CoSt-BR: a Language Resource for Conversational Stance Detection

Felipe Fonseca, Ivandré Paraboni, Luciano Digiampietri
Escola de Artes, Ciências e Humanidades - Universidade de São Paulo
03828-000 – São Paulo – SP – Brazil
{felipe.fonseca, ivandre, digiampietri}@usp.br

Abstract

Stance detection is the computational task of determining the attitude (e.g., for, against, neutral) expressed in text toward a specific target topic. In its more conventional form, the task focuses on isolated, context-free input utterances. Conversational stance detection, by contrast, analyzes messages embedded within dialogue threads, enabling the interpretation of responses in relation to preceding discourse, and takes into account a greater variety of stance relations (e.g., support, deny, query, comment, etc.). Despite growing research attention, however, conversational stance detection remains relatively under-resourced and largely limited to the English language. To address these gaps, this study introduces CoSt-BR, a new corpus for conversational stance detection composed of a large set of annotated Reddit discussions in Brazilian Portuguese. In addition, the paper also reports benchmark results obtained using various computational methods, including supervised and prompt-based strategies, applied to the corpus data, providing baseline references for future research in this area.

Keywords: Stance detection, Conversational stance, Corpora

1. Introduction

Stance detection is the computational task of determining the attitude (e.g., in favor, against, neutral, etc.) expressed in a given text towards a specific target topic (AlDayel and Magdy, 2021). For example, given the target 'gambling taxation', the utterance "Gambling money should be channeled to social services" conveys a stance in favor of the target.

In its more standard task definition, stance detection primarily focuses on the interpretation of single utterances presented without contextual information, such as individual social media posts (Mohammad et al., 2016). This constitutes a well-established NLP task with a substantial body of available resources, most of which are dedicated to the English language (AlDayel and Magdy, 2021; Alturayeif et al., 2023).

Conversational stance detection, by contrast, involves interpreting messages situated within a conversational context, such as discussion forum exchanges (Derczynski et al., 2017), and typically includes a broader set of possible labels (e.g., *Support*, *Deny*, *Query*, and *Comment* in (Derczynski et al., 2017)). Studies of this kind remain comparatively under-resourced despite increasing research interest in the field. This is particularly true given the emergence of several conversation-aware formulations of the task, driven by newly developed dialogue-centered corpora (Gorrell et al., 2018; Derczynski et al., 2017; Zheng et al., 2022; Lavrouk et al., 2024), which, once again, are often focused on the English language.

Building on these observations, this study introduces a novel conversational stance corpus for the Portuguese language, hereafter referred to as CoSt-BR (Conversational Stance in Brazilian Portuguese). The corpus comprises a large collection of Reddit conversations discussing controversial topics (e.g., abortion, sexuality, drug use, etc.) and extends the set of available languages addressed by existing resources such as RumourEval (Derczynski et al., 2017) and Stanceosaurus (Zheng et al., 2022).

The remainder of this article is organized as follows. Section 2 reviews related work on conversational stance corpora. Section 3 introduces CoSt-BR and presents its main features. Section 4 describes the corpus annotation. Section 5 reports a set of preliminary experiments conducted using the corpus and provides baseline results for future studies. Finally, Section 6 presents conclusions and outlines directions for future research.

2. Related work

Conversational stance was first introduced as an NLP task in the RumourEval shared task, also known as SemEval 2017 Task 8 (Derczynski et al., 2017). The accompanying dataset comprised Twitter conversations organized in dialogue threads, enabling the interpretation of individual replies in relation to preceding messages. Each message was annotated with one of four stance categories: *Support*, *Deny*, *Query*, and *Comment*. In addition, the task provided contextual rumor metadata for each topic. The training data included 297 source tweets initiating conversations and 4,222 response tweets (4,519 tweets in total) across eight rumors, whereas the test set comprised 28 conversations covering 20 rumors already seen in the training

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8141–8146
11–16 May 2026.

data and 8 previously unseen ones.

The second version of RumourEval (also referred to as SemEval 2019 Task 7) improved upon the initial edition by correcting errors in the original dataset, refining the evaluation methodology through the adoption of the F1-score metric instead of accuracy, and incorporating Reddit threads to enhance conversational diversity. This updated dataset contains 6,702 examples for training and 1,872 examples for testing.

The Stanceosaurus corpus (Zheng et al., 2022) addresses the closely related problem of fact-checking and comprises 28,083 tweets associated with 251 misinformation claims in English, Hindi, and Arabic. Corpus instances are organized as claim-centric Twitter conversations, in which each claim functions as a rumor under discussion. A key distinction from RumourEval lies in its definition of five stance categories: *Irrelevant*, *Supporting*, *Refuting*, *Discussing*, and *Querying*. The newly introduced *Irrelevant* class designates posts that are off-topic with respect to the initial claim. Furthermore, the Stanceosaurus 2.0 corpus (Lavrouk et al., 2024) extends the original corpus by incorporating data in Russian and Spanish.

The SCRum-9 corpus (Li et al., 2025) is a multilingual, claim-centred Twitter dataset comprising 7,156 instances, and constitutes the only available resource of its kind to include Portuguese text (alongside eight additional languages: Czech, German, English, Spanish, French, Hindi, Polish, and Russian). The Portuguese subset contains 1,000 instances. Annotations follow a rumor-verification scheme with four stance labels: *Confirming*, *Rejecting*, *Questioning*, and *Commenting*.

The main distinctions between SCRum-9 and the present work are twofold. First, SCRum-9 is focused on rumor fact-checking, whereas the present study aims to address conversational stance in a broader sense. Second, the SCRum-9 corpus lacks a multi-turn conversational structure, as it includes only direct replies to source tweets. This design choice is based on the assumption that deeper reply chains tend to become less informative or drift away from the original rumor topic. Consequently, SCRum-9 provides at most one level of replies and does not capture the richer hierarchical conversational context necessary for conversational stance detection.

Finally, with respect to our target language – Brazilian Portuguese – we notice that, although stance corpora do exist such as UstanceBR (Pereira et al., 2026) and UlyssesSD (Maia et al., 2022), these are generally not organized as conversation threads and therefore do not support conversational stance detection.

3. Corpus design

The CoSt-BR corpus was constructed from publicly available user comments on *Reddit*, specifically from the Brazilian communities *r/brasil*, *r/brasilivre*, and *r/FilosofiaBAR*. These communities were selected for their prominence as some of the most active Portuguese-speaking spaces on the platform. To ensure relevance and engagement, only comments from *threads* listed under the *hot* tab, representing recent and high-visibility discussions, were included. Furthermore, eligible threads contained at least 100 comments and addressed topics likely to evoke strong opinions, namely abortion, sexuality, drug use, and rumors concerning Brazil's *Pix* electronic payment system. These topics were chosen due to their tendency to elicit explicit stances from participants, facilitating the identification of agreement and disagreement within discussions.

This procedure resulted in a total of 11 conversation *threads* collected between September 2023 and August 2024, comprising 2,266 posts. Following the approach adopted in RumourEval (Gorrell et al., 2018), stances within each conversation were defined relative to the first comment, acknowledging that a single *thread* may give rise to multiple conversational branches. Data collection adhered to the platform's terms of use, and no personally identifiable information was retained.

In the CoSt-BR framework, five stance categories adapted from previous work in conversational stance detection (Derczynski et al., 2017; Gorrell et al., 2018; Zheng et al., 2022) are considered: *Supporting*, *Refuting*, *Discussing*, *Querying*, and *Irrelevant*. These are defined as follows:

- **Supporting**: the message explicitly or implicitly agrees with the claim.
- **Refuting**: the message explicitly or implicitly disagrees with the claim.
- **Discussing**: the message is related to the topic of the claim, but it remains neutral.
- **Querying**: the message requests additional information about the claim, aiming to clarify doubts.
- **Irrelevant**: the message addresses a subject other than the topic of the claim, or is not related to the claim.

The corpus structure follows the framework adopted in RumourEval (Derczynski et al., 2017), where the first comment t of each conversation is treated as the central object or target of the stance, and all subsequent comments are evaluated with respect to t. As there are cases where the initial comment is phrased as a question – thus lacking a clear proposition toward which later comments

You are a helpful assistant that given a context message and a follow up message (human message), create a new message that encapsulate what the follow up sentence means regarding the context message.

The context message is the following: context.

Your job is to create a new message that conveys what the follow up message is saying but taking in consideration the context message. Remember, the generated message must be in Brazilian Portuguese. If any of the message have swear or is offensive, ignore the offensive part and create a message that only keeps the important information.

Let's begin!

Just generate the new message like you are the user that created the follow up message. Do not include anything else. Do not write saying you or você in Portuguese. Write everything in first person.

Follow up question: input

Figure 1: Prompt used for target text generation.

could express a stance – the first comment for each thread was reformulated into an explicit declarative statement.

Comment reformulation was carried out using *Llama 3.1 8B*[1] , followed by manual post-editing to ensure accuracy and naturalness. Figure 1 illustrates the prompt used for this process, where the blue context field corresponds to the title of the current Reddit thread and the red input field represents the first comment of a given conversation.

Messages generated for each conversation were reviewed by a human annotator to ensure that the text was relevant to the discussion and could appropriately serve as a target claim for the entire conversation.

In this setting, conversational contexts were defined according to the concept of *diffusion* described in (Largeron et al., 2021), understood as an ordered sequence of messages within a conversation. A diffusion begins with an initial post on social media (in the present corpus, this corresponds to the first comment that initiates the discussion) and ends with a leaf comment, including all intermediate messages that connect the original post to that leaf. In this representation, every comment is associated with a diffusion, that is, a path linking the comment back to the original post, and not only the leaf nodes.

Figure 2 illustrates a conversation fragment from the corpus, where each element u denotes a comment within a conversation thread. The element $claim$ represents the target claim against which stances are classified, and the element $u1$ corresponds to the initial comment that triggered the discussion. The conversation comprises the comments $(u1, ..., u5)$, where, for example, the diffusion of $u2$ is $u1, u2$, and that of $u4$ is $u1, u3, u4$.

Based on these definitions, conversational stance detection in the CoSt-BR corpus consists of leveraging the diffusion δ_j to determine the stance of a given comment u_j.

[1] https://groq.com/

Claim: Nowadays abortion is allowed in cases of rape, therefore Rosa Weber's comment that motherhood is a choice is valid given that it is already permitted in those scenarios.

u1: And abortion is already allowed in cases of rape. **[Supporting]**

—**u2:** No. It is allowed in cases of an allegation of rape, currently there is no obligation to report to the police, in fact the hospital is prohibited from reporting and initiating a police investigation. In other words, if a woman goes there and lies that she was raped she can have an abortion with no consequences. **[Refuting]**

—**u3:** But conservative cuckolds already want to ban it even in those cases. Even if a child is diagnosed with a rare disease that will pose a risk of death to the mother, and with a chance of being born dead, they want to prohibit abortion. **[Supporting]**

———**u4:** Strawman fallacy... **[Irrelevant]**

———**u5:** If it is born, that means it does not represent a risk. If it then died, sad and let it be buried. Besides, you want to use 0.0001% of cases to justify the murder of innocents. **[Refuting]**

Figure 2: A fragment of a conversation from the CoSt-BR corpus (translated to english).

4. Corpus annotation

The corpus was annotated by five judges. These were undergraduate students, three of whom identified as male and two as female. The final label for each example was determined using a majority voting scheme. Since not all examples were annotated by all judges (see Table 1), ties occasionally occurred when multiple labels received an equal number of votes. In such cases, an additional judge, not involved in the original annotation task, was responsible for assigning the final label.

Table 2 summarizes the number of judges involved in the annotation of different portions of the data, with 62% of the data being annotated by all five judges.

Annotators	Data %
2	2%
3	8%
4	28%
5	62%

Table 1: Annotation coverage.

Inter-annotator agreement was computed by using Fleiss' kappa (Warrens, 2010). To this end, instances with fewer than four annotations were removed from the analysis and, conversely, those that were annotated by all five raters had one random annotation left out. Thus, the agreement analysis covered about 90% of the corpus instances, as illustrated in the previous Table 1.

The Fleiss' kappa score for the 5-class annotation scheme was found to be 0.323. Descriptive statistics of the final dataset are summarized in Table 2.

The corpus data is available on https://github.com/felipepenhorate/cost-br.

Class	Train	Test
Supporting	647	295
Refuting	452	170
Querying	29	23
Discussing	177	153
Irrelevant	229	91

Table 2: Corpus descriptive statistics.

5. Evaluation

As a means to assess the practical use of the CoSt-BR data and to provide reference results for future work, we evaluated a number of conversational stance detection models on the corpus data. These models and their results are described in the following sections.

5.1. Conversational stance detection models

We designed four conversational stance detection models, referred to as St-BERT, GPT-OSS-prompt, Phi-4-prompt, and Phi-4-IFT, which are described in the following.

The St-BERT model is a standard BERT-based stance classifier that fine-tunes the Portuguese BERT model BERTimbau (Souza et al., 2020), following the approach used in the evaluation of the Stanceosaurus dataset (Zheng et al., 2022). To address class imbalance, we employed Class-Balanced Focal Loss during training (Cui et al., 2019).

The GPT-OSS-prompt and Phi-4-prompt models adapt the prompt-engineering method proposed in (de Fonseca et al., 2023) for conversation-level stance classification. Specifically, the original prompt was modified to include the claim associated with each example and to explicitly model the stance relationship between the utterance and the claim.

The prompt template used for both models is shown in Figure 3, where brown tokens represent the claim, blue tokens indicate the conversational context, green tokens denote the current message, and red tokens correspond to the model output (i.e., the predicted stance class).

The prompt was submitted to both the GPT-OSS (OpenAI et. al., 2025) and Phi-4 models (Abdin et al., 2024), with reasoning effort set to High, resulting in the GPT-OSS-prompt and Phi-4-prompt variants.

The Phi-4-IFT model builds on the same prompt, but replaces the zero-shot or few-shot setting with instruction fine-tuning (Wei et al., 2022). In this approach, Phi-4 was fine-tuned on prompt–response pairs derived from the training data, allowing it to follow the task-specific instructions more consistently.

You are an evaluator tasked with analyzing a conversation about a claim on Reddit. Your job is to determine whether the current message is Supporting, Refuting, Querying, Discussing, or Irrelevant in relation to the subject of the claim. You must respond with only one of the following stances: Supporting, Refuting, Querying, Discussing, or Irrelevant.

When a user supports the claim, they MUST be classified as Supporting. This can be shown through their own messages or when they express agreement with another user who supports the claim.

When a user questions the validity of the claim, considers it false, or asks others to refute it, they MUST be classified as Refuting. This can be shown through their own messages or when they express agreement with another user who denies the claim.

When a user asks for new information regarding the claim, they MUST be classified as Querying.

When a user does not add new information about the claim and does not express support or opposition toward another user, they MUST be classified as Discussing.

When a user talks about something unrelated to the claim or the conversation, they MUST be classified as Irrelevant

The claim being discussed is the following "I don't understand how it is possible to remove a fundamental right like same-sex marriage, which harms no one. This is a setback for Brazilian society."

Conversation:

Elijah said "[URL-about-marriage]"
Elijah is Discussing the subject
Keith answered Elijah "Quais são os beneficios do governo? Quem casa é cartorio, igreja tem o livre arbitrio para casar ou não casais homossexuais."

Given the conversation, what is Amelia stance regarding the claim? Think step by step and answer with ONLY one word: Supporting, Refuting, Querying, Discussing or Irrelevant.
Answer: Supporting

Figure 3: Example of prompt filled with data from CoSt-BR (translated to english).

The results of the instruction fine-tuned model are compared directly with those of the prompt-only variants to evaluate which strategy offers the best stance classification performance.

5.2. Results

Table 3 shows the results obtained by the models St-BERT, GPT-OSS-prompt, Phi-4-prompt and Phi-4-IFT described in the previous section.

Model	Precision	Recall	F1
St-BERT	42.1	36.5	36.9
GPT-OSS-prompt	46.6	53.2	48.4
Phi-4-prompt	47.9	52.7	46.1
Phi-4-IFT	67.8	56.7	58.3

Table 3: Overall results.

From the results in Table 3, we notice that the Instruction Fine-Tuning strategy outperforms the alternatives by a considerable margin. Table 4 provides a breakdown of these results for each class.

Class	Precision	Recall	F1	#examples
Discussing	67.4	18.9	29.5	153
Irrelevant	57.2	64.8	60.8	91
Querying	93.3	60.8	73.6	23
Refuting	61.7	54.1	57.6	170
Supporting	59.2	84.7	69.7	295

Table 4: Phi-4-IFT results per class.

Results in Table 4 show that the *Discussing* class represents a significant source of error in the task

as measured by the F1 score. More details are illustrated by the confusion matrix for model Phi-4-IFT in Table 5.

	D	R	Q	S	I
D	29	26	1	81	16
R	7	92	0	62	9
Q	4	2	14	2	1
S	2	25	0	250	18
I	1	4	0	27	59

Table 5: Phi-4 IFT confusion matrix: (D)iscussing; (R)efuting; (Q)uerying; (S)upporting; (I)rrelevant.

Table 5 shows that instances of the *Discussing* class are frequently misclassified as *Supporting*, which is the majority class in the corpus. This suggests a tendency of the model to default to the dominant class when uncertainty arises.

6. Final remarks

This article introduced CoSt-BR, a novel corpus for conversational stance detection comprising 2,266 examples of Reddit discussions manually annotated using a five-class scheme. The corpus builds upon prior work such as the English RumourEval conversational stance dataset (Derczynski et al., 2017) and, to the best of our knowledge, constitutes the first resource of its kind developed for Brazilian Portuguese.

In addition to presenting the corpus, we reported initial benchmark results obtained using a range of baseline systems, including fine-tuning, prompt-based, and instruction fine-tuning approaches. These results are intended to serve as reference points and to encourage further research leveraging this dataset.

Our preliminary experiments point to several promising directions for future work beyond traditional stance classification methods. Notably, our findings indicate that generative models with instruction fine-tuning may achieve superior performance for this task. However, more conventional approaches—such as BERT-based fine-tuning—also warrant closer examination, as their performance, while comparatively less robust, may benefit from further optimization and adaptation.

7. Acknowledgements

This paper was partially funded by the CNPq project number 302164/2025-1.

8. Bibliographical References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. 2024. Phi-4 technical report. *arXiv*, 2412.08905.

Abeer AlDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.

Nora Alturayeif, Hamzah Luqman, and Moataz Ahmed. 2023. A systematic review of machine learning techniques for stance detection and its applications. *Neural Comput. Appl.*, 35(7):5113–5144.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In *CVPR*.

Felipe de Fonseca, Ivandré Paraboni, and Luciano Digiampietri. 2023. Contextual stance classification using prompt engineering. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 33–42, Porto Alegre, RS, Brasil. SBC.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, Vancouver, Canada. Association for Computational Linguistics.

Genevieve Gorrell, Kalina Bontcheva, Leon Derczynski, Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2018. Rumoureval 2019: Determining rumour veracity and support for rumours.

Christine Largeron, Andrei Mardale, and Marian-Andrei Rizoiu. 2021. Linking the dynamics of user stance to the structure of online discussions. In *Advances in Intelligent Data Analysis XIX - 19th International Symposium on Intelligent Data Analysis, IDA 2021, Porto, Portugal, April 26-28, 2021, Proceedings*, volume 12695 of *Lecture Notes in Computer Science*, pages 275–286. Springer.

Anton Lavrouk, Ian Ligon, Tarek Naous, Jonathan Zheng, Alan Ritter, and Wei Xu.

2024. Stanceosaurus 2.0: Classifying stance towards russian and spanish misinformation.

Yue Li, Jake Vasilakes, Zhixue Zhao, and Carolina Scarton. 2025. Scrum-9: Multilingual stance classification over rumours on social media.

Dyonnatan Ferreira Maia, Nádia Felix Felipe da Silva, Ellen Polliana Ramos Souza, Augusto Sousa Nunes, Lucas Caetano Procópio, Guthemberg da S Sampaio, Márcio de Souza Dias, Adrio Oliveira Alves, Dyéssica F Maia, Ingrid Alves Ribeiro, Fabíola Souza Fernandes Pereira, and André Carlos Ponce de Leon Ferreira de Carvalho. 2022. Ulyssessd-br: stance detection in brazilian political polls.

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.

OpenAI et. al. 2025. gpt-oss-120b & gpt-oss-20b Model Card. *arXiv*, 2508.10925.

Camila Pereira, Matheus Pavan, Sungwon Yoon, Ricelli Ramos, Pablo Costa, Laís Cavalheiro, and Ivandré Paraboni. 2026. UstanceBR: a social media language resource for stance prediction. *Language Resources and Evaluation*, 60(14).

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: Pretrained bert models for brazilian portuguese. In *Intelligent Systems*, pages 403–417, Cham. Springer International Publishing.

Matthijs J. Warrens. 2010. Inequalities between multi-rater kappas. *Advances in Data Analysis and Classification*, 4:271–286.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

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.