cosine embedding similarity. Regardless of how targets are predicted, an additional BERT model is fine-tuned to predict stance given a target and a document. They form their benchmark dataset by collating four existing English stance datasets.

Yan et al. (2025) simplify the TC approach by using the same underlying BERT encoder for both target prediction and stance prediction. They train two different classifier heads for the model: one for predicting the target when given only a document and another for predicting the stance given both a document and a target.

Akash et al. (2025) introduce Open-Target Stance Detection (OTSD), a scenario similar to

Target-Stance Extraction with Target Generation. They validate the effectiveness of using large-language models for target prediction and stance classification without any pretraining. OTSD differs from TSE: OTSD emphasizes open-target discovery without a fixed mapping step, relying on human evaluation to verify target quality. TSE on the other hand requires mapping the predicted target to a pool of predetermined targets during evaluation. Our study follows the TSE setting.

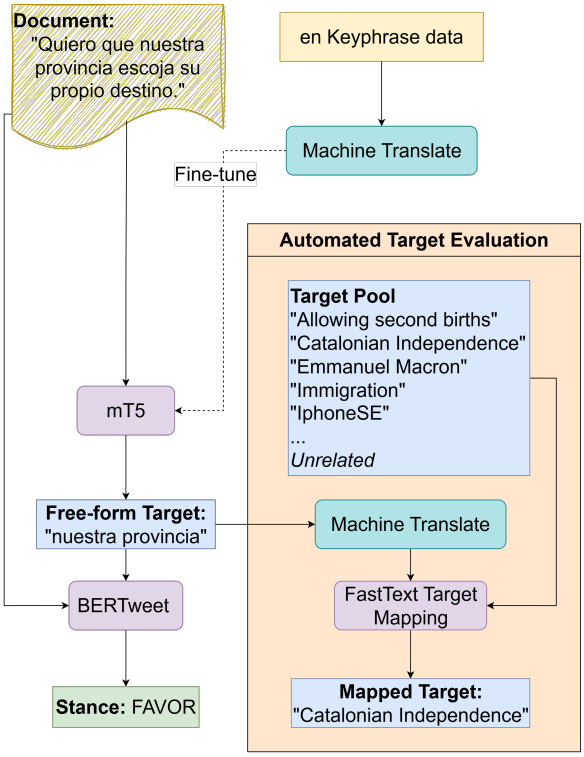

Figure 2: Overview of Multilingual Target-Stance Extraction. A machine-translated keyword corpus is used to fine-tune an mT5 sequence model for target prediction. A BERTweet model predicts the stance given the original document and the target prediction. Using FastText embedding similarity, the free-form predicted target is mapped to a set of fixed targets for evaluation purposes.

Mather et al. (2021) introduce a novel approach to stance detection that requires no machine learning models but some semi-automated resource construction. They provide a more explainable grounding for stance as a belief regarding the target plus a sentiment towards that belief. While this work does not use such a notion of stance, both works are able to predict targets in addition to stance.

Like Li et al. (2023b), we collate existing stance corpora to form a benchmark TSE dataset. Distinct from prior multilingual stance work, our study addresses TSE, which requires predicting both targets and stances. Extending earlier TSE studies, the pipeline operates in a multilingual setting. Additionally, the Target Generation mode is adopted exclusively for predicting targets. The non-English stance corpora used here each have very distinct target sets. A simple classification head (i.e., the task-specific output layer) can easily memorize the mapping between each corpus and its targets—even if all language cues are removed by translating all texts into the same language.

Building on prior work in multilingual stance detection and English-only TSE, the following section describes our design for a multilingual TSE benchmark and pipeline.

3. Methodology

Described here are the problem definition for Target-Stance Extraction (§3.1), the multilingual TSE algorithm (§3.2), and the data used for training and evaluation (§3.3).

3.1. Problem Definition

Consider an input document x, a groundtruth (GT) target t, and a groundtruth stance $y \in \{\text{Against}, \text{Favor}, \text{Neutral}\}$. The goal of stance detection is to predict a stance $\hat{y}$ given (x, t). The goal of target-stance extraction is to predict both $\hat{y}$ and a target $\hat{t}$ given only x. We impose the additional constraint that the language of $\hat{t}$ matches that of x, to ensure the TSE system remains truly multilingual.

3.2. TSE Pipeline

We operationalize multilingual TSE with a two-stage pipeline (Figure 2). An mT5 sequence transduction model (Xue et al., 2021) first generates a free-form target for the input document, and a BERTweet (Nguyen et al., 2020) stance classifier then predicts stance given the document and the generated target. The mT5 model is fine-tuned on a machine-translated keyphrase-generation corpus, and the stance model is fine-tuned on labeled stance examples. Due to the keyphrase generation data used to fine-tune the mT5 model, a single output sequence can have multiple keyphrases. During postprocessing, exact duplicates are removed to ensure unique candidate targets.

To perform an automated evaluation of generated targets, there needs to be a way to map them to a set of known targets. Following Li et al. (2023b), all possible targets from the stance corpora are collated into a target pool. Because most targets in the pool do not co-occur across languages, we apply the mapping process using only English verbalizations to prevent the system from exploiting language-specific cues.

Target	Verbalization
catalonia	Catalonian Independence
immigration	Immigration
lepen	Marine LePen
macron	Emmanuel Macron
sardinia	Sardinian Independence
firecracker	Setting off firecrackers during the Spring Festival [a]
iphone	IphoneSE [b]
russia	Russia's counter-terrorism operations in Syria [c]
secondbirth	Allowing second births
shenzhen	Shenzhen bans motorcycles and imposes electricity restrictions [d]

[a] LLM: "Firecracker Spring Festival"; Manual: "Firecrackers"
[b] LLM: "iPhone SE"
[c] LLM: "Russian counterterrorism in Syria"; Manual: "Russia"
[d] LLM: "Shenzhen motorcycle electricity"; Manual: "Shenzhen Laws"

Table 1: Target labels and their verbalizations. Same verbalization is used across all three pools (Full, LLM, Manual) unless indicated otherwise.

Lang	Target	Against	Favor	Neutral
ca	catalonia	3988	3902	2158
	Unrelated	–	–	2087
es	catalonia	4105	4104	1868
	Unrelated	–	–	2093
et	immigration	1175	489	1597
	Unrelated	–	–	677
fr	macron	308	91	131
	lepen	466	65	55
	Unrelated	–	–	231
it	sardinia	1770	785	687
	Unrelated	–	–	673
zh	firecracker	250	250	100
	iphone	209	245	146
	russia	250	250	100
	secondbirth	200	260	140
	shenzhen	300	160	126
	Unrelated	–	–	620

Table 2: Sample counts by language, target label, and stance label (Against, Favor, Neutral). *Unrelated* samples may only be Neutral.

The free-form target is translated by an MT model into English. A set of pretrained FastText embeddings (Bojanowski et al., 2017) are used to map this translated target to a pool entry. For each target in the pool, the cosine similarity is computed between the prediction's embedding and the target's embedding. If no cosine similarity exceeds a predefined threshold τ, the prediction is mapped to a special *Unrelated* target. Otherwise, the prediction is mapped to the target with the highest cosine similarity. In practice, the sequence transduction model returns multiple predictions in its output sequence; the single prediction with the highest similarity to any pool entry is retained.

Note that while the mapped target is the target used when evaluating model performance, an end user of the system would only see the generated free-form target. Additionally, it is the free-form target that is provided to the BERTweet model for stance classification.

3.3. Data

We construct a multilingual TSE benchmark consisting of Spanish, Catalan, French, Italian, Estonian, and Mandarin samples. Samples labeled Favor, Against, or Neutral are drawn from the following existing stance corpora:

- Catalonian Independence Corpus (**cic**) *(Zotova et al., 2020)*
- French Election Corpus (**e-fra**) *(Lai et al., 2020)*
- SardiStance Corpus (**sardi**) *(Cignarella et al., 2020)*
- Estonian Immigration Corpus (**et-imm**) *(Mets et al., 2024)*
- NLPCC 2016 Stance Corpus (**nlpcc**) *(Xu et al., 2016)*

For the mapping process, the available targets from each stance corpus are combined to form a pool. Table 1 lists each target label along with its corresponding textual verbalization. Since the verbalizations used for targets can greatly influence the outcome of the mapping, three target pools are considered: Full, LLM, and Manual. Full refers to the original, precise target names from each dataset. LLM verbalizations are obtained by prompting the 20b-parameter OSS ChatGPT (OpenAI, 2025) to trim each target name to three words or fewer. Manual verbalizations are created by manually trimming the target names to shorter phrases. For most targets, the verbalization is identical across all pools.

An additional *Unrelated* class is used to provide true negatives for TSE evaluation and to avoid forcing a spurious target match when a document does not express stance toward any target. *Unrelated* samples are derived as follows: for Estonian, additional non-target posts from **et-imm** are used; for Mandarin, posts are drawn from C-Stance (Zhao et al., 2023), with target annotations ignored; and for Catalan, French, Italian, and Spanish, sentences are sampled from GlobalVoices (Tiedemann, 2012). For each language, the *Unrelated* portion is set to approximately 17% of the samples to align with prior TSE practice (Li et al., 2023b). Table 2 provides the number of samples for each language, target, and stance label.

Beyond the TSE benchmark data, 64,000 English samples are drawn from GlobalVoices for training

the FastText embeddings. This size is comparable to the ∼50,000 samples used by Li et al. (2023b) to train their embeddings. The keyphrase generation corpus KPTimes (Gallina et al., 2019) is used for target-predictor training. All the development samples and 1,000,000 training samples are translated into the six languages (1/6 of the corpus for each language).

To comply with the licenses of the component datasets, we do not directly distribute the benchmark; instead, our codebase provides scripts to automatically retrieve the datasets from their original repositories.

4. Experiments

Evaluation covers target prediction, stance detection, and overall performance. For target prediction, we report per-class F1 along with the micro- and macro-averages (F_{mic}, F_{mac}) across classes.

Following Li et al. (2023b), the macro average of F1 score over the Favor and Against classes, as shown in Eq. 1, is used to evaluate the stance classifier. This metric implicitly captures performance on the Neutral class.

$$F_{\text{avg}} = \frac{F_{\text{against}} + F_{\text{favor}}}{2} \tag{1}$$

Any samples with the *Unrelated* target are excluded from this evaluation as their stance is irrelevant. Note there are still samples that have a Neutral label but not the *Unrelated* target. F_{avg} is reported for each (language, target) pair, as well as a macro average $F_{\text{mac stance}}$.

To evaluate the TSE system as a whole, the TSE F1 metric from Li et al. (2023b) is used. Any sample with the *Unrelated* target is defined to be a negative sample; otherwise it has an actual target and is considered a positive sample. The system must predict the correct target and stance for a positive sample to obtain a true positive. Predicting any target but *Unrelated* for a negative sample yields a false positive. Predicting the *Unrelated* target for a positive sample yields a false negative. Predicting the wrong target or wrong stance for a positive sample counts both as a false positive and a false negative. From these definitions precision, recall, and F1 score can be computed. We report global, per-language, and macro-averaged F1 scores ($F_{\text{mac tse}}$).

FastText embeddings (256-d) are trained for 500 epochs on the English GlobalVoices data. For target generation, a pretrained mT5 model (Xue et al., 2021) from HuggingFace[2] is fine-tuned on the machine-translated keyphrase generation corpus for 24 hours with a batch size of 32. Validation occurs every 500 batches, and the checkpoint with the lowest validation cross-entropy is retained. Based on preliminary experiments, we choose a cosine similarity threshold of $\tau = 0.35$ for mapping free-form targets to fixed ones.

For stance classification, a BERTweet model (Nguyen et al., 2020) from HuggingFace[3] is fine-tuned for five epochs with a batch size of 32; training takes approximately 20 minutes. The checkpoint with the highest validation F_{avg} is retained and—following Li et al. (2023b)—*Unrelated* samples are excluded from stance training.

Five-fold cross-validation is performed, and results are averaged across folds. Splits are stratified to preserve—within each fold—the proportions of language, target, and stance labels observed in the full dataset. Cross-validation applies only to the stance data (not the keyphrase data), so the same keyphrase model is used across all five folds. All machine translation, both of the keyphrase corpus and of the generated targets, is done with a pretrained M2M100 model[4] (Fan et al., 2020).

One NVIDIA B200 GPU is used for keyphrase model fine-tuning, target prediction, and target translation; five NVIDIA L4 GPUs (one for each fold) are used for the remainder of the pipeline.

Target	Full	LLM	Manual
catalonia	17.96	17.97	**17.99**
immigration	**42.60**	25.66	25.31
lepen	34.77	34.73	**34.81**
macron	17.96	17.97	**17.99**
sardinia	19.86	**20.78**	20.72
firecracker	**4.25**	2.47	2.87
iphone	**42.60**	25.66	25.31
russia	6.70	30.37	**30.43**
shenzhen	18.58	**38.62**	0.00
secondbirth	0.58	0.56	**0.59**
Unrelated	22.19	**27.81**	25.37
F_{mic}	20.94	**24.19**	22.75
F_{mac}	22.14	**27.82**	21.24

Table 3: Target prediction F1 score using different target pools. Results averaged across five folds.

5. Results and Analysis

The target prediction accuracy for individual partitions of the dataset is shown in Table 3. Across target pools, performance is lower for the targets *firecracker* and *secondbirth*. Shorter verbalizations

[2] https://huggingface.co/google/mt5-base

[3] https://huggingface.co/vinai/bertweet-base

[4] https://huggingface.co/facebook/m2m100_418M

	Full		LLM		Manual	
	Mapped	GT	Mapped	GT	Mapped	GT
ca	8.25	74.45	**8.72**	74.04	8.69	74.33
es	8.24	72.25	8.43	71.79	**8.52**	71.96
et	36.25	58.54	40.98	57.74	**41.31**	57.47
fr	19.56	70.43	19.60	67.72	**20.54**	70.49
it	8.63	56.60	**9.76**	56.14	9.60	57.00
zh	16.80	45.75	**18.14**	46.02	13.36	46.61
All	12.78	67.22	**13.85**	66.73	13.43	67.11
$F_{\text{mac tse}}$	16.29	63.00	**17.60**	62.24	17.00	62.98

Table 4: TSE F1 scores for predictions by target pool (Full, LLM, Manual), comparing mapped and GT targets. Results averaged across five folds. The best result for mapped targets is **bold** while that for GT targets is underlined.

improve performance for the *russia* target, while results for the *shenzhen* legislation target vary considerably across pools. Performance varies even for targets sharing identical verbalizations across pools (e.g., *immigration*).

Language consistency between input documents and generated targets is further verified using the Lingua language detection library (Staab, 2025). Table 5 shows the per-language match rate and its macro average Avg_{lang}. For completeness, every target candidate in the model's output sequence is evaluated, not just the one that scores the highest embedding similarity with a fixed target. This evaluation occurs before target mapping, so target pools are not yet applied. Mandarin scores highest because its distinct script provides a clear signal to the language detector, while Catalan scores lowest primarily due to the Lingua model classifying the output as one of the other Romance languages. Overall accuracy supports training the keyphrase generator on machine-translated data.

Table 6 reports stance classifier performance across different (language, target) pairs. Results are more consistent than for target prediction since the classifier trains directly on stance labels, whereas the target predictor learns from a keyphrase-generation corpus, not GT targets.

Table 4 shows Target-Stance Extraction performance by language. Under the TSE F1 scheme, a sample counts as a true positive only when both the target and stance are correctly predicted for a non-*Unrelated* case. Following Li et al. (2023b), a ceiling condition replaces predicted targets with GT targets. Performance is lowest for Catalan, Spanish, and Italian, reflecting poor recall on *catalonia* and *sardinia*. The baseline achieves F1 scores between 0.10 and 0.20, whereas models on the original English TSE benchmark report F1 scores between 0.30 and 0.40 (Li et al., 2023b). This gap illustrates the challenge of multilingual TSE and identifies target prediction as the bottleneck.

ca	es	et	fr	it	zh	Avg_{lang}
71.28	78.59	87.14	77.76	83.14	91.90	81.64

Table 5: Language match rate of generated targets.

Lang-Target	Full	LLM	Manual
ca-catalonia	75.50	75.08	**75.73**
es-catalonia	**69.85**	69.08	69.36
et-immigration	27.25	**28.64**	26.83
fr-lepen	49.33	51.09	**53.29**
fr-macron	43.17	**45.00**	40.82
it-sardinia	50.98	50.81	**51.97**
zh-firecracker	40.25	41.17	**41.98**
zh-iphone	**42.01**	40.76	34.53
zh-russia	40.55	**44.98**	40.62
zh-shenzhen	49.92	**50.20**	48.21
zh-secondbirth	34.97	32.13	**41.80**
$F_{\text{mac stance}}$	47.62	**48.09**	47.74

Table 6: F_{avg} scores for stance models across target pools (five-fold average).

Target predictions for randomly selected samples are shown in Table 7. None of the generated targets for the Catalan nor the Spanish sample explicitly mention Catalonia, making mapping to *Catalonian Independence* difficult. The same holds for the Italian sample and *Sardinian Independence*. The single generated target for the Estonian sample explicitly mentions north African nations as well as a migration crisis, easily mapping to *Immigration*. For the French sample, the model notes the named entity *Marine LePen* and explicitly generates a corresponding target that is mapped correctly. One of the generated Mandarin candidates explicitly mentions Russia, allowing it to map to the Russian counterterrorism target.

Lang	Document	Generated Targets	Mapped Target	Groundtruth Target
ca	🔊Trapote insisteix en que era el cap de secció qui decidia en quines localitats s'intervenia #JudiciProcés	super bowl;trapote; futbol;nfl;super bowl 2015	*Unrelated*	Catalonian Independence
es	4. Y si Arrimadas no quiere ir de dos y aspira al 1. Un discurso de la primera mujer presidenta de España. La ambición de Rivera es desmesurada, aunque quizás fuera el camino para tocar poder.	política y gobierno;rivera mariana;elecciones presidenciales de 2008	Russia's counter-terrorism operations in Syria	Catalonian Independence
et	Tuneesia ja Liibüa on peamised punktid, kust migrandid asutavad end ohtlikule mereretkele Põhja-Aafrikast Euroopasse, eelkõige Itaaliasse.	tuneesia, liibüa, euroopa, rahvusvahelised suhted, rahvusvahelised suhted, lähis-ida ja aafrika rände kriis, euroopa rände kriis, euroopa rände kriis	Immigration	Immigration
fr	screenshot exclusif provenant du téléphone de Marine Le Pen	marine le pen;marine le pen communications inc;computer et internet	Marine LePen	Marine LePen
it	Datemi un nome! Un solo nome di un leghista che abbia fatto qualcosa di buono o intelligente riconosciuto dal mondo o dagli italiani. Non ce ne. E quindi loro sono semplicemente uno scarto. Una razza stupida. Gente che vale meno dello sterco del mulo di mio nonno. Fanculo	scienza e tecnologia; razza e etnia;nomi geografici	*Unrelated*	Sardinian Independence
zh	中俄战略伙伴关系，不是少数别有用心的中国人，以及冒充中国人的美日等西方网络水军、第五纵队，所能离间分化的	中俄战略伙伴关系;中国;军事	Russia's counter-terrorism operations in Syria	Russia's counter-terrorism operations in Syria

Table 7: Randomly selected target predictions using the Full pool for verbalizations. The chosen candidate is underlined.

6. Conclusions and Future Directions

This work introduces the first multilingual TSE benchmark and a baseline system spanning six languages. The baseline serves as a reference point for future TSE models. Although our approach achieves lower metric values than English-only results (Li et al., 2023b), the resource establishes a crucial starting point for multilingual TSE. We are also the first to demonstrate the sensitivity of TSE's main metric to target verbalizations. Most errors arise from incorrect target predictions; in contrast, stance detection on these corpora is well studied.

Three complementary avenues for improvement are identified: (a) *Multilingual keyphrase corpora*; (b) *Target mapping and embeddings*; and (c) *Broader evaluation datasets*.

(a) Multilingual keyphrase corpora. Fine-tuning the keyphrase model on genuine multilingual keyphrase data—rather than machine-translation text—will likely improve performance. Available resources exist, though they tend to be domain-specific (e.g., law (Salaün et al., 2024); scientific research (Piedboeuf and Langlais, 2022)). As a complementary avenue, the generative target predictor could be replaced with an extractive span model; however, span extraction cannot produce targets that do not appear in the input.

(b) Target mapping and embeddings. There are other means of strengthening the mapping step, for example, by adopting higher-quality multilingual contextual embeddings for similarity computation (Akash et al., 2025). When translation remains in the loop, using higher-quality MT can reduce drift.

Alternatively, during keyphrase generation training, the model could be fine-tuned to produce English keyphrases from multilingual input. Although this reduces the multilinguality of the target predictor, it removes the need for a translation step during mapping. In terms of data preparation, this would require keeping the KPTimes keyphrase labels in English and translating only the corresponding input documents.

Even under this setup, FastText embeddings would still be required for target mapping. Future work could also systematically examine target verbalizations, as automated evaluation is likely sensitive to them—even in a monolingual setting.

(c) Broader evaluation datasets. Adapting zero-shot stance resources to TSE (Allaway and McKeown, 2020; Zhao and Caragea, 2024) would provide broader coverage and stronger generalization checks. To move beyond a small, fixed target pool, evaluation could test growing pools, metrics at multiple pool sizes, and open-world settings that allow unseen targets.

Taken together, these directions provide a concrete path toward stronger multilingual TSE systems and more informative evaluations.

Limitations

While our benchmark does cover six different languages, four of them (Catalan, French, Italian, Spanish) are Romance languages, potentially limiting the applicability of our results to other groups of languages. Additionally, with the exception of the Catalonian Independence target, no targets are shared across languages. Even within a specific language, using a different set of targets than

those in our study could lead to different levels of performance, particularly during target mapping. Annotating a new multilingual dataset specifically designed for TSE, with a wider variety of languages and targets, would address these issues inherent in collating existing stance corpora.

Ethical Considerations

While our benchmark is formed from publicly available stance corpora, we do not redistribute the data and thus adhere to copyright restrictions and social network privacy policies.

Regarding application, target-stance extraction is an excellent tool for good actors investigating public opinion on relevant issues. Unfortunately, this same tool could enable bad actors on social media to find individuals expressing a particular opinion and target them for harassment, or worse. One future line of work to mitigate this is adversarial Target-Stance Extraction: altering a document to preserve its meaning while causing a misclassification from the target predictor or stance classifier.

Bibliographical References

Abu Ubaida Akash, Ahmed Fahmy, and Amine Trabelsi. 2025. Can large language models address open-target stance detection? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 971–985, Vienna, Austria. Association for Computational Linguistics.

Emily Allaway, Malavika Srikanth, and Kathleen McKeown. 2021. Adversarial learning for zero-shot stance detection on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4756–4767, Online. Association for Computational Linguistics.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885, Austin, Texas. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Hongzhou Chen, Ke Yan, Mustafa Raad Kadhim, Kui Wu, and Ling Tian. 2024. SentKB-BERT: Sentiment-filtered Knowledge-based Stance Detection. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. ISSN: 2161-4407.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *CoRR*, abs/2010.11125.

Ygor Gallina, Florian Boudin, and Beatrice Daille. 2019. KPTimes: A large-scale dataset for keyphrase generation on news documents. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 130–135, Tokyo, Japan. Association for Computational Linguistics.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. Few-Shot Cross-Lingual Stance Detection with Sentiment-Based Pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10729–10737. Number: 10.

Zihao He, Negar Mokhberian, and Kristina Lerman. 2022. Infusing knowledge from Wikipedia to enhance stance detection. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 71–77, Dublin, Ireland. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Ang Li, Bin Liang, Jingqian Zhao, Bowen Zhang, Min Yang, and Ruifeng Xu. 2023a. Stance detection on social media with background knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15703–15717, Singapore. Association for Computational Linguistics.

Yingjie Li, Krishna Garg, and Cornelia Caragea. 2023b. A new direction in stance detection: Target-stance extraction in the wild. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10071–10085, Toronto, Canada. Association for Computational Linguistics.

Bin Liang, Yonghao Fu, Lin Gui, Min Yang, Jiachen Du, Yulan He, and Ruifeng Xu. 2021. Target-adaptive Graph for Cross-target Stance Detection. In *Proceedings of the Web Conference 2021*, WWW '21, pages 3453–3464, New York, NY, USA. Association for Computing Machinery.

Bin Liang, Qinglin Zhu, Xiang Li, Min Yang, Lin Gui, Yulan He, and Ruifeng Xu. 2022. JointCL: A joint contrastive learning framework for zero-shot stance detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 81–91, Dublin, Ireland. Association for Computational Linguistics.

Rui Liu, Zheng Lin, Yutong Tan, and Weiping Wang. 2021. Enhancing zero-shot and few-shot stance detection with commonsense knowledge graph. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3152–3157, Online. Association for Computational Linguistics.

Yun Luo, Zihan Liu, Yuefeng Shi, Stan Z. Li, and Yue Zhang. 2022. Exploiting sentiment and common sense for zero-shot stance detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7112–7123, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Brodie Mather, Bonnie J Dorr, Owen Rambow, and Tomek Strzalkowski. 2021. A general framework for domain-specialization of stance detection: A covid-19 response use case. *The International FLAIRS Conference Proceedings*, 34.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card.

Peter F. M. Staab. 2025. lingua. Computer software. Version 2.1.1.

Penghui Wei and Wenji Mao. 2019. Modeling Transferable Topics for Cross-Target Stance Detection. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, pages 1173–1176, New York, NY, USA. Association for Computing Machinery.

Wan Wei, Xiao Zhang, Xuqin Liu, Wei Chen, and Tengjiao Wang. 2016. pkudblab at SemEval-2016 task 6 : A specific convolutional neural network system for effective stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 384–388, San Diego, California. Association for Computational Linguistics.

Chang Xu, Cécile Paris, Surya Nepal, and Ross Sparks. 2018. Cross-target stance classification with self-attention networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 778–783, Melbourne, Australia. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Naiyu Yan, Shaobin Huang, and Rongsheng Li. 2025. BCLTC: Bi-directional curriculum learning based tasks collaboration for target-stance extraction. *Information Processing & Management*, 62(4):104137.

Guido Zarrella and Amy Marsh. 2016. MITRE at SemEval-2016 task 6: Transfer learning for stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 458–463, San Diego, California. Association for Computational Linguistics.

Chenye Zhao and Cornelia Caragea. 2025. Bilingual zero-shot stance detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29900–29919, Vienna, Austria. Association for Computational Linguistics.

Qinglin Zhu, Bin Liang, Jingyi Sun, Jiachen Du, Lanjun Zhou, and Ruifeng Xu. 2022. Enhancing Zero-Shot Stance Detection via Targeted Background Knowledge. In *Proceedings of the 45th International ACM SIGIR Conference on Research*

and Development in Information Retrieval, SIGIR '22, pages 2070–2075, New York, NY, USA. Association for Computing Machinery.

Language Resource References

Emily Allaway and Kathleen McKeown. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931, Online. Association for Computational Linguistics.

Alessandra Teresa Cignarella, Mirko Lai, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2020. SardiStance @ EVALITA2020: Overview of the Task on Stance Detection in Italian Tweets. In Valerio Basile, Danilo Croce, Maria Maro, and Lucia C. Passaro, editors, *EVALITA Evaluation of NLP and Speech Tools for Italian - December 17th, 2020*, pages 177–186. Accademia University Press, Torino.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. Will-they-won't-they: A very large dataset for stance detection on Twitter. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1724, Online. Association for Computational Linguistics.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance detection in COVID-19 tweets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1596–1611, Online. Association for Computational Linguistics.

Mirko Lai, Alessandra Teresa Cignarella, Delia Irazú Hernández Farías, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2020. Multilingual stance detection in social media political debates. *Computer Speech & Language*, 63:101075.

Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021. P-stance: A large dataset for stance detection in political domain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2355–2365, Online. Association for Computational Linguistics.

Bin Liang, Ang Li, Jingqian Zhao, Lin Gui, Min Yang, Yue Yu, Kam-Fai Wong, and Ruifeng Xu. 2024. Multi-modal stance detection: New datasets and model. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12373–12387, Bangkok, Thailand. Association for Computational Linguistics.

Mark Mets, Andres Karjus, Indrek Ibrus, and Maximilian Schich. 2024. Automated stance detection in complex topics and small languages: The challenging case of immigration in polarizing news media. *PLOS ONE*, 19(4):e0302380. Publisher: Public Library of Science.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Fuqiang Niu, Zebang Cheng, Xianghua Fu, Xiaojiang Peng, Genan Dai, Yin Chen, Hu Huang, and Bowen Zhang. 2024. Multimodal multi-turn conversation stance detection: A challenge dataset and effective model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 3867–3876, New York, NY, USA. Association for Computing Machinery.

Frédéric Piedboeuf and Philippe Langlais. 2022. A new dataset for multilingual keyphrase generation. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Olivier Salaün, Frédéric Piedboeuf, Guillaume Le Berre, David Alfonso-Hermelo, and Philippe Langlais. 2024. EUROPA: A legal multilingual keyphrase generation dataset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12718–12736, Bangkok, Thailand. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Jannis Vamvas and Rico Sennrich. 2020. X-stance: A multilingual multi-target dataset for stance detection. *CoRR*, abs/2003.08385.

Ruifeng Xu, Yu Zhou, Dongyin Wu, Lin Gui, Jiachen Du, and Yun Xue. 2016. Overview of NLPCC Shared Task 4: Stance Detection in Chinese

Microblogs. In *Natural Language Understanding and Intelligent Applications*, pages 907–916, Cham. Springer International Publishing.

Chenye Zhao and Cornelia Caragea. 2024. EZ-STANCE: A large dataset for English zero-shot stance detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15697–15714, Bangkok, Thailand. Association for Computational Linguistics.

Chenye Zhao, Yingjie Li, and Cornelia Caragea. 2023. C-STANCE: A large dataset for Chinese zero-shot stance detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13369–13385, Toronto, Canada. Association for Computational Linguistics.

Jonathan Zheng, Ashutosh Baheti, Tarek Naous, Wei Xu, and Alan Ritter. 2022. Stanceosaurus: Classifying stance towards multicultural misinformation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2132–2151, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Elena Zotova, Rodrigo Agerri, Manuel Nuñez, and German Rigau. 2020. Multilingual stance detection in tweets: The Catalonia independence corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1368–1375, Marseille, France. European Language Resources Association.

Appendix A. LLM Prompt

When prompting GPT-OSS (OpenAI, 2025) to trim the target verbalizations, we use the following prompt:

```
I'm developing an algorithm to classify
a document's stance (favor, against,
neutral) toward the following ten
targets.  Can you reduce each of these
targets to 2-3 words (and leave them
completely unchanged if they're already
in that range) while still capturing the
original intent?
Emmanuel Macron
Marine LePen
Immigration
IphoneSE
Russia's counter-terrorism operations in
Syria
Allowing second births
Setting off firecrackers during the
Spring Festival
Shenzhen bans motorcycles and imposes
electricity restrictions
Sardinian Independence
Catalonian Independence
```