Semi-automatic corpus expansion: the case of stance prediction

Camila Farias Pena Pereira and **Ivandré Paraboni**
University of São Paulo (EACH-USP)
Av Arlindo Bettio 1000, São Paulo, Brazil
camilafpp@usp.br, ivandre@usp.br

Abstract

Stance prediction – the task of determining the attitude or position (e.g., for or against) towards a particular topic in a given text – usually relies on annotated corpora as training data and, since topics are in principle unlimited, so is the need for labelled data about every single topic of interest. As a means to ameliorate some of these difficulties, this work adapts a corpus expansion method developed for sentiment analysis to stance prediction by making use of BERT. The method is then applied to a large (46K) stance corpus covering six topics of political interest, obtaining a 9.9% increase in number of instances. Results from both automatic and human evaluation suggest that adding automatically labelled instances to the original dataset does not harm classification accuracy, and that the automatically generated labels are mostly correct.

1 Introduction

Stance prediction (SP) (Aldayel and Magdy, 2021; Kucuk and Can, 2020) aims to determine the attitude or position (e.g., for or against) towards a particular topic in a given text. The task allows identifying, for instance, whether an individual or group is agreeing or disagreeing with a particular statement, taking a particular stance on a possibly controversial or hateful topic (da Silva et al., 2020) or, more generally, how a piece of text may reflect upon the intended target (e.g., by being for or against it). The latter, also known as target-based SP, is the focus of the present work.

SP usually takes the form of a supervised machine learning task based on annotated corpora (e.g., social media posts manually labelled with for/against information about a particular target), and it is in principle analogous to sentiment analysis (SA), that is, the task of determining positive/negative sentiment in text (Zhang and v Wang, 2018). However, SA is arguably a more shallow NLP task since SA models may in principle use any sufficiently close domain (e.g., movies reviews as in '*the film was terrible*') as training data to infer sentiment in other domains (e.g., product reviews, as in '*the smartphone battery was terrible*'). SP, by contrast, is much more target-dependent, and it is usually necessary to create a new target-specific model from scratch. This means that we may need a new training corpus for every target topic of interest. Consider the following examples.

> (i) *Sure hydroxycloroquine is the right thing to do. You may still die from covid-19, but never from malaria!*
>
> (ii) *If the Sinovac vaccine is so effective, why not even a single European country is using it?*

Both examples convey a stance against a medicine or treatment that has been discussed within the context of the covid-19 pandemic. However, in addition to mixing positive (e.g., 'right'), and negative (e.g., 'die') terms, we notice that both statements have little else in common, for instance, in terms of vocabulary or structure. As a result, a training corpus of stance towards one topic will not necessarily help build a prediction model of stance towards the other and, more importantly, since the number of possible target topics for SP is arguably unlimited, so is the need for labelled data on every single topic of interest.

As a means to ameliorate some of these difficulties, the present work addresses a corpus expansion strategy originally developed for the somewhat shallower sentiment analysis task (Brum and das Graças Volpe Nunes, 2018), and which has been presently adapted for stance prediction in the Portuguese language (dos Santos and Paraboni, 2019; Pavan et al., 2020) with the aid of BERT (Devlin et al., 2019). This strategy has been applied to a corpus expansion experiment, and intrinsic and human evaluation results are reported.

This paper is organised as follows. After a brief overview of existing work on stance corpus resources in Section 2, the present work is divided into two main parts. In the first part, presented in Section 3, we describe the stance corpus to be expanded automatically, the classifier models to be taken as the basis for the expansion method, their individual results and model interpretation. In the second part, described in Section 4, our attention turns to the actual corpus expansion method, describing its architecture and its results (Section 5). Finally, Section 6 presents our main conclusions and suggestions of future work.

2 Related work

Table 1 summarises a number of recent NLP studies that produced larger (over 4,000 instances) corpora for target-based SP, categorised according to text genre, target language (Ar=Arabic, Ca=Catalan, De=German, Du=Dutch, En=English, Fr=French, It=Italian, Pt=Portuguese, Sp=Spanish, *=others), number of instances, and labelling method (t=text-level, u=user-level, p=label propagation).

Regarding the text genre of the existing resources, we notice that Twitter and other social media are common and, as expected, the target language of choice is usually English. We notice also that the number of instances has increased significantly since the original SemEval corpus release (top row of the table), but the larger resources in Magdy et al. (2016); Geiss et al. (2022) are not labelled at the individual text level, resorting instead to label propagation or user-level labelling.

In the case of our target language – Portuguese – we have identified only two relevant studies. The work in Pavan et al. (2023) presents a relatively small corpus of crowd sourced essays about topics of a moral nature (e.g., abortion legislation, same sex marriage, etc.) manually labelled with stance information for classification purposes (e.g., (Flores et al., 2022)). The much larger *UstanceBR* Twitter corpus (Pavan and Paraboni, 2022; Pereira et al., 2023), on the other hand, will be taken as the starting point to our present expansion experiments.

3 Data

We use the UstanceBR Portuguese corpus (Pereira et al., 2023) of labelled tweets conveying 24,995 stances in favour and further 21,857 stances against six topics favoured by either liberal or conservative users, and which are taken as train and test data for our stance classifiers described in the next sections. In addition to that, by following the same procedure described in Pereira et al. (2023), we collected a set of 194,899 unlabelled tweets to be taken as the basis for the corpus expansion experiments described in Section 4. These consist of tweets that happen to mention a keyword of interest (e.g., 'Globo'), but which may or (more often) may not convey an actual stance towards the intended target.

Descriptive statistics of the labelled and unlabelled datasets are summarised in Table 2.

As a means to illustrate the tasks at hand, a standard logistic regression classifier based on TF-IDF counts was built for each task. Table 3 shows the ten most important word features for the positive class (for) of each of the six targets, and weights representing the change of the evaluation score when the corresponding feature is shuffled, as computed with ELI5[1].

To a great extent, the most important features for each classification task are intuitively associated with discourse in support for the corresponding target. These include, for instance, frequent discussions about Lula's trial, praise to Bolsonaro's government, or appreciation for popular Globo's shows. Moreover, after some scrutiny, even less obvious results turned out to be consistent. This is the case of Church goers, among whom the publication of messages as in, e.g., 'I am going to the church tomorrow' seems to be a common expression of faith, and which explains the prominence of 'amanhã' (tomorrow) in the Church topic. On the other hand, as expected from purely data-driven methods of this kind, some features do not seem to be associated with a particular target or stance in any obvious way, and may simply reflect the distribution of this particular dataset. Examples of this kind include the prominent use of 'está' (is) in the Bolsonaro topic, among others.

4 Corpus expansion

From the labelled portion of the data described in the previous section, we built a standard stance classifier, hereby called SM.BERT (softmax BERT) using BERTabaporu (da Costa et al., 2023), a BERT model trained on 237 million tweets in Portuguese. SM.BERT training was performed in one epoch with a batch size of 8, and with a maximum sequence length of 128, and output class labels (for/against) were obtained with the aid of softmax.

Using SM.BERT as a basis, we envisaged a method to (semi-) automatically expand the

[1] https://eli5.readthedocs.io/en/latest/

Ref.	Genre	Language	Instances (k)	Labelling
(Mohammad et al., 2016)	twitter	En	4.2	t
(Magdy et al., 2016)	twitter	En	336.3	p
(Taulé et al., 2017)	twitter	Ca,Sp	10.8	t
(Darwish et al., 2017)	twitter	Ar	33.0	t
(Sobhani et al., 2017)	twitter	En	4.5	t
(Conforti et al., 2020)	twitter	En	51.3	t
(Pavan et al., 2023)	essays	Pt	4.1	t,u
(Mutlu et al., 2020)	twitter	En	14.4	t
(Allaway and McKeown, 2020)	opinions	En	23.6	t
(Lai et al., 2020)	twitter	En,Fr,It,Sp,Ca	14.4	t
(Glandt et al., 2021)	twitter	En	6.1	t
(Jaziriyan et al., 2021)	twitter	Ar	9.6	t
(Geiss et al., 2022)	reddit	En	2,717	u
(Chen et al., 2022)	twitter	En,Fr,De,Du,Sp,*	17.9	t
(Pereira et al., 2023)	twitter	Pt	86.8	t,u

Table 1: Corpora for target-based stance prediction.

Class	Instances
Against	24,995
For	21,857
Unlabelled	194,899

Table 2: Data descriptive statistics.

UstanceBR corpus by adding tweets taken from the unlabelled portion of the data through supervised self-training (Zhu, 2005). This is analogous to the method used in the CasSUL sentiment analysis framework (Brum and das Graças Volpe Nunes, 2018), but (a) presently adapted to the SP task with the aid of BERT instead of count-based (e.g., bag-of-words) text representations, and (b) including a method intended to preserve class balance.

As in Brum and das Graças Volpe Nunes (2018), our approach consists of taking a subset of unlabelled instances from a suitable dataset, and then tentatively labelling the intended corpus with the aid of existing classifiers. The classifiers' output is sorted according to the perceived confidence level, and only the N% most likely instances (i.e., those instances whose probability is above a minimum N threshold value) are added to the corpus. Finally, the expanded corpus is taken as an input for re-training the classifiers in the next round of corpus expansion. This is repeated until overall F1 scores obtained by the classifiers show a significant decrease, suggesting that adding further training data beyond that point would be unhelpful.

Since our unlabelled data largely consists of factual information or otherwise text that simply happens to mention a keyword of interest (e.g., 'church') without any particular value judgement, we estimate that about 90% of unlabelled instances do not convey any stance at all. Thus, in our pilot experiments, we initially considered threshold values of 1%, 5%, 10%, 25% and 40% to select newly classified instances at each round but, as the computational costs of fine-tuning a new BERT model at every round turned out to be prohibitive, our present BERT results are based on the selection of 1% of instances with 5 iterations only.

A significant difference between the present approach and CasSul is that the latter selects all the most likely classifier outputs, which leads to a class imbalance that may have a cumulative effect on the re-training of the classifiers in the next round of corpus expansion. In our current approach, by contrast, the training data is kept constantly class-balanced across rounds by splitting results according to the predicted label (for/against), and by selecting the N% best results from each class separately. This should arguably ensure greater classification accuracy. The top portion of Table 4 shows the number of iterations performed using BERT, and the number of instances that were added to the corpus.

5 Evaluation

We followed the procedure described in the previous section to select 1% of instances during 5 iterations. This added 4648 semi-automatically labelled tweets to the corpus, corresponding to a

Weight	Word feature	Weight	Word feature
	Lula		Bolsonaro
3.586	presidente (president)	5.307	presidente (president)
2.877	moro (a judge's name)	3.760	nosso (our)
2.820	contra (against)	3.360	mídia (media)
2.602	provas (evidence)	2.875	está (is)
2.509	golpe (coup d'état)	2.741	imprensa (press)
2.390	livre (free)	2.724	esquerda (left)
2.308	coração (heart)	2.710	parabéns (congratulations)
2.220	lula	2.642	povo (the people)
2.029	perseguição (persecution)	2.633	stf (the supreme court)
1.942	julgamento (judgement)	2.595	apoio (support)
	Hydroxychloroquine		Sinovac
3.996	anos (years)	3.449	gado (cattle)
3.930	hidroxicloroquina	3.351	bolsonaro
3.923	china	3.021	doses
3.805	vidas (lives)	2.769	butantan (a vaccine producer)
3.698	esquerda (left)	2.352	coronavac (Sinovac)
3.425	governadores (governors)	2.165	bozo (Bolsonaro, derogatory)
3.271	chinês (Chinese)	2.042	mil (thousand)
3.118	globo (Globo TV)	1.984	vacinas (vaccines)
3.101	azitromicina (an antibiotic)	1.833	gente (guys)
2.803	uip (a health authority)	1.822	instituto (institute)
	Globo TV		Church
3.464	amo (I love)	2.989	vou (I am going to church)
3.270	na (on Globo TV)	2.527	ir (go to church)
2.725	parabéns (congratulations)	2.506	saudade (longing)
2.507	série (series)	2.409	nossa (our)
2.487	filme (film)	2.259	amanhã (tomorrow)
2.369	obrigada (thanks)	2.185	hoje (today)
2.251	novela (soap opera)	1.999	maria (Mary)
2.040	passando (broadcasting)	1.952	fui (I went to church)
2.008	plantão (breaking news report)	1.857	indo (going to church)
1.771	bbb (Big Brother Brazil)	1.843	senhor (Lord)

Table 3: Ten most important word features for each stance target.

	Lula	Bolsonaro	Hydrox.	Sinovac	Globo TV	Church
# of Iterations	1	1	1	3	3	5
# of Added instances	275	320	352	1074	897	1730
Original corpus F1	0.76	0.80	0.80	0.81	0.81	0.84
Expanded corpus F1	0.78	0.80	0.80	0.81	0.83	0.85

Table 4: BERT corpus expansion statistics (top) and F1 results (bottom).

	Lula	Bolsonaro	Hydrox.	Sinovac	Globo TV	Church
Agreement %	86.0	90.0	97.0	69.0	91.0	81.0
Marked as 'none' %	9.0	5.0	0.0	2.0	9.0	12.0

Table 5: Agreement between human judges and the corpus expansion method.

9.9% increase. As a means to asses the quality of the added data, we compared models trained from the original corpus data with their counterparts built from the expanded data. Results based on the test portion of the *UstanceBR* corpus are shown in the bottom portion of Table 4, suggesting that the inclusion of semi-automatically labelled instances did not harm performance. In fact, some classes even show a small increase in F1 scores even though the original models had already been optimised for the current dataset.

As a means to further asses the present method, we also performed a human evaluation task. This made use of 100 randomly selected sets of class-balanced instances for each of the six target topics, making 600 evaluation instances in total. Each subset was evaluated by two judges. In case of disagreement, a third judge made the final decision. Unlike our binary (for/against) classifiers, human judges were given the opportunity to choose also a third ('none') label. This was intended to represent cases in which they could not provide a clear for/against answer. Thus, in the present evaluation, the expansion method is to be penalised not only when making explicit mistakes, but also when the text stance is unclear. Agreement results are summarised in Table 5.

Agreement between judges and the expansion method ranged from 69% to 97%, and disagreement generally stemmed from the ambiguity of certain out-of-context tweets, as in '*Great! I hope everyone will be vaccinated soon. Just one question, though: has anyone heard of Sinovac being used anywhere else in the world?*'. In situations of this kind, it is unclear whether the message represents a genuine stance in favour of Sinovac, or whether there is implicit sarcasm. Other than that, we notice also that most cases of disagreement stem from unclear stance (marked as 'none' by the judges), which were beyond the capabilities of our present binary classifiers.

6 Final remarks

This paper presented a SP corpus expansion experiment based on BERT classifiers. The expanded corpus has been subject to both intrinsic and human evaluation, and results suggest that adding automatically labelled instances to the original corpus does not decrease classification accuracy, and that the added instances are mostly correct.

The present work leaves a number of opportunities for improvement. Among these, we notice that the human annotation has only being used as a starting point. It may however be useful to include a human evaluation step in the predict-select cycle as well, which would help prevent the inclusion of noise in the subsequent classifier.

7 Acknowledgements

The present research has been supported by the São Paulo Research Foundation (FAPESP grant #2021/08213-0).

References

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

Emily Allaway and Kathleen R. McKeown. 2020. Zero-shot stance detection: A dataset and model using generalized topic representations. In *EMNLP-2020 proceedings*, pages 8913–8931, Online. Assoc. for Computational Linguistics.

Henrico Bertini Brum and Maria das Graças Volpe Nunes. 2018. Semi-supervised sentiment annotation of large corpora. In *PROPOR-2018 proceedings*, pages 385–395.

Ninghan Chen, Xihui Chen, and Jun Pang. 2022. A multilingual dataset of covid-19 vaccination attitudes on twitter. *Data in Brief*, 44:108503.

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 *ACL-2020 proceedings*, pages 1715–1724, Online. Assoc. for Computational Linguistics.

Pablo Botton da Costa, Matheus Camasmie Pavan, Wesley Ramos dos Santos, Samuel Caetano da Silva, and Ivandré Paraboni. 2023. BERTabaporu: assessing a genre-specific language model for Portuguese NLP. In *Recents Advances in Natural Language Processing (RANLP-2023)*, pages 217–223, Varna, Bulgaria.

Samuel Caetano da Silva, Thiago Castro Ferreira, Ricelli Moreira Silva Ramos, and Ivandré Paraboni. 2020. Data driven and psycholinguistics motivated approaches to hate speech detection. *Computación y Systemas*, 24(3):1179–1188.

Kareem Darwish, Walid Magdy, and Tahar Zanouda. 2017. Improved stance prediction in a user similarity feature space. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 145–148, New York, USA. Assoc. for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-2019 proceedings*, pages 4171–4186, Minneapolis, USA. Assoc. for Computational Linguistics.

Wesley Ramos dos Santos and Ivandré Paraboni. 2019. Moral Stance Recognition and Polarity Classification from Twitter and Elicited Text. In *Recents Advances in Natural Language Processing (RANLP-2019)*, pages 1069–1075, Varna, Bulgaria. INCOMA Ltd.

Arthur Marçal Flores, Matheus Camasmie Pavan, and Ivandré Paraboni. 2022. User profiling and satisfaction inference in public information access services. *Journal of Intelligent Information Systems*, 58(1):67–89.

Henri-Jacques Geiss, Flora Sakketou, and Lucie Flek. 2022. OK boomer: Probing the socio-demographic divide in echo chambers. In *10th International Workshop on Natural Language Processing for Social Media*, pages 83–105, Seattle, Washington USA. Assoc. for Computational Linguistics.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance Detection in COVID-19 Tweets. In *ACL-2021 proceedings*, pages 1596–1611, online. Assoc. for Computational Linguistics.

Mohammad Mehdi Jaziriyan, Ahmad Akbari, and Hamed Karbasi. 2021. ExaASC: A General Target-Based Stance Detection Corpus in Arabic Language. In *11th International Conference on Computer Engineering and Knowledge (ICCKE)*, pages 424–429, Mashhad, Iran. IEEE.

Dilek Kucuk and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys*, 53(1):1–37.

M. Lai, A. T. Cignarella, D. I. Hernandez Farias, C. Bosco, V. Patti, and P. Rosso. 2020. Multilingual stance detection in social media political debates. *Computer Speech and Language*, 63.

Walid Magdy, Kareem Darwish, Norah Abokhodair, Afshin Rahimi, and Timothy Baldwin. 2016. #ISISisNotIslam or #deportallmuslims? predicting unspoken views. In *8th ACM Conference on Web Science*, pages 95–106, New York, NY, USA. Assoc. for Computing Machinery.

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

E.C. Mutlu, T. Oghaz, J. Jasser, E. Tutunculer, A. Rajabi, A. Tayebi, O. Ozmen, and I Garibay. 2020. A stance data set on polarized conversations on twitter about the efficacy of hydroxychloroquine as a treatment for covid-19. *Data in brief*, 33(106401).

Matheus Camasmie Pavan, Vitor Garcia dos Santos, Alex Gwo Jen Lan, Jo ao Trevisan Martins, Wesley Ramos dos Santos, Caio Deutsch, Pablo Botton da Costa, Fernando Chiu Hsieh, and Ivandré Paraboni. 2023. Morality classification in natural language text. *IEEE transactions on Affective Computing*, 14(1):857–863.

Matheus Camasmie Pavan, Wesley Ramos dos Santos, and Ivandré Paraboni. 2020. Twitter Moral Stance Classification using Long Short-Term Memory Networks. In *BRACIS-2020 proceedings LNAI 12319*, pages 636–647. Springer.

Matheus Camasmie Pavan and Ivandré Paraboni. 2022. Cross-target stance classification as domain adaptation. In *Advances in Computational Intelligence - MICAI 2022. LNAI 13612*, pages 15–25. Springer.

Camila Pereira, Matheus Pavan, Sungwon Yoon, Ricelli Ramos, Pablo Costa, Laís Cavalheiro, and Ivandré Paraboni. 2023. UstanceBR: a multimodal language resource for stance prediction. *arXiv:2312.06374*.

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A dataset for multi-target stance detection. In *EACL-2017 proceedings*, pages 551–557, Valencia, Spain. Assoc. for Computational Linguistics.

Mariona Taulé, Maria Antònia Martí, Francisco Manuel Rangel Pardo, Paolo Rosso, Cristina Bosco, and Viviana Patti. 2017. Overview of the task on stance and gender detection in tweets on catalan independence at IberEval 2017. In *IberEval-2017 proceedings*, pages 157–177, Murcia, Spain. CEUR-WS.org.

Lei Zhang and and Bing Liu v Wang. 2018. Deep learning for sentiment analysis: A survey. *WIREs Data Mining and Knowledge Discovery*, 8(4):e1253.

Xiaojin Jerry Zhu. 2005. Semi-supervised learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences.