Stanceosaurus 2.0: Classifying Stance Towards Multicultural Misinformation

Jonathan Zheng[*], **Anton Lavrouk**[*], **Tarek Naous, Ashutosh Baheti, Ian Ligon, Alan Ritter, Wei Xu**
College of Computing
Georgia Institute of Technology
{jonathanqzheng, antonlavrouk, tareknaous, abaheti3, iligon3}@gatech.edu; {alan.ritter, wei.xu}@cc.gatech.edu

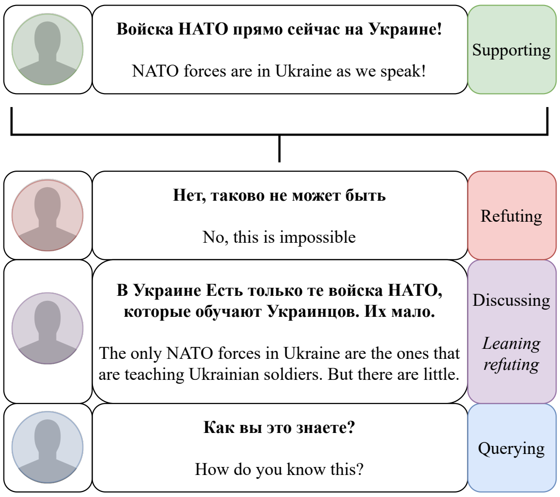
1 Introduction

We present Stanceosaurus, a new corpus of 31,904 tweets in English, Hindi, Arabic, Russian, and Spanish annotated with stance towards 292 misinformation claims. As far as we are aware, it is the largest corpus annotated with stance towards misinformation claims. The claims in Stanceosaurus originate from 21 fact-checking sources that cover diverse geographical regions and cultures. Unlike existing stance datasets, we introduce a more fine-grained 5-class labeling strategy with additional subcategories to distinguish implicit stance. Pre-trained transformer-based stance classifiers that are fine-tuned on our corpus show good generalization on unseen claims and regional claims from countries outside the training data. Cross-lingual experiments demonstrate Stanceosaurus' capability of training multilingual models, achieving 53.1 F1 on Hindi, 50.4 F1 on Arabic, 43.9 F1 on Russian, and 43.8 F1 on Spanish without any target language fine-tuning. Finally, we show how a domain adaptation method can be used to improve performance on Stanceosaurus using additional RumourEval-2019 data. We make Stanceosaurus publicly available to the research community and hope it will encourage further work on misinformation identification across languages and cultures.

2 Dataset Construction

Our corpus consists of social media posts manually annotated for stance toward claims from multiple fact-checking websites across the world. We carefully designed the data collection and annotation scheme to ensure better quality and coverage.

Misinformation Claims: Misinformation claims for English, Arabic, Hindi, Russian, and Spanish are chosen from fact checking websites in that language, with an exception of Russian, where misinformation claims are translated from English. This is due to Russian control of their internet, where any non-biased Russian language fact-checking site would be blocked. We select from a diverse range of websites to mitigate any potential bias, and generally speaking, claims are collected as randomly as possible.



Войска НАТО прямо сейчас на Украине!

NATO forces are in Ukraine as we speak!

Нет, таково не может быть

No, this is impossible

В Украине Есть только те войска НАТО, которые обучают Украинцов. Их мало.

The only NATO forces in Ukraine are the ones that are teaching Ukrainian soldiers. But there are little.

Как вы это знаете?

How do you know this?

Figure 1: Example of a data point (tweet and context) in the Russian Stanceosaurus dataset. For the claim "NATO forces are currently fighting in Ukraine", we have an example tweet chain demonstrating various stances.

Tweet Collection: For better coverage of diverse topics, we invested substantial effort in creating customized queries with varied keywords and time ranges for each claim to retrieve tweets. We also trace the entire reply chain in both directions, so Stanceosaurus includes relevant tweets that may not contain keywords from claims. A reply chain in both direction allows our annotators to make more informed decisions when annotating the data. Data was collected before the name change from Twitter to X and the implementation of stricter Twitter API policies.

Annotation: We define our 5-way stance annotation as follows:

[*] Equal contribution.

- **Irrelevant**: unrelated to the claim
- **Supporting**: explicitly affirms the claim is true
- **Refuting**: explicitly asserts the claim is false
- **Querying**: questions the veracity of the claim
- **Discussing**: provide neutral information on the context or veracity of the claim.

Then, we can adapted this schema to a 3-way stance classification with *supporting*, *refuting*, and *other* as labels. This is done by additionally annotating leanings for the Discussing category. Then, with those leanings, we get the following 3-way schema:

- **Supporting:** Supporting, Discussing$_{sup}$;
- **Refuting:** Refuting, Discussing$_{ref}$;
- **Other:** Irrelevant, Querying, Discussing$_{other}$.

This allows for additional flexibility in the use of Stanceosaurus. In terms of annotators, we hired four native speakers for English, two for Hindi, two for Arabic, two for Russian, and one for Spanish (second one on the way). For the languages with more than one annotator, the annotator agreements were all within acceptable range.

3 Experimental Results

Throughout these experiments, we use cross-entropy loss, weighted cross-entropy loss, and class-balanced focal loss (Baheti et al. 2021, Cui et al. 2019), which down-weights easy examples and focuses more on difficult ones.

Stance Detection for Unseen Claims We find that training BERTweet$_{large}$ on the train set of Stanceosaurus English and using the aforementioned class-balanced focal loss yields a macro F1 of about 61 on unseen claims. If we break this down into the 3-way categorization, we reach a macro F1 of 68 using the same set-up. Furthermore, in terms of classwise results, we find that for the classes of supporting, refuting, discussing, querying, and irrelevant, we record classwise F1 scores of 60.6, 61.1, 64.9, 45.8, and 74.1 respectively.

Zero Shot Cross-Lingual Transfer Truly multicultural stance identification requires models that are capable of operating across languages. This is especially important, since multiple sources have shown that misinformation in non-English languages is rampantly spreading on the internet. To demonstrate the feasibility of identifying the stance towards misinformation claims in a zero-shot cross-lingual setting, when no training data in the target language is available, we fine-tune models on Stanceosaurus' English training set and use all the annotated Hindi/Arabic/Russian/Spanish data as the test set. We achieve 53.1 F1 on Hindi, 50.4 F1 on Arabic, 43.9 F1 on Russian, and 43.8 F1 on Spanish without any target language fine-tuning. Note that the Russian and Spanish numbers are lower as they were only tested on mBERT, while Hindi and Arabic's best results came from XLM-R$_{large}$.

Combining Stanceosaurus + RumourEval We use EasyAdapt (Daumé III 2007, Bai et al. 2021) to fine-tune BERTweet$_{large}$ on the combination of RumourEval and Stanceosaurus. RumourEval (Gorrell et al., 2019) is the main "competing" stance-based misinformation dataset. BERTweet$_{large}$ with EasyAdapt achieves 67.4 Macro F1 for Stanceosaurus and 65.8 Macro F1 for RumourEval, outperforming the in-domain model performance for Stanceosaurus and matching the in-domain model performance of RumourEval.

Stance Detection for Unseen Countries The English dataset comprises 97 international and 93 regional claims. We test BERTweet's ability to generalize toward regional claims by training on international claims. Performance on the regional data varies widely between sources. Poynter and AFP Fact Check New Zealand, two sources with the most international data, have the best F1s at 63.0 and 63.5 respectively.

4 Conclusion

We introduce Stanceosaurus, a new corpus of 31,904 social media messages annotated with stance towards 292 misinformation claims originating from 21 multicultural fact-checking sources. To the best of our knowledge, Stanceosaurus is the largest stance dataset yet. Stanceosaurus contains consistent annotations across claims and languages, and stance classifier models trained on our dataset can perform well on unseen claims and languages. Our work represents a step towards the development of accurate models that can track the spread of misinformation online across diverse languages and cultures.

References

Ashutosh Baheti, Maarten Sap, Alan Ritter, and Mark Riedl. 2021. Just say no: Analyzing the stance of neural dialogue generation in offensive contexts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4846–4862, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fan Bai, Alan Ritter, and Wei Xu. 2021. Pre-train or annotate? domain adaptation with a constrained budget. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5002–5015, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic. Association for Computational Linguistics.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, Minneapolis, Minnesota, USA. Association for Computational Linguistics.