Stanceosaurus 2.0: Classifying Stance Towards Russian and Spanish Misinformation

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

Abstract

The Stanceosaurus corpus (Zheng et al., 2022) was designed to provide high-quality, annotated, 5-way stance data extracted from Twitter, suitable for analyzing cross-cultural and cross-lingual misinformation. In the Stanceosaurus 2.0 iteration, we extend this framework to encompass Russian and Spanish. The former is of current significance due to prevalent misinformation amid escalating tensions with the West and the violent incursion into Ukraine. The latter, meanwhile, represents an enormous community that has been largely overlooked on major social media platforms. By incorporating an additional 3,874 Spanish and Russian tweets over 41 misinformation claims, our objective is to support research focused on these issues. To demonstrate the value of this data, we employed zero-shot cross-lingual transfer on multilingual BERT, yielding results on par with the initial Stanceosaurus study with a macro F1 score of 43 for both languages. This underlines the viability of stance classification as an effective tool for identifying multicultural misinformation.

1 Introduction

Misinformation on social media is a highly multicultural phenomenon (Roozenbeek et al., 2020). In the ongoing Russia-Ukraine conflict, Russian-language misinformation and propaganda are important weapons used by both sides to influence the opinions of Internet users across the globe. Meanwhile, Spanish-language misinformation is surging unchecked through virtually every online community.[1] With these issues in mind, we seek to create a dataset that can help identify Spanish and Russian misinformation beyond a binary yes/no approach. We do this by expanding the Stanceosaurus dataset (Zheng et al., 2022) to include Spanish and Russian tweets annotated using a 5-way stance labeling schema (Gorrell et al. 2018, Schiller et al. 2021), thus creating *Stanceosaurus 2.0*. By fine-tuning multilingual BERT (Devlin et al., 2019), we experiment with zero-shot cross-lingual transfer, demonstrating the potential for *Stanceosaurus 2.0* to help drive forward misinformation research on Spanish and Russian. Furthermore, recent Twitter policies have made it clear that the site is moving away from account-based labeling of misinformation.[2] Our dataset presents the opportunity to identify potential misinformation on a per-tweet basis, allowing users to see relevant context for potentially misleading tweets. Some may argue that in recent times, Twitter (now X at the time of revision) has taken a far more "hands-off" approach to misinformation. While this may or may not be true, this dataset can be used on social media platforms that are different from Twitter/X. One can get around the tweet length limit by simply concatenating various tweets, etc. In the following sections, we discuss what exactly Russian and Spanish misinformation entail and why they are so important.

Russian Misinformation Misinformation and propaganda are crucial to Russian political warfare. Part of so-called “active measures”, they are designed to “weaken the West [and] to drive wedges in the Western community alliances of all sorts, particularly NATO ...” (Alexander, 2017). When Russia launched a full-scale invasion of Ukraine in February of 2022,[3] both sides of the conflict engaged in hybrid warfare, putting an equal focus on the information front and global deception.[4] With propaganda machines in full force, the war in Ukraine has spawned many new misinformation claims. In this context, although a Russian stance dataset is present in Lozhnikov et al. (2018), it is limited, and our research aims to modernize

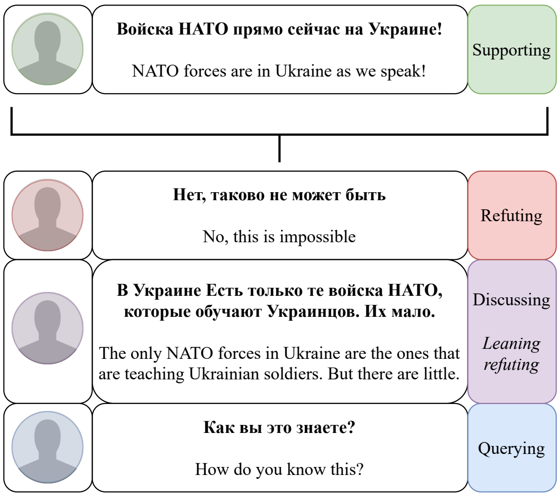
[1] The Guardian

[2] Twitter

[3] CNBC

[4] The Atlantic

Proceedings of the Ninth Workshop on Noisy and User-generated Text (W-NUT 2024), pages 31–43
March 22, 2024 ©2024 Association for Computational Linguistics



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.

Russian stance data to include wartime misinformation. This is because Russian misinformation "then" and "now" are two different beasts. Potential feasibility for the idea of Russian Stanceosaurus can be seen via the findings in Park et al. (2022) which, among many interesting things, identified Twitter as a platform with a significant amount of Russian-language discussion regarding wartime events. The series of Solopova et al. (2023a) and Solopova et al. (2023b), which showcases great results on pro-Kremlin propaganda detection, also potentially implies feasibility of our stance-based approach.

Spanish Misinformation Misinformation is rampant in the Spanish-speaking world, surging through various online communities (Bonnevie et al., 2023). Despite being the fourth most spoken language in the world and an enormous medium for the spread of information worldwide both true and false, misinformation in Spanish is far more of a problem than in English[5]. This problem is further exacerbated when it is ignored; Facebook whistleblower Frances Haugen revealed an enormous disconnect between the proportion of users who speak Spanish and the amount of spending committed to anti-misinformation resources in this language[6]. Unsurprisingly, works such as Posadas-Durán et al. (2019) and Abonizio et al. (2020) attempt to help solve this crucial problem, particularly framing the problem as detecting fake news. These studies show that this kind of claim-based misinformation detection works quite well. Our approach was inspired by such studies. To our knowledge, there are two existing Spanish stance datasets. One is Zotova et al. (2020), a valuable but singular claim-limited collection of Spanish-language stance data. The other is Toledo-Ronen et al. (2020), which creates a wonderful Spanish stance dataset, but based on arguments and not misinformation claims. We aim to expand the set of Spanish misinformation via the five-way and three-way classification framework of Stanceosaurus.

2 Stanceosaurus 2.0: Details

In order to facilitate the study of Russian and Spanish misinformation, we have created a 5-way stance classification dataset in accordance with the guidelines established by Zheng et al. (2022). These stance categories are `Irrelevant`, `Supporting`, `Refuting`, `Querying`, and `Discussing`. The stance of a tweet is based on the misinformation claim it is discussing. An example of various misinformation claim related tweets and their stance categorization can be seen in Figure 1. Details on the five stance categories (and how they can be merged to 3 categories) are listed in appendix C.

2.1 Data Collection

Misinformation Claims We derived 18 examples of Russian-language misinformation, with 13 from the European Union initiative, *euvsdisinfo*, and manually translated them into Russian using a bilingual Russian/English speaker (both fluent). Despite criticism of fact-checking methodology (Giorio, 2018), euvsdisinfo is to our knowledge the best source of prominent misinformation which can be found on Russian-language Twitter, especially considering that there is no reliable Russian fact-checking website (this is the reason why we had to translate the claims to Russian). Nonetheless, to mitigate this bias, we supplemented these misinformation claims with 5 claims from the Western media. Again, the absence of claims from Russian sources or Russian-language fact-checking sites is notable. We re-iterate that identifying misinformation is challenging when Russian media is largely state controlled[7]. Any source that does fact-checking that might disagree with the state media would most likely get taken down or blocked by Roskomnadzor. Therefore,

[5] Washington Post

[6] The Guardian

[7] ForeignPolicy.com

for this study, we selected a ground truth based on western-leaning sources to assess Russian misinformation claims. For the Spanish corpus, we collected 23 misinformation claims from reputable Spanish-language fact-checking websites *Verificado*, *Chequeado*, *Newtral*, and *ChequeaBolivia*, including claims with various veracity ratings. The selection of both Spanish and Russian claims was guided by the volume of relevant Twitter discourse. The detailed Russian and Spanish claims are listed in Appendix A and B, respectively.

Tweet Collection & Reply Chains For both languages, we collect tweets using the Twitter API. Queries (Appendix A and B) are manually curated and iteratively refined in order to capture as many relevant tweets as possible while allowing for diversity in stance categories. This refinement was done by scraping data for a claim using a certain query, and then sampling 20 tweets and hand annotating them. Queries were then modified, and the process was repeated until a reasonably equal distribution of stances was achieved. Alongside tweets from queries, we also collect additional tweets for context, including those "above"(preceding) and "below"(following) in the reply chain. Similar to the original Stanceosaurus paper, context tweets were included to potentially help models make classification decisions. Annotation was done by sampling 50 tweets for each claim, adding up to 100 context tweets, which were then evaluated. These quantities were chosen based on the number of available annotators. All three annotators are college-educated students who are fluent in their respective languages. There were two annotators for Russian and one for Spanish. Future versions of this work will include another annotator for Spanish. The 5-class Cohen Kappa for the Russian data is 77.4. Reproducibility criteria for this process are discussed in Appendix D. A basic overview of corpus statistics is in Table 1. More detailed corpus statistics can be found in F. As mentioned in D, the dataset can be requested directly from the authors. Since the tweet ID's it contains can link tweets back to their authors, it is important that these tweets are used only for academic purposes, especially given that some of them present controversial political opinions.

2.2 Russian Corpus

Russian Twitter According to a Statista study (Statista, 2023a), only 5% of Russians surveyed reported using Twitter. This makes sense, as before this study, "Facebook, Instagram, and Twitter were all blocked by the Russian state in early March 2022 when the laws on antiwar activity came into force" (McCarthy et al., 2023). Thus, the only way to access Twitter in Russia is through a VPN. From this information, we gather that Russian speakers on Twitter either use a VPN, are native to another country where Russian is a common language, or live abroad. Acknowledging this is important, as it provides a population for the Twitter users that were sampled and presents a limitation of the data to be addressed with future work.

#tweets	**Refute**	**Support**	**Irrel.**	**Query**	**Discuss**	**Total**
Russian	119	332	999	50	407	1907
Spanish	270	302	1036	16	342	1966

Table 1: Corpus statistics of Stanceosaurus 2.0.

Code Switching There are instances of tweets that contain different languages. It is fairly common to see acronyms such as "HIMARS" and "NATO" being written in both English and Russian interchangeably. A brief analysis using regular expressions finds that tweets containing characters from the English alphabet make up about 12 percent of all tweets. Furthermore, sometimes the Russian language is phonetically written in the Latin alphabet. This poses a challenge when querying for relevant Tweets, which we addressed by accounting for as many code-switched variations as was reasonable. Furthermore, in the sampled reply chains, it was common to see the mixing of languages, especially between Russian and Ukrainian. Fortunately, since every single tweet was hand annotated by a fluent Russian speaker with proficient knowledge of Ukrainian, it was not difficult to differentiate between the two languages, as well as any other language that uses the Cyrillic alphabet.

Obscenities Due to the politically charged nature of the Russian misinformation claims, many tweets contain large amounts of cursing, which is known as *mat* (pronounced maht). It is argued that *mat* "is not merely an accumulation of obscenities, but rather constitutes a set of refined, complex structures", hinting at a "potentially limitless quantity of expressions" (Dreizin and Priestly, 1982). A cursory analysis indicates that around 10 percent of all Russian tweets collected contain some sort of obscenity. Context is key when trying to understand Russian obscenities, and this may prove to be quite confusing for a language model to interpret.

2.3 Spanish Corpus

Circumventing Filters Particularly when discussing the COVID-19 vaccine, many tweets include language that is most likely obscured to circumvent misinformation filters. When searching "vacuna" (Spanish for "vaccine"), Twitter sends users a warning to check federal websites for information related to the pandemic.[8] Accordingly, the queries had to be adjusted to include numerous alternative spellings for the word for vaccine (including 'vacuno', 'vacun@', 'vakuna', 'cacunados', 'v@cunad0s', 'kakuna', etc.).

Social Media Usage The decision to utilize Twitter for this corpus was driven by its accessible API and publicly shareable text-centric content for open and ethical NLP research. It is worth noting that within the Spanish-speaking realm, Twitter ranks behind Facebook, Instagram, and TikTok in terms of social media usage (Statista 2023b, StatCounter 2023). Additionally, more Hispanics use WhatsApp than any other race or ethnicity,[9] and significant volumes of misinformation spread on private channels such as WhatsApp[10] where misinformation detection is much more difficult and misinformation is less likely to be corrected by the public.

Code Switching Mixing Spanish and English together in a single tweet is common, particularly in Spanish-speaking communities in Northern Mexico and the USA. The spread of misinformation in bilingual communities is a unique challenge of particular importance in the United States, where more than one-third of all Hispanic adults self-identify as bilingual in English and Spanish (Pew, 2015).

3 Automatic Stance Detection Using Stanceosaurus 2.0

Zero-Shot Cross-Lingual Transfer In accordance with the original Stanceosaurus paper (Zheng et al., 2022), we conduct a zero-shot cross-lingual transfer experiment on our data. This entails training a model on the English Stanceosaurus dataset of 20,707 tweets and then evaluating it on the Russian and Spanish sets. We believe that this is the best way to evaluate Stanceosaurus 2.0 since we assume that there is little to no stance-based training data available for Russian and Spanish (something we observed during our research, and can be seen in Section 1 where we discuss related work). Also, various studies such as Pires et al. (2019) and Artetxe et al. (2020) have shown zero-shot cross-lingual transfer to be an effective approach in many languages, including Russian and Spanish.

Multilingual BERT Multilingual BERT (Devlin et al., 2019), or mBERT, has been shown to be very competitive in the zero-shot setting that we have described (Wu and Dredze 2019, Libovický et al. 2019). We believe that mBERT is a simple baseline that indicates the quality of our dataset and model performance. For our experiments, we follow the original Stanceosaurus paper (Zheng et al., 2022) and use the five stance label schema. To create model input, we format our strings using special tokens as follows: "[CLS] claim [SEP] text".

Loss Functions Similar to the original Stanceosaurus (Zheng et al., 2022), we examine three different loss functions: cross-entropy loss, weighted cross-entropy loss (Cui et al., 2019), and class-balanced focal loss (Baheti et al., 2021). While the cross-entropy loss is a baseline commonly used in classification tasks, we use weighted cross-entropy to modify this baseline to account for imbalanced classes by assigning more weights to classes with fewer samples. Class-balanced focal loss is an alternative method to account for imbalanced classes. It down-weights easy examples and focuses more on difficult ones (Cui et al., 2019).

Results The results of our experiments can be seen in Table 2. One can compare these results to English performance on BERT$_{BASE}$ for unseen claims from the original Stanceosaurus paper (Zheng et al., 2022), as well as the same zero-shot cross-lingual transfer experiment on Hindi and Arabic. These extra experiments are also shown in 2, but they are clearly marked as the contribution of the authors of the original Stanceosaurus paper. Both Russian and Spanish datasets performed similarly to models for English to Hindi and English to Arabic transfer experiments in the original Stanceosaurus (Zheng et al., 2022). The weighted loss functions performed better overall, and both languages achieved an F1 score of around 43. Reproducibility criteria for our experiments can be seen in appendix E.

[8] This is no longer the case following recent changes to Twitter policy.

[9] Insider Intelligence

[10] Harvard Kennedy School

Russian (our contribution)

Loss	Precision	Recall	F1
CE	53.55±0.8	35.33±0.7	36.15±1.3
Weighted CE	44.38±0.2	42.84±0.5	42.09±0.1
CBFL	45.60±1.5	46.98±2.0	43.94±0.2

Spanish (our contribution)

Loss	Precision	Recall	F1
CE	50.26±1.9	40.86±0.7	41.81±1.0
Weighted CE	54.12±0.4	42.65±0.5	43.75±0.4
CBFL	51.26±2.2	44.15±0.9	43.83±1.0

Hindi (Zheng et al., 2022)

Loss	Precision	Recall	F1
CE	52.1±2.9	39.4±2.0	40.8±2.5
Weighted CE	55.0±4.2	42.4±1.4	44.3±1.8
CBFL	53.0±3.4	44.1±1.7	45.3±1.5

Arabic (Zheng et al., 2022)

Loss	Precision	Recall	F1
CE	44.8±4.0	40.1±2.5	40.0±2.0
Weighted CE	44.1±3.3	40.7±1.6	39.7±1.7
CBFL	46.1±2.6	44.7±1.1	43.1±0.2

English on BERT$_{BASE}$ (Zheng et al., 2022)

Loss	Precision	Recall	F1
CE	51.1±1.1	50.5±2.0	50.4±1.6
Weighted CE	50.5±1.9	52.7±1.1	51.3±1.3
CBFL	50.6±1.3	55.7±2.1	52.5±1.0

Table 2: Russian and Spanish experiments. Models are trained on English Stanceosaurus and then evaluated on either Russian or Spanish in our work. F1 is measured as macro F1. Results are taken as the average of 3 experiments, with error being one standard deviation. English, Arabic, and Hindi experiments are taken directly from Stanceosaurus (Zheng et al., 2022) as a comparison benchmark.

4 Conclusion

We introduce Stanceosaurus 2.0, an extension of the 5-way stance dataset Stanceosaurus (Zheng et al., 2022). Our dataset includes 18 Russian misinformation claims (1907 tweets) and 23 Spanish misinformation claims (1966 tweets). Our dataset is modern and up to date given the recent slough of misinformation and current events. It also contains Russian and Spanish, which as shown previously, are two languages in which misinformation thrives, and efforts to combat it are limited. Our zero-shot cross-lingual transfer experiments show that our dataset performs at similar levels to that of Hindi and Arabic in the original Stanceosaurus, with a macro F1 score of about 43. This means that there is potential to continue refining models and algorithms to create a somewhat reliable stance classifier using transformer-based models like mBERT. Future versions of this work will entail experiments on more models, as well as a second annotator for the Spanish version.

Limitations

The Veracity of Fact-Checked Claims One of the biggest limitations of our work is the fact that fact-checking is often not as black-and-white as it seems and is generally a practice that suffers from many limitations (Uscinski and Butler, 2013). It is very difficult to find objective truths that are verified to a degree of absolute precision for a work like this. This is doubly so for political-leaning claims, such as the claims in the Russian dataset.

Russian Misinformation Claims An unfortunate limitation of the Russian language is that there are no Russian fact-checking websites that would provide reasonably objective fact-checking, at least as far as we are aware. This is most likely due to the level of control that the Russian government has over the Russian internet (Polyakova and Meserole, 2019). This lack of resources means that Russian claims were hand-picked. This could introduce author bias, and may not be an accurate representation of the Russian internet, as claims were mostly all found on the heavily western-leaning website euvsdisinfo, as discussed in section 2.

Russian Twitter As mentioned in section 2.2, Twitter is not the most used social media, and this could introduce various biases into our data. Future work could involve the social media website VKontakte, which as mentioned earlier, is the most popular in Russia. However, some problems could arise due to state-owned entities being shareholders[11].

Spanish Twitter Likewise, Twitter is far from the most popular social media network in Latin America. More work should be done to analyze misinformation on Facebook and WhatsApp in the Hispanosphere. Despite favoring small-group communication, WhatsApp persists as a medium for rapid misinformation dissemination in Latin America (Nobre et al., 2022).

Spanish Queries As mentioned in section 2.3, numerous obstacles made it difficult to query for

[11] Reuters

relevant Tweets in Spanish. From properties inherent to the Spanish language like a highly inflectional morphology to broader social factors including the prevalence of code-mixing and filter circumvention, care had to be taken when querying Twitter's API to find relevant Tweets without biasing the data in any one direction (Pfaff, 1979). Future work might include broad queries to procure larger datasets that can then be manually cleaned to include more relevant Tweets.

Code Switching As mentioned in both sections 2.2 and 2.3, both languages experienced a decent amount of code-switching, whether it be in the context or the tweet itself. It has been shown before that dealing with code-switching is not an easy task (Winata et al., 2021). However, recently there has been a large number of code-switching datasets that have become available (Jose et al., 2020). Potential further research may include creating stance datasets exclusively on code-switched datasets.

Tweet Deletion A feature of the obscured version of the dataset (the version we plan on giving out in most cases) is that it only features tweet IDs. However, if someone deletes a tweet, that tweet will be gone from the obscured dataset. This maintains the user's right to remove their content without it still being a database. However, this may be an issue for researchers using this dataset a long time after the tweets were originally collected.

Ethics Statement

Working With Social Media Data Mining social media data from Twitter users without their consent is at best ethically problematic (Taylor and Pagliari, 2018). Unfortunately, this kind of data would not exist without this technique. However, our publicly available dataset only contains tweet IDs and does not include actual tweets and usernames. Furthermore, social media data can contain harmful biases towards certain groups, as moderating social media can be extremely difficult (Ganesh and Bright, 2020). We encourage a thorough review of the data and its context before deploying in a production environment.

Data Annotation We recognize that some of the tweets that have been annotated deal with sensitive topics and contain some hateful language, especially in the Russian dataset, given its political nature. We recognize that annotators need to be warned of this before they start annotating.

Propaganda Analysis An issue with analyzing propaganda and misinformation is that this analysis can potentially fall into the wrong hands. For example, using this dataset to analyze the effectiveness of Russian propaganda can inform the source of the propaganda exactly what they could improve on.

Acknowledgments

The authors would like to thank Dennis Pozhidaev for his help with data annotation and evaluation. This research is supported by the NSF (IIS-2052498) and IARPA via the HIATUS program (2022-22072200004). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of NSF, ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Hugo Queiroz Abonizio, Janaina Ignacio de Morais, Gabriel Marques Tavares, and Sylvio Barbon Junior. 2020. Language-independent fake news detection: English, portuguese, and spanish mutual features. *Future Internet*, 12(5).

Keith B. Alexander. 2017. Disinformation: A primer in russian active measures and influence campaigns. Prepared statement, United States Senate Select Committee on Intelligence.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

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.

E Bonnevie, V Ricciulli, M Fields, and R O'Neill. 2023. Lessons learned from monitoring spanish-language vaccine misinformation during the covid-19 pandemic. *Public Health Rep*, 138(4):586–592. PMID: 37102367; PMCID: PMC10140774.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of* *the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

F. Dreizin and T. Priestly. 1982. A systematic approach to russian obscene language. *Russ Linguist*, 6:233–249.

Bharath Ganesh and Jonathan Bright. 2020. Countering extremists on social media: Challenges for strategic communication and content moderation.

L. Giorio. 2018. *War on Propaganda or PRopaganda War?: A case study of fact-checking and (counter)propaganda in the EEAS project EUvsDisinfo*. Dissertation, Uppsala University, Jagiellonian University.

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

Navya Jose, Bharathi Raja Chakravarthi, Shardul Suryawanshi, Elizabeth Sherly, and John P. McCrae. 2020. A survey of current datasets for code-switching research. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 136–141.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2019. How language-neutral is multilingual bert?

Nikita Lozhnikov, Leon Derczynski, and Manuel Mazzara. 2018. Stance prediction for russian: Data and analysis.

Lauren A. McCarthy et al. 2023. Four months of "discrediting the military": Repressive law in wartime russia. *Demokratizatsiya: The Journal of Post-Soviet Democratization*.

Gabriel Peres Nobre, Carlos H.G. Ferreira, and Jussara M. Almeida. 2022. A hierarchical network-oriented analysis of user participation in misinformation spread on whatsapp. *Information Processing & Management*, 59(1):102757.

Chan Young Park, Julia Mendelsohn, Anjalie Field, and Yulia Tsvetkov. 2022. Challenges and opportunities in information manipulation detection: An examination of wartime russian media.

Pew. 2015. A majority of english-speaking hispanics in the u.s. are bilingual. Accessed: 2023-06-11.

Carol W. Pfaff. 1979. Constraints on language mixing: Intrasentential code-switching and borrowing in spanish/english. *Language*, 55(2):291–318.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert?

Alina Polyakova and Chris Meserole. 2019. Exporting digital authoritarianism: The russian and chinese models. *Policy Brief, Democracy and Disorder Series*, pages 1–22.

Juan-Pablo Posadas-Durán et al. 2019. Detection of fake news in a new corpus for the spanish language. *Journal of Intelligent & Fuzzy Systems*, 36(5):4869–4876.

Jon Roozenbeek, Claudia R. Schneider, Sarah Dryhurst, John Kerr, Alexandra L. J. Freeman, Gabriel Recchia, Anne Marthe van der Bles, and Sander van der Linden. 2020. Susceptibility to misinformation about covid-19 around the world. *R. Soc. open sci.*, 7:201199.

B. Schiller, J. Daxenberger, and I. Gurevych. 2021. Stance detection benchmark: How robust is your stance detection? *Künstl Intell*, 35:329–341.

V. Solopova, OI. Popescu, C. Benzmüller, et al. 2023a. Automated multilingual detection of pro-kremlin propaganda in newspapers and telegram posts. *Datenbank Spektrum*, 23:5–14.

Veronika Solopova, Christoph Benzmüller, and Tim Landgraf. 2023b. The evolution of pro-kremlin propaganda from a machine learning and linguistics perspective. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 40–48, Dubrovnik, Croatia. Association for Computational Linguistics.

StatCounter. 2023. Social media stats spain. Accessed: June 11, 2023.

Statista. 2023a. Ranking of social media platforms in russia q3 2022, by user share. Accessed: 2023-05-26.

Statista. 2023b. Social media usage in latin america - statistics & facts. Accessed: June 11, 2023.

Joanna Taylor and Claudia Pagliari. 2018. Mining social media data: How are research sponsors and researchers addressing the ethical challenges? *Research Ethics*, 14(2):1–39.

Orith Toledo-Ronen, Matan Orbach, Yonatan Bilu, Artem Spector, and Noam Slonim. 2020. Multilingual argument mining: Datasets and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 303–317, Online. Association for Computational Linguistics.

Joseph E. Uscinski and Ryden W. Butler. 2013. The epistemology of fact checking. *Critical Review*, 25(2):162–180.

Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2021. Are multilingual models effective in code-switching?

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert.

Jonathan Zheng, Ashutosh Baheti, Tarek Naous, Wei Xu, and Alan Ritter. 2022. Stanceosaurus: Classifying stance towards multilingual misinformation.

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.

A Russian Claims and Queries

Russian claims and queries can be found in figure 3.

B Spanish Claims and Queries

Spanish claims and queries can be found in figures 4 and 5.

C Stance Categorization

The following is a description of each stance:

- **Supporting:** Tweets that directly support the fact that a claim is true.
- **Refuting:** Tweets that refute the veracity of a claim.
- **Querying:** Questions the veracity of a claim.
- **Discussing:** Provides neutral information on the context or truth of a claim.
- **Irrelevant:** Not relevant to the given claim.

If a tweet is labeled as discussing, then to enable 3-way stance classification, the tweet is also given a leaning. The following is a description of each leaning:

- **Supporting:** The tweet has an indirect positive bias when discussing the claim.
- **Refuting:** The tweet has an indirect negative bias when discussing the claim.
- **Other:** The tweet does not have any sort of bias.

With this information, we can now construct our guidelines for 3-way stance categorization as well:

- **Supporting:** Merge supporting with discussing$_{supporting}$.
- **Refuting:** Merge refuting with discussing$_{refuting}$.
- **Other:** Merge irrelevant, querying, and discussing$_{other}$.

D Dataset Reproducibility Criteria

- Using the twitter API, up to 150 tweets were pulled for each claim using the queries listed in figures 3, 4, and 5. Context for each tweet was also retrieved. Context in this case means the entire reply chain from the root tweet down to the pulled tweet, as well as any immediate replies.
- Quality control was done by an extensive iteration of Twitter API queries. We aimed to make queries such that the distribution of stance categories was reasonably even, although this proved to be difficult with the "Querying" category.
- With these tweets in hand, up to 50 tweets were sampled for each claim for annotation. Context tweets were also annotated. Up to 50 parent context tweets were sampled and up to 50 context children tweets were sampled for each claim.
- Claims were annotated in accordance with details given in appendix C. Russian tweets were double annotated, while Spanish tweets currently only have a single annotator, but we are working to find another annotator at the moment.
- Tweets were pre-processed to remove duplicates using lexical similarity.
- The context chains were then reconstructed and formatted in json to match the original Stanceosaurus paper (Zheng et al., 2022).
- The dataset can be requested from the authors using the emails given in the paper. Since the data is potentially sensitive (tweets of political nature) we need to make sure that anyone who uses these tweets is doing so solely out of academic intent.

E Experiment Reproducibility Criteria

- **Model:** bert-base-multilingual-uncased
- **Computing Infrastructure:** 4 Nvidia Titan X GPUs. NVIDIA-SMI 460.84. Driver Version 460.84. CUDA version 11.2. Running on CentOS linux 7. Conda version 7. Package versions listed in `requirements.txt` file in code used.
- **Average Training Time:** Per experiment, around 40 minutes
- **Evaluation Metrics:** Best evaluation of the development set per training run
- **Number of Experiments:** Each row in 2 was

done 3 times. Results are the mean ± the standard deviation. Random seeds for the three runs were 10, 20, and 30.

- **Hyperparameters:** Hyperparameters were chosen based off of best performing hyper parameters in the original Stanceosaurus model, and then manually tuned.
  - **Learning Rate:** 3e-5
  - **Batch Size:** 8 per GPU, so 32 total
  - **Class Balanced Focal Loss:** Similar to the original paper, we tune β and γ between $[0.1, 1)$ and $[0.1, 1.1]$ respectively.
  - The rest are defaulted to what is used in the code. Run commands are included with code.
- Code zip file can be accessed upon request.

F Corpus Statistics

The distribution of labels and tweet types for Russian Spanish are shown in tables 3 and 4 respectively. A visual representation of the tweets (not context or replies) for Russian Spanish is shown in figures 2(a) and 2(b) respectively.

G Annotation Logistics

Annotators were American college students paid 18 dollars an hour. Each annotator was fluent in the language they were annotating. All annotators were recruited as people the authors directly knew. Verbally, annotators were told the scope of the paper and given the abstract.

H Use of AI assistants

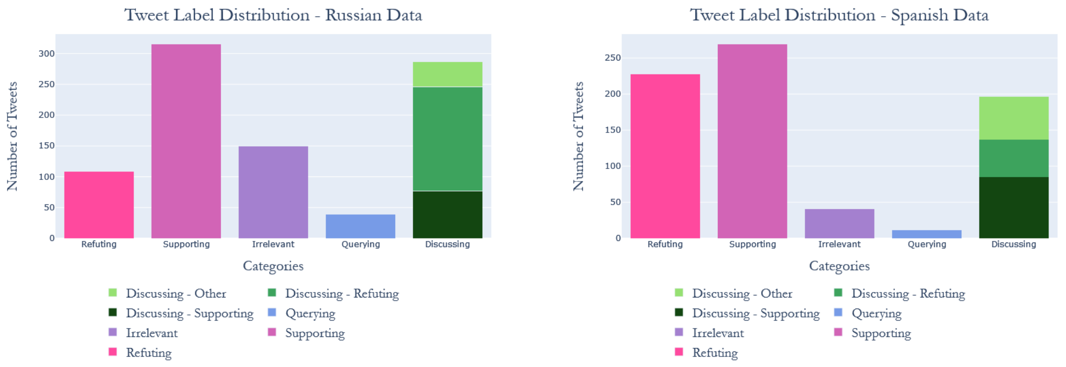
AI assistants were used by the authors of this paper in order to proofread the paper. Occasionally, an AI assistant was asked to rephrase some text, just to generate some ideas on sentence flow. Work was never directly copied, and model output was used as inspiration.

	Refuting	Supporting	Irrelevant	Querying	$\mathrm{D}_{supporting}$	$\mathrm{D}_{refuting}$	D_{other}	Total
Tweets	109	315	149	39	77	169	41	899
Context	5	15	738	9	51	40	7	865
Replies	5	2	112	2	6	15	1	143
Total	119	332	999	50	134	224	49	1907

Table 3: Russian Corpus Statistics.

	Refuting	Supporting	Irrelevant	Querying	$\mathrm{D}_{supporting}$	$\mathrm{D}_{refuting}$	D_{other}	Total
Tweets	228	269	418	12	85	52	60	1124
Context	15	21	370	2	18	13	4	443
Replies	27	12	248	2	76	18	16	399
Total	270	302	1036	16	179	83	80	1966

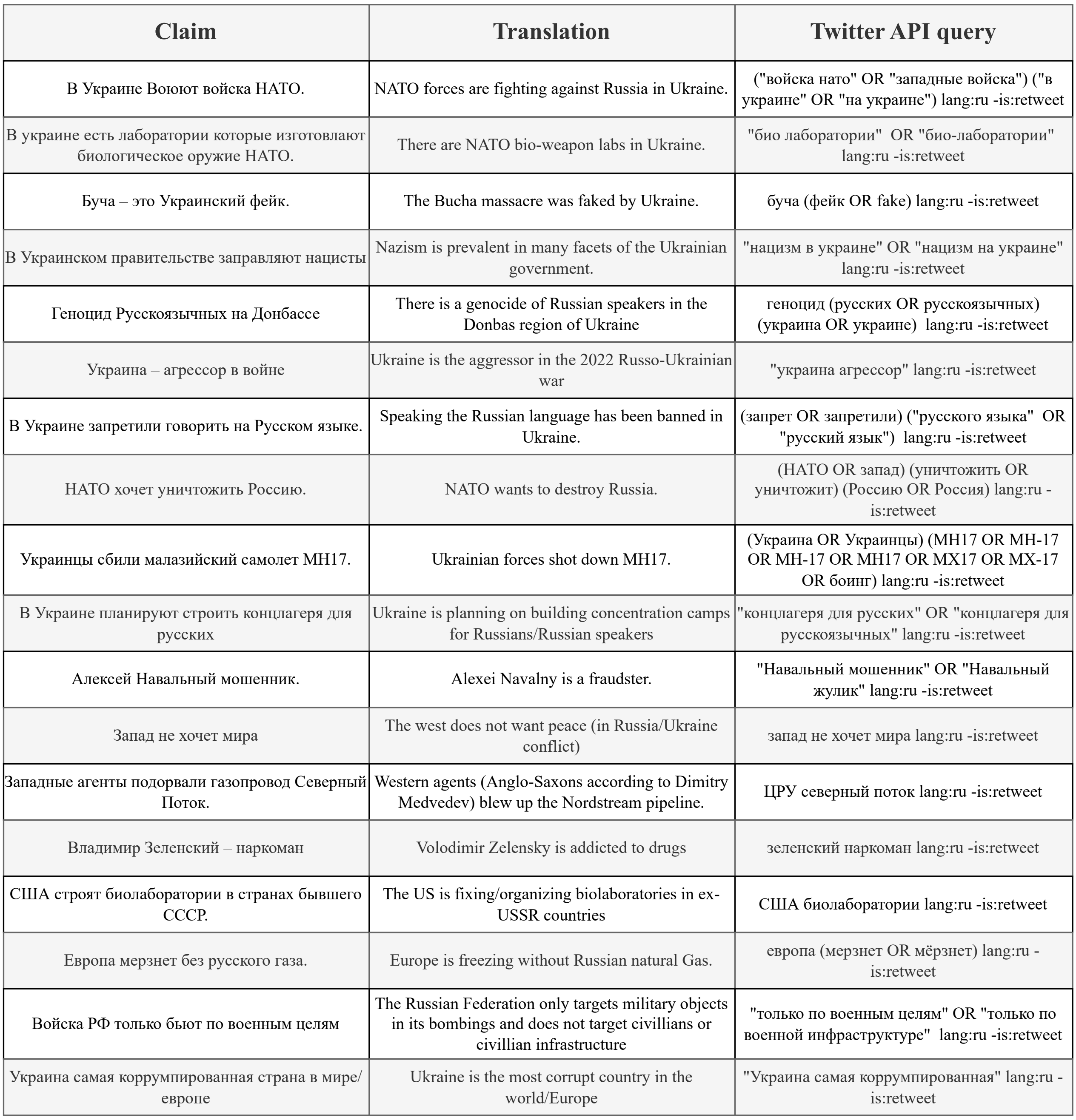
Table 4: Spanish Corpus Statistics.



(a)

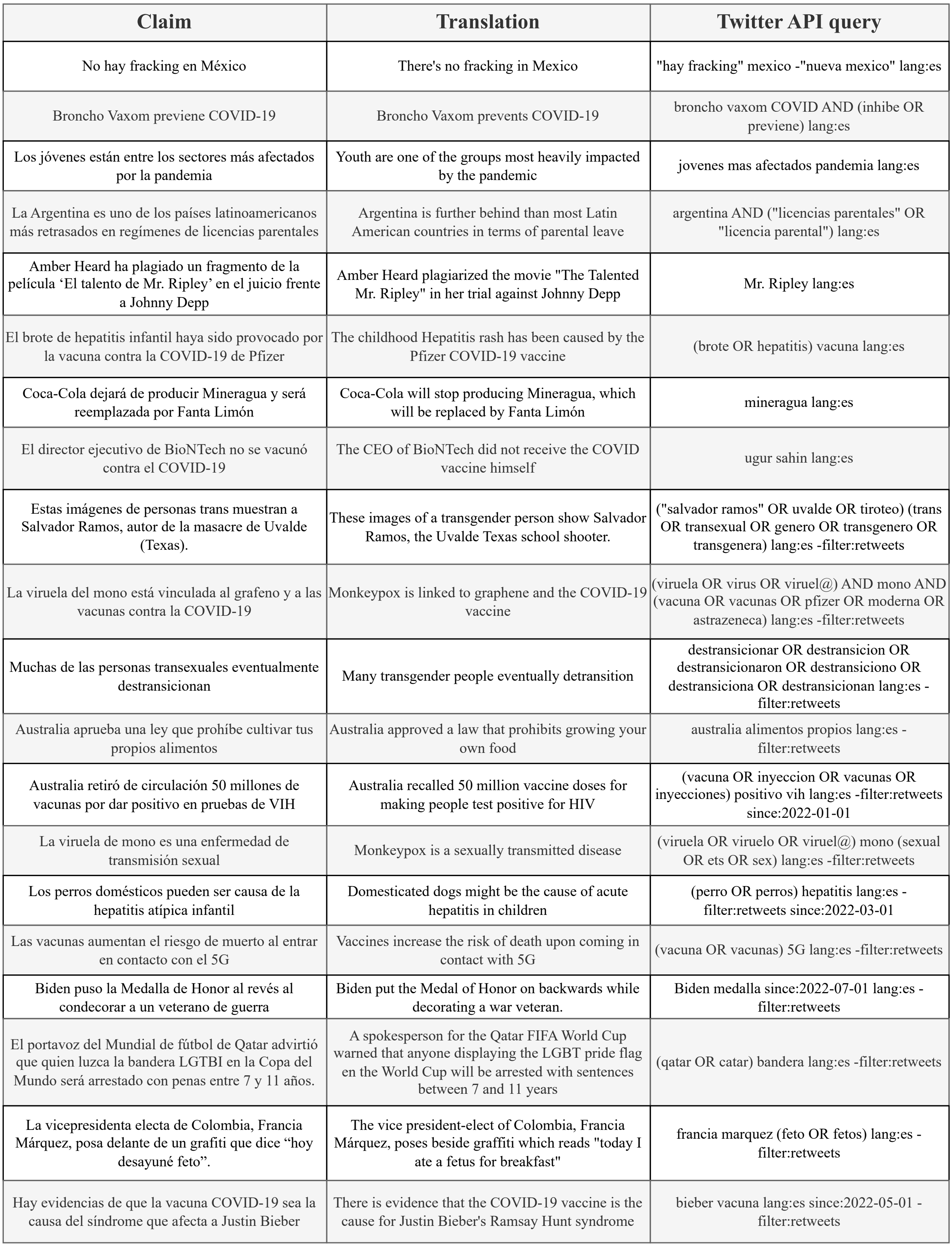
(b)

Figure 2: Label distribution for tweets (by query, not context) in the (a) Russian dataset and (b) Spanish dataset.



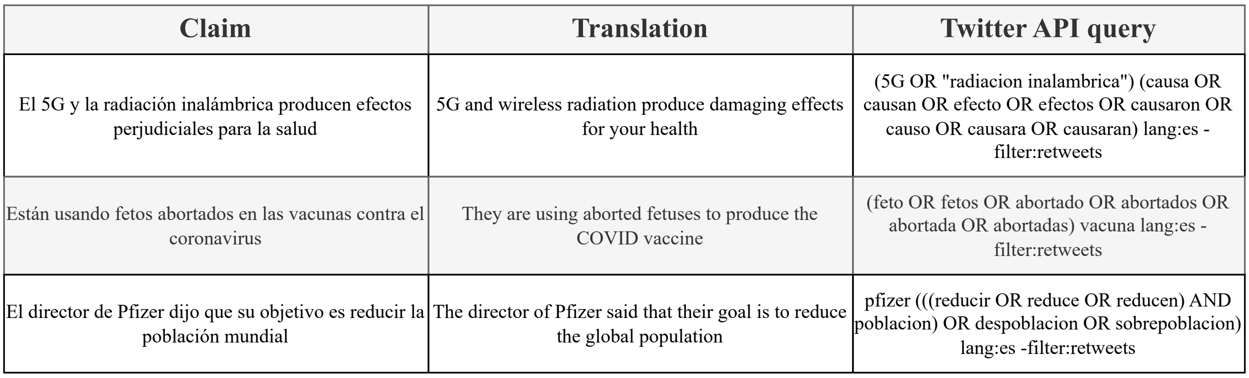
Claim	Translation	Twitter API query
В Украине Воюют войска НАТО.	NATO forces are fighting against Russia in Ukraine.	("войска нато" OR "западные войска") ("в украине" OR "на украине") lang:ru -is:retweet
В украине есть лаборатории которые изготовлают биологическое оружие НАТО.	There are NATO bio-weapon labs in Ukraine.	"био лаборатории" OR "био-лаборатории" lang:ru -is:retweet
Буча – это Украинский фейк.	The Bucha massacre was faked by Ukraine.	буча (фейк OR fake) lang:ru -is:retweet
В Украинском правительстве заправляют нацисты	Nazism is prevalent in many facets of the Ukrainian government.	"нацизм в украине" OR "нацизм на украине" lang:ru -is:retweet
Геноцид Русскоязычных на Донбассе	There is a genocide of Russian speakers in the Donbas region of Ukraine	геноцид (русских OR русскоязычных) (украина OR украине) lang:ru -is:retweet
Украина – агрессор в войне	Ukraine is the aggressor in the 2022 Russo-Ukrainian war	"украина агрессор" lang:ru -is:retweet
В Украине запретили говорить на Русском языке.	Speaking the Russian language has been banned in Ukraine.	(запрет OR запретили) ("русского языка" OR "русский язык") lang:ru -is:retweet
НАТО хочет уничтожить Россию.	NATO wants to destroy Russia.	(НАТО OR запад) (уничтожить OR уничтожит) (Россию OR Россия) lang:ru -is:retweet
Украинцы сбили малазийский самолет MH17.	Ukrainian forces shot down MH17.	(Украина OR Украинцы) (МН17 OR МН-17 OR MH-17 OR MH17 OR MX17 OR MX-17 OR боинг) lang:ru -is:retweet
В Украине планируют строить концлагеря для русских	Ukraine is planning on building concentration camps for Russians/Russian speakers	"концлагеря для русских" OR "концлагеря для русскоязычных" lang:ru -is:retweet
Алексей Навальный мошенник.	Alexei Navalny is a fraudster.	"Навальный мошенник" OR "Навальный жулик" lang:ru -is:retweet
Запад не хочет мира	The west does not want peace (in Russia/Ukraine conflict)	запад не хочет мира lang:ru -is:retweet
Западные агенты подорвали газопровод Северный Поток.	Western agents (Anglo-Saxons according to Dimitry Medvedev) blew up the Nordstream pipeline.	ЦРУ северный поток lang:ru -is:retweet
Владимир Зеленский – наркоман	Volodimir Zelensky is addicted to drugs	зеленский наркоман lang:ru -is:retweet
США строят биолаборатории в странах бывшего СССР.	The US is fixing/organizing biolaboratories in ex-USSR countries	США биолаборатории lang:ru -is:retweet
Европа мерзнет без русского газа.	Europe is freezing without Russian natural Gas.	европа (мерзнет OR мёрзнет) lang:ru -is:retweet
Войска РФ только бьют по военным целям	The Russian Federation only targets military objects in its bombings and does not target civillians or civillian infrastructure	"только по военным целям" OR "только по военной инфраструктуре" lang:ru -is:retweet
Украина самая коррумпированная страна в мире/европе	Ukraine is the most corrupt country in the world/Europe	"Украина самая коррумпированная" lang:ru -is:retweet

Figure 3: Russian Claims and Queries



Claim	Translation	Twitter API query
No hay fracking en México	There's no fracking in Mexico	"hay fracking" mexico -"nueva mexico" lang:es
Broncho Vaxom previene COVID-19	Broncho Vaxom prevents COVID-19	broncho vaxom COVID AND (inhibe OR previene) lang:es
Los jóvenes están entre los sectores más afectados por la pandemia	Youth are one of the groups most heavily impacted by the pandemic	jovenes mas afectados pandemia lang:es
La Argentina es uno de los países latinoamericanos más retrasados en regímenes de licencias parentales	Argentina is further behind than most Latin American countries in terms of parental leave	argentina AND ("licencias parentales" OR "licencia parental") lang:es
Amber Heard ha plagiado un fragmento de la película 'El talento de Mr. Ripley' en el juicio frente a Johnny Depp	Amber Heard plagiarized the movie "The Talented Mr. Ripley" in her trial against Johnny Depp	Mr. Ripley lang:es
El brote de hepatitis infantil haya sido provocado por la vacuna contra la COVID-19 de Pfizer	The childhood Hepatitis rash has been caused by the Pfizer COVID-19 vaccine	(brote OR hepatitis) vacuna lang:es
Coca-Cola dejará de producir Mineragua y será reemplazada por Fanta Limón	Coca-Cola will stop producing Mineragua, which will be replaced by Fanta Limón	mineragua lang:es
El director ejecutivo de BioNTech no se vacunó contra el COVID-19	The CEO of BioNTech did not receive the COVID vaccine himself	ugur sahin lang:es
Estas imágenes de personas trans muestran a Salvador Ramos, autor de la masacre de Uvalde (Texas).	These images of a transgender person show Salvador Ramos, the Uvalde Texas school shooter.	("salvador ramos" OR uvalde OR tiroteo) (trans OR transexual OR genero OR transgenero OR transgenera) lang:es -filter:retweets
La viruela del mono está vinculada al grafeno y a las vacunas contra la COVID-19	Monkeypox is linked to graphene and the COVID-19 vaccine	(viruela OR virus OR viruel@) AND mono AND (vacuna OR vacunas OR pfizer OR moderna OR astrazeneca) lang:es -filter:retweets
Muchas de las personas transexuales eventualmente destransicionan	Many transgender people eventually detransition	destransicionar OR destransicion OR destransicionaron OR destransiciono OR destransiciona OR destransicionan lang:es -filter:retweets
Australia aprueba una ley que prohíbe cultivar tus propios alimentos	Australia approved a law that prohibits growing your own food	australia alimentos propios lang:es -filter:retweets
Australia retiró de circulación 50 millones de vacunas por dar positivo en pruebas de VIH	Australia recalled 50 million vaccine doses for making people test positive for HIV	(vacuna OR inyeccion OR vacunas OR inyecciones) positivo vih lang:es -filter:retweets since:2022-01-01
La viruela de mono es una enfermedad de transmisión sexual	Monkeypox is a sexually transmitted disease	(viruela OR viruelo OR viruel@) mono (sexual OR ets OR sex) lang:es -filter:retweets
Los perros domésticos pueden ser causa de la hepatitis atípica infantil	Domesticated dogs might be the cause of acute hepatitis in children	(perro OR perros) hepatitis lang:es -filter:retweets since:2022-03-01
Las vacunas aumentan el riesgo de muerto al entrar en contacto con el 5G	Vaccines increase the risk of death upon coming in contact with 5G	(vacuna OR vacunas) 5G lang:es -filter:retweets
Biden puso la Medalla de Honor al revés al condecorar a un veterano de guerra	Biden put the Medal of Honor on backwards while decorating a war veteran.	Biden medalla since:2022-07-01 lang:es -filter:retweets
El portavoz del Mundial de fútbol de Qatar advirtió que quien luzca la bandera LGTBI en la Copa del Mundo será arrestado con penas entre 7 y 11 años.	A spokesperson for the Qatar FIFA World Cup warned that anyone displaying the LGBT pride flag en the World Cup will be arrested with sentences between 7 and 11 years	(qatar OR catar) bandera lang:es -filter:retweets
La vicepresidenta electa de Colombia, Francia Márquez, posa delante de un grafiti que dice "hoy desayuné feto".	The vice president-elect of Colombia, Francia Márquez, poses beside graffiti which reads "today I ate a fetus for breakfast"	francia marquez (feto OR fetos) lang:es -filter:retweets
Hay evidencias de que la vacuna COVID-19 sea la causa del síndrome que afecta a Justin Bieber	There is evidence that the COVID-19 vaccine is the cause for Justin Bieber's Ramsay Hunt syndrome	bieber vacuna lang:es since:2022-05-01 -filter:retweets

Figure 4: Part 1 of Spanish Claims and Queries



Claim	Translation	Twitter API query
El 5G y la radiación inalámbrica producen efectos perjudiciales para la salud	5G and wireless radiation produce damaging effects for your health	(5G OR "radiacion inalambrica") (causa OR causan OR efecto OR efectos OR causaron OR causo OR causara OR causaran) lang:es -filter:retweets
Están usando fetos abortados en las vacunas contra el coronavirus	They are using aborted fetuses to produce the COVID vaccine	(feto OR fetos OR abortado OR abortados OR abortada OR abortadas) vacuna lang:es -filter:retweets
El director de Pfizer dijo que su objetivo es reducir la población mundial	The director of Pfizer said that their goal is to reduce the global population	pfizer (((reducir OR reduce OR reducen) AND poblacion) OR despoblacion OR sobrepoblacion) lang:es -filter:retweets

Figure 5: Part 2 of Spanish Claims and Queries