UlyssesSD-Br: Stance Detection in Brazilian Political Polls

Dyonnatan F. Maia[1(✉)], Nádia F. F. Silva[1], Ellen P. R. Souza[2], Augusto S. Nunes[3], Lucas C. Procópio[3], Guthemberg da S. Sampaio[2], Márcio de S. Dias[3,4], Adrio O. Alves[3], Dyéssica F. Maia[6], Ingrid A. Ribeiro[7], Fabíola S. F. Pereira[5], and André P. de L. F. de Carvalho[3]

[1] Institute of Informatics, Federal University of Goiás, Goiânia, GO, Brazil
dyonnatan@discente.ufg.br, nadia.felix@ufg.br
[2] Centro de Informática, Federal University of Pernambuco, Recife, PE, Brazil
ellen.ramos@ufrpe.br, gss6@cin.ufpe.br
[3] Institute of Mathematics and Computer Science, University of São Paulo, São Carlos, SP, Brazil
{augustonunes,lucascbsi020,adrio20,andre}@usp.br
[4] Federal University of Catalão, Catalão, GO, Brazil
marciodias@ufcat.edu.br
[5] Federal University of Uberlândia, Uberlândia, MG, Brazil
fabiola.pereira@ufu.br
[6] Pontifical Catholic University of Goiás, Goiânia, Brazil
20202013000134@pucgo.edu.br
[7] Faculdade de Ceilândia, University of Brasilia, Brasília, DF, Brazil
ribeiro.ingrid@aluno.unb.br

Abstract. Political bill comments published in digital media may reveal the issuer's stances. Through this, we can identify and group the polarity of these public opinions. The automatic stance detection task involves viewing the text and the target topic. Due to the diversity and emergence of new bills, the challenge approached is to estimate the polarity of a new topic. Thus, this paper evaluates cross-target stance detection with many-to-one approaches in a collected Portuguese dataset of the political pool from the Brazilian Chamber of Deputies website. We proposed a new corpus for the bills' opinion domain and tested it in several models, where we achieved the best result with the mBERT model in classification with the joint input topic and comment method. We verify that the mBERT model successfully handled cross-target tasks with this corpus among the tested algorithms.

Keywords: Stance detection · Political comments · Cross-target

1 Introduction

Stance detection treats to identify and classify the polarity in a text towards a topic [6,7]. Moreover, the polarity classification subtask evaluates whether the

G. Marreiros et al. (Eds.): EPIA 2022, LNAI 13566, pp. 85–95, 2022.
https://doi.org/10.1007/978-3-031-16474-3_8

text supports or is contrary to the topic. So, the input can be composed of the tuple (topic, text), and the output may be a ternary class (against, favour, none) or divided into identification (stance, none) and classification (against, favour). [6]. For political bill comments, the topic's representation can be the bill target itself or the subtopics that affect it. Thus, the topic could be any political, social or economic subject. When new bills get the public attention in Brazil, people can discuss the content and express opinions in public comments, social media debates, and specialised sites. This text data can reveal the yearnings of that population sample, so if we automatically determine their stance, then one advantage is that the author's bill may better comprehend its polarity of acceptance.

To automatically detect the texts' stance on each bill, it is interesting that the method detects points of view from new topics because the interest in these stances is more related to the popularity and deadline for the bill vote. This kind of problem may be treated as cross-target with many-to-one approach, which means it is tested the generalisation model capability with other targets/topics in the same context [7], many-to-one means that we will consider many topics to model training and evaluates each new topic.

Therefore, this paper presents a labelled corpus of surveys about Brazilian political bills extracted from the Chamber of Deputies website and evaluates some models by focusing on checking the cross-target capability with many-to-one approach. As a result, we verify that the BERT-based model overcomes the other tested models. We also make our corpus and models available for use[1] In the end, it discoursed the identified approaches' limitations and research opportunities.

The research reported in this paper has the following contributions: (i) A collected corpus from an online platform that enables all Brazilian citizens to interact and express their opinions concerning bills being discussed by the parliament; (ii) We discuss our annotation protocol and provide statistics about the stance detection corpus; (iii) We evaluate and compare Logistic Regression (LR), Naive Bayes (NB), Support Vector Machine (SVM), Feed-Forward Network (FNN) classical machine learning approaches and a pretrained multilingual BERT-based (mBERT) [4] deep learning model; and (iv) we make our considerations about the results.

This paper is organised into the following sections: Sect. 2 contains related works for stance detection applied to the Portuguese language and the cross-target approaches attached to stance detection. Section 3 relates the process of data collection and annotation and describes the resulting corpus. Section 4 describe the applied methods to generate the models. Section 5 contains the experiments and their results. Finally, Sect. 6 is a conclusion of this work with its contribution and future works.

[1] Dataset and code available at https://github.com/Dyonnatan/UlyssesSD-Br.

2 Related Works

Since the release of SemEval Task 6b [6], we had works exploring traditional methods to achieve the cross-target task, besides approaches that used neural networks with word embeddings and pre-trained models [7].

For architectures applied in Portuguese corpora, we have some baseline studies [8,9,12] that tested the Logistic Regression (LR), Naive Bayes (NB), Support Vector Machine (SVM), Feed-Forward Network (FNN), BiLSTM with FFN, BiLSTM with Attention Mechanism and FFN. Those are known techniques in the English domain, and newer methods have succeeded, whereas the Portuguese language has a drawback of low resources compared to English.

Nevertheless, earlier works in the English language indicate that contextual language representation achieves better results than static representation [11]. The BERT-like model [5] is the prior baseline with some variants; some works [2,13] have an implemented model that consists of a BERT base model where topic and text are joint as input and fine-tuned for the task. The Allaway et al. (2021) [2] tested BERT-joint compared to BERT-sep. Both had a FFN with two layers to the classification output, and BERT-joint showed a better result.

In Reuver et al. [11] study, they proposed to verify whether stance detection is topic-independent and cross-topic generalisable. They conclude that the topic matters despite BERT showing better results than other tested models. The domain, linguistic characteristics and socio-cultural context are some of the main challenges.

We shall remark that our research differs from the existing works due to the aspects: We proposed an annotated corpus with several topics in the Portuguese language to achieve the stance detection cross-target with many-to-one approach in Brazilian Political Polls. We verify how our proposed model performs on elicited and Twitter corpora [12] adjusting for the cross-target with many-to-one task. We verify the capability of our trained model to evaluate on Santos and Paraboni's work [12] with that platform source variation.

3 UlyssesSD-Br Corpus

3.1 Data Collection

The Chamber of Deputies of Brazil website has a section about public opinion polls on political bills. There is a comment field in the poll where participants can write their opinions about the bill content with positive and negative points. Also, there is an option to download the comments on the page poll, whose data are collected anonymised.

Usually, famous political bills receive a nickname for referring to them, so these nicknames can be used to identify the main bill discussion. They are used to express or to represent an idea at the human comprehension level in replacement of the bills' formal names like "PL 10934/2004" to "Budget Guidelines Law", a translation of *"Lei de Diretrizes Orçamentárias"*. So, because this resource can be founded in the comment and brings words with semantic meaning, we picked

them as the target topic. We collected some known nicknames and linked each one to the comments that contain a citation to the nickname, generating a link between the comment and the related topic.

3.2 Annotation

We divided three teams with three volunteers each one for the annotation process. The volunteers were composed of undergraduates and graduates students randomly grouped. The possible classes were defined based on Mohammad (2017) [7]: Favor, to represent the comment towards the topic favourable. Against is a comment stance against the topic. None, when the topic is cited but there is no stance related to the topic. Furthermore, based on the Confortili et al. (2020) [3] was also added the not related class when the comment has no reference to the topic.

The annotation was divided into different documents per group, providing one document per day, varying in 50 to 100 comments per document. If at least one annotator gets confused or does not understand the comment, it is discarded.

In the first phase of conduction, we dedicated the first day to instructions and preparation testing with 50 comments, and a follow-up was carried out on the progress to the next two days. In the last phase, on the fourth day, we execute a new instruction to help with the emerged questions about the text interpretation. The annotation continues until the 13th day of work. Finally, we got the result data and kept the comments agreeing with at least two of the three annotators.

3.3 Data Analysis

We collected 215,712 comments in Portuguese language from 5,266 bills. Where 5093 bills have less than 100 comments, and 2,374 bills have just one comment. Only the four most popular have more than 10000 comments, whereas the most popular PL 3019/2020 have 26065 comments. There were 856 bills collected with their nicknames. They could have more than one nickname, like "*Lei de Diretrizes Orçamentárias*" is also "LDO".

After the annotation 1935 comments were accepted where it was discarded topics with fewer than 5 comments, resulting in 20 topics. Table 2 shows the stances and comments amount from this generated corpus.

4 Experimental Setup

We split the data by topics where the test has "subsistence allowance", "CLT", "LOAS", "Public Servants", and the training has all the other 16 topics that correspond to 22.4% and 77.6%, respectively. Table 3 shows the distribution of the label (Table 1).

Table 1. Examples of topics and comments collected.

Topic	Comment	Stance
CLT	*Falta de respeito com o trabalhador. Retrocessos na CLT.* [Lack of respect for the worker. Setbacks in the CLT.]	Favour
Estatuto do Desarmamento [Disarmament Statute]	*Nós já escolhemos sobre ter o direito e isso foi usurpado pelo "Estatuto do Desarmamento".* [We have already chosen about having the right and this has been usurped by the "Disarmament Statute"]	Against
Servidores Públicos [Public Servants]	*Retira a estabilidade dos futuros servidores públicos e não é justo.* [It takes away the stability of future public servants and it's not fair.]	Favour
Reforma Trabalhista [Labor Reform]	*Reduz direito dos trabalhadores e vai pior a crise brasileira. Chega de reforma trabalhista.* [It reduces workers' rights and will worsen the Brazilian crisis. Enough of labor reform.]	Against
Contratação [Hiring]	*Processo de contratação de servidores comissionados.* [Process of hiring commissioned servants.]	None

Table 2. Comments stance per topic

Topic	Favor	Against	None	NR	Total
Desarmamento [Disarmament]	83	273	24	2	382
Servidores Públicos [Public Servants]	185	46	35	0	266
Contratação [Hiring]	77	164	19	2	262
Código Penal [Penal Code]	194	19	38	0	251
Estatuto do Desarmamento [Disarmament Statute]	8	130	23	0	161
Reforma Administrativa [Administrative Reform]	9	101	4	1	115
Reforma Tributária [Tax Reform]	90	1	7	1	99
CLT	17	55	11	1	84
Reforma Trabalhista [Labor Reform]	1	78	3	0	82
Ajuda de custo [Subsistence allowance]	15	29	4	3	51
Reforma Previdenciária [Pension Reform]	6	29	0	0	35
LOAS	11	7	14	0	32
Partidos Políticos [Political Parties]	0	17	8	1	26
Seguro-Desemprego [Unemployment Insurance]	18	3	2	0	23
Porte de Armas [Possession of Arms]	14	0	2	0	16
Estatuto da OAB [OAB Statute]	5	4	5	0	14
Salário Mínimo [Minimum Wage]	11	0	3	0	14
LDB	7	1	1	0	9
Lei Maria da Penha [Maria da Penha Law]	0	5	1	0	6
Código de Defesa do Consumidor [Consumer Protection Code]	7	0	0	0	7
Overall	787	973	214	16	1935

The input test was followed by joint the topic with the comment (topic + comment). For both identification and classification subtasks, we apply the same methods. For NB, LR, SVM, RF and MLP, we tested some combinations from 1 to 5 n-grams, word and char representation, we found out the (1, 2) n-grams with char tokenisation and removing the diacritic by converting the text to Unicode format was the best configuration. The Majority score (Maj) is based on considering the majority class of each topic as the predicted label.

Table 3. Stance distribution by split corpus

	Favor	Against	None/NR	Total
Train	530	825	147	1502
Test	228	137	69	433

We generate the bag of words and TF-IDF features, compute chi-squared and select the features with a 0.95 p-value, resulting in 175 features. The models were trained with the default Scikit-learn framework [10] setup and received the joint input in a binary classification (stance, none) for the recognition task, where the not related class was also included for the none class. For the classification task, the same corpus was applied, but only comments with stance were included. Furthermore, the classes for this task were in favour and against.

For the BERT model, we used the PyTorch Transformers library version [14] getting pre-trained multilingual BERT base [4] and apply to pair sentence classification, in which the input is composed by joint the tokenised topic with the tokenised comment ($< CLS > topic < SEP > comment < SEP >$) and the output strategy remains the same from other models. We ran the train for 10 epochs and applied AdamW optimiser with a weight decay rate of 0.01 and a learning rate of 2e-5.

For the elicited and Twitter corpora, we also split the train and test set by topic to check the cross-target stance detection validation. We use the "Same sex marriage" and "Church tax exemptions" for testing and the other six topics for training for elicited corpus. We chose another two topics for the Twitter corpus because it does not have these topics, so we selected "Racial quotas and Drugs" legalisation and the other three topics for training. We also verify the mBERT model trained on UlyssesSD-Br to evaluate both corpora to check its performance on these topics from different corpora.

We use the metric weight-averaged F1 for the test validation to compare the models, aiming to minimise the impact of unbalanced data on the scores. To verify the score by polarity we use the macro F1.

5 Results and Discussion

5.1 Experiments in UlyssesSD-Br Corpus

Table 4 shows the weighted average F1 score evaluated in the identification subtask by each model on political bills comments, the models represented were applied the BoW features because they perform better than TF-IDF for the subtask. BERT model outperforms the other models. We can notice that all models, except BERT, have at least one topic with a score fewer than the majority class, which means they do not fully outpass the majority baseline (Maj), just the contextual word representation model did.

Table 4. Weighted-average F1 on the test identification set

Topic	Maj	NB	LR	MLP	SVM	RF	mBERT	Suport
S. allowance	0.799	0.779	0.806	0.777	0.762	0.742	**0.904**	51
CLT	0.791	0.817	0.656	0.522	0.560	0.573	**0.901**	84
LOAS	0.405	0.359	0.521	0.557	0.557	0.585	**0.875**	32
Public Servants	0.620	0.615	0.503	0.349	0.430	0.281	**0.973**	266

Table 5 shows the weighted average F1 score evaluated in the classification subtask by each model on political bills comments. Here we can see the BERT model with score superiority in all topics. The "subsistence allowance" and "CLT" topics have the same majority label (against); otherwise, the other two topics have the favour as the majority class. We can notice that the static models performed similar to the majority for two topics, but the topics with opposite class labels were poorly performed, indicating bias in some common set of tokens as identified polarity. Only the NB TF-IDF version outperforms "Public Servants" between the static token representation models.

Table 5. Weighted-averaged F1 on the test classification set

Topic	Maj	NB	NB*	LR	MLP	SVM	RF	mBERT	Suport
S.allowance	0.524	0.524	0.609	0.524	0.588	0.524	0.571	**0.887**	44
CLT	0.662	0.662	0.650	0.618	0.648	0.648	0.708	**1.000**	72
LOAS	0.464	0.218	0.218	0.218	0.492	0.425	0.615	**0.943**	18
Public Servants	0.712	0.092	0.760	0.092	0.413	0.134	0.364	**0.991**	231

* NB TF-IDF with word tokenization model

Table 6 shows the detailed macro-averaged F1 evaluated in the classification subtask by BERT model. We can verify that despite the unbalanced data, the model evaluates the polarity labels with a non-discrepant score, showing that

the model handled well with unbalanced topic polarity and amount, but overall it has more proportional polarity. The "CLT" topic was fully predicted correctly, and "Subsistence allowance" had the lowest, the only one with results below 0.9 in macro-weighted F1.

Table 6. Macro-averaged F1 on the test classification set for mBERT model

Topic	Against	Favor	All	Suport
Subsistence allowance	0.912	0.839	0.875	44
CLT	1.000	1.000	1.000	72
LOAS	0.923	0.956	0.940	18
Public Servants	0.978	0.994	0.986	231

5.2 Experiments in Elicited and Twitter Corpora

The elicited corpus has a little different context where people argue their opinion about some moral topics, but that relationship between moral topics and stance polarity may also be found in UlyssesSD-Br, once some dealt topics have moral points. Both elicited, and Twitter corpora have more texts by topic than UlyssesSD-Br and also implicit topics. However, it has significative fewer topics than UlyssesSD-Br; this is important to evaluate cross-target with many-to-one approach because we expect that, with more topics, the model has more capability to generalise and thus perform better in unknown topics.

Table 7 summarise the F1 evaluation on elicited corpus by mBERT model trained on Elicited train set for stance identification and also another model trained on the Twitter corpus.The Table 8 summarise mBERT model trained for stance polarity for both corpora.

Table 7. The mBERT weighted-average and macro (against, favor, all) F1 score on the elicited and Twitter corpus identification subtask.

Corpus	Topic	Weighted	Stance	None	All	Suport
Elicited	Same sex marriage	0.870	0.911	0.196	0.553	510
	Church tax exemptions	0.681	0.745	0.418	0.581	510
Twitter	Racial quotas	0.765	0.886	0.245	0.565	3,200
	Drugs legalisation	0.673	0.848	0.169	0.508	1,998

We notice the drop in results, two significant differences in these corpora from UlyssesSD-Br are the number of topics and the implicit topics are more present here too, which suggests that the amount of topic matter for this task, but we need to consider that we do not verify the linguistics phenomena issues for the model, that also impact on the results.

Table 8. The mBERT weighted-average and macro (against, favor, all) F1 score on the elicited and Twitter corpus classification subtask.

Corpus	Topic	Weighted	Against	Favor	All	Suport
Elicited	Same sex marriage	0.877	0.000	0.957	0.478	481
	Church tax exemptions	0.042	0.000	0.269	0.135	411
Twitter	Racial quotas	0.551	0.692	0.338	0.515	604
	Drugs legalisation	0.457	0.493	0.438	0.466	516

5.3 Experiments in Elicited and Twitter Corpora Using the UlyssesSD-Br Model Knowledge

The Table 9 and Table 10 shows the same trained mBERT applied to UlyssesSD-Br tested in the entire elicited and Twitter corpus, respectively. We can verify the model cannot perform so well compared to Table 4, but considering it has been evaluated on unknown corpora seems the model generalisation could perform the cross-target with many-to-one approach in this situation.

Table 9. mBERT weighted-average and macro F1 on the elicited corpus identification subtask

Topic	Weighted	Stance	None	All	Suport
Abortion legalisation	0.650	0.771	0.195	0.483	510
Same sex marriage	0.769	0.807	0.139	0.473	510
Gun ownership	0.690	0.856	0.125	0.491	510
Racial quotas	0.617	0.759	0.196	0.477	510
Church tax exemptions	0.709	0.831	0.202	0.517	510
Drugs legalisation	0.667	0.818	0.251	0.535	510
Criminal age	0.456	0.513	0.257	0.385	510
Death penalty	0.641	0.820	0.126	0.473	510

Table 10. mBERT weighted-average and macro (stance, none, all) F1 on the Twitter corpus identification subtask

Topic	Weighted	Stance	None	All	Suport
Abortion legalisation	0.682	0.835	0.051	0.443	3194
Racial quotas	0.592	0.687	0.186	0.436	3200
Drugs legalisation	0.393	0.425	0.303	0.364	1998
Death penalty	0.439	0.733	0.013	0.373	2563

6 Conclusion

From all tested algorithms, we find that the multilingual BERT-base model in a sentence pair classification, which tokenises the topic and text and joins them for the input model, achieves the best results in overcoming the baseline strategy in the identification and classification phases. We verify that only the contextual word representation model outperforms the majority baseline strategy for all topics. Our model is evaluated in a proposed annotated corpus based on Portuguese comments on Brazilian political polls from the Chamber of Deputies' Bills' opinion website section.

In future work, we plan to investigate other methods to achieve the cross-topic stance detection task and surpass the previous results, such as zero-shot and few-shot stance detection [1,2]. In addition, we will analyse in more detail the what are the semantic and linguistic phenomena barriers for the models.

Acknowledgements. We would like to thank the Ditec (Diretoria de Inovação e Tecnologia da Informação) from the Chamber of Deputies of Brazil for the support.

References

1. Allaway, E., McKeown, K.: Zero-shot stance detection: a dataset and model using generalized topic representations. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 8913–8931. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.emnlp-main.717, https://aclanthology.org/2020.emnlp-main.717
2. Allaway, E., Srikanth, M., McKeown, K.: Adversarial learning for zero-shot stance detection on social media. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4756–4767. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.naacl-main.379, https://aclanthology.org/2021.naacl-main.379
3. Conforti, C., Berndt, J., Pilehvar, M.T., Giannitsarou, C., Toxvaerd, F., Collier, N.: Will-they-won't-they: a very large dataset for stance detection on Twitter. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 1715–1724. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.157, https://aclanthology.org/2020.acl-main.157
4. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805 (2018). http://arxiv.org/abs/1810.04805
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 [cs] (2019). http://arxiv.org/abs/1810.04805, arXiv: 1810.04805
6. Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., Cherry, C.: SemEval-2016 Task 6: detecting stance in tweets. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pp. 31–41. Association for Computational Linguistics, San Diego (2016). https://doi.org/10.18653/v1/S16-1003, http://aclweb.org/anthology/S16-1003

7. Mohammad, S.M., Sobhani, P., Kiritchenko, S.: Stance and sentiment in tweets. ACM Trans. Internet Technol. **17**(3), 26:1–26:23 (2017). https://doi.org/10.1145/3003433
8. Pavan, M.C., et al.: Morality classification in natural language text. IEEE Trans. Affect. Comput., 1 (2020). https://doi.org/10.1109/TAFFC.2020.3034050
9. Pavan, M., dos Santos, W., Paraboni, I.: Twitter moral stance classification using long short-term memory networks. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12319 LNAI, pp. 636–647 (2020). https://doi.org/10.1007/978-3-030-61377-8-45
10. Pedregosa, F., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
11. Reuver, M., Verberne, S., Morante, R., Fokkens, A.: Is stance detection topic-independent and cross-topic generalizable? - a reproduction study. arXiv:2110.07693 [cs] (2021). arXiv: 2110.07693
12. Santos, W., Paraboni, I.: Moral stance recognition and polarity classification from Twitter and elicited text. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), pp. 1069–1075. INCOMA Ltd., Varna (2019). https://doi.org/10.26615/978-954-452-056-4-123
13. Vamvas, J., Sennrich, R.: X-stance: a multilingual multi-target dataset for stance detection. CoRR abs/2003.08385 (2020). https://arxiv.org/abs/2003.08385
14. Wolf, T., et al.: Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45. Association for Computational Linguistics (2020). https://www.aclweb.org/anthology/2020.emnlp-demos.6