

CoFE: A New Dataset of Intra-Multilingual Multi-target Stance Classification from an Online European Participatory Democracy Platform

Valentin Barriere

CENIA

Vicuña Mackenna 4860

Macul, Chile

valbarrierepro@gmail.com

Guillaume Jacquet

Publications Office of the EU

2 rue Mercier

Luxembourg, Luxembourg

name.surname@ec.europa.eu

Léo Hemamou

Sanofi R&D France*

1 av Pierre Brossolette

Chilly-Mazarin, France

l.hemamou@gmail.com

Abstract

Stance Recognition over proposals is the task of automatically detecting whether a comment on a specific proposal is in favor of this proposal, against this proposal or that neither inference is likely. The dataset that we propose to use is an online debating platform inaugurated in 2021, where users can submit proposals and comment over proposals or over other comments. It contains 4.2k proposals and 20k comments focused on various topics. Every comment and proposal can come written in another language, with more than 40% of the proposal/comment pairs containing at least two languages, creating a unique intra-multilingual setting. A portion of the data (more than 7k comment/proposal pairs, in 26 languages) was annotated by the writers with a self-tag assessing whether they are in favor or against the proposal. Another part of the data (without self-tag) has been manually annotated: 1,206 comments in 6 morphologically different languages (fr, de, en, el, it, hu) were tagged, leading to a Krippendorff’s α of 0.69. This setting allows defining an intra-multilingual and multi-target stance classification task over online debates.

1 Introduction and Related Works

Stance recognition is a relevant tool for many real-life applications, from misinformation detection (Hardalov et al., 2021a) or poll verification (Joseph et al., 2021) to large-scale citizen consultation project (Barriere et al., 2022). Some recent work focused on tweets either in a non-interactive manner, like the SemEval-2016 task (Mohammad et al., 2016; Li et al., 2021), or by including the interactions between the users and applying stance detection over the whole thread (Gorrell et al., 2019). When working on online debates, authors employed linguistics-based methods inside debates using pre-defined opposed targets such as “*iPhone vs BlackBerry*” (Somasundaran and Wiebe, 2009), classifying ideological debates (Somasundaran and Wiebe, 2010) and on social justice subjects such

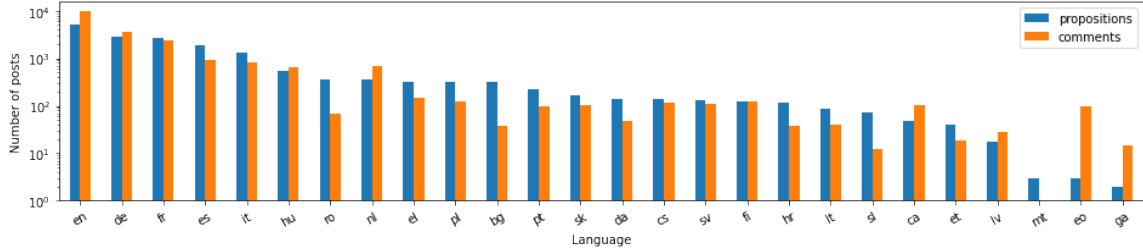
as “*Abortion*” or “*Gay Rights*”. They then used hybrid models, i.e. machine learning models employing linguistic cues as features (Abbott et al., 2011; Barriere et al., 2018). They were followed by more complex probabilistic graphical systems (Walker et al., 2012; Sridhar et al., 2015; Barriere, 2017), allowing to model the dynamics of the debate and the disagreements between speech turns, and finally deep neural methods (Augenstein et al., 2016; Allaway and McKeown, 2020). Sakketou et al. (2022) studied the dynamics of the stances on eight controversial topics in online debates.

On multilingual stance analysis over tweets, Lai et al. (2020) present a model using mainly high-level linguistic features like stylistic, structural, affective or contextual knowledge, but no dense contextual vectors. Hardalov et al. (2021b) proposed a few-shot cross-lingual neural model, by aggregating different language datasets altogether.

Stance annotated datasets are often restricted to a few targets of concepts (Hardalov et al., 2021b). In Vamvas and Sennrich (2020), the authors propose the **X-stance** dataset, containing 67k comments over 150 political issues in 3 languages. Their approach was to reformulate the target in a natural question in order to easily train one multilingual multi-target model on the entire dataset. Similarly, in the *procon* dataset, containing 6,019 comments over 419 controversial issues, each target was also reformulated as a question (Hosseinia et al., 2020). However, none of these datasets contains interactive data. On contrary, Barriere et al. (2022) presented the **Debating Europe (DE)** dataset, a multi-target, multi-lingual stance classification over online debates, integrating the interactive context inside a model. In all the presented works, the language of the comments and propositions are the same, which can be seen as *intra-monolingual*.

Positioning and Motivation Stance recognition is generally restricted to tasks targeting a few defined entities or concepts (Hardalov et al., 2021b;

Figure 1: Number of posts and comments per language, using ISO 3166-1 alpha-2 country codes.



Li et al., 2021). In the proposed dataset, the targets are proposals that can be written in any language, making the task more difficult due to the high variability in terms of topics and in terms of languages.

The work the most similar to ours is the one of Vamvas and Sennrich (2020), where they proposed a somewhat similar framework with the XStance dataset. But in their case, the data they release is restricted to 3 languages and one (small) country only. Another similar work is the one of Barriere et al. (2022), with the Debating Europe (DE) dataset, which contains only 2 languages with intra-monolingual discussions, and annotations just for English only. We differ principally from related works by the multilingual aspect: in our dataset the comments and the propositions in the same discussion can be written in different languages (see examples Table 1). For this reason we name this aspect, specific to our dataset, *intra-multilingual*. To the best of the authors’ knowledge, having several different languages inside the same online debate is specific to our dataset and could not be found in the literature.

The first motivation of this work relates to the lack of an appropriate intra-multilingual multi-target stance-annotated debate dataset. In the context of a citizen consultation project, various questions are asked and contributors can either answer these questions or express their stance by commenting on prior comments made by other users, in a discussion. We created such a corpus, together with the appropriate annotation schema and guidelines. It is also important to note that restricting a dataset to one language could induce nationality or cultural bias.

Contributions The contributions of this paper are the following. Firstly, we propose a new dataset of stance in intra-multilingual online debates, containing binary self-annotations from the users in 34% of the cases. Secondly, we annotate more than 1200 comments in 6 different languages, and obtained a high inter-annotator agreement of 0.69 using Krippendorff’s α .

In the proposed dataset, we want to address the issue of classifying whether a comment is *Pro*, *Against* or *Neutral* towards the proposal it is commenting on. The novelty of this proposed dataset remains in the use of intra-multilingual data and highly variable target. Firstly, the structure of the platform makes it possible for users speaking different languages to interact on the same proposal page, hence the comments and the proposal are not necessarily written in the same language. Secondly, there are many proposals on the CoFE platform, hence the target of the comment (i.e. the proposal) is highly varying in terms of topic and vocabulary.

2 CoFE Dataset

2.1 CoFE Participatory Democracy Platform

The raw data is composed of contemporary questions that are debated in the **Conference on the Future of Europe**¹ (CoFE). CoFE is an online platform in which any user can write a proposal in any of the EU24 languages.² For each proposal, any other user can comment and/or endorse a proposal or another comment. All the texts are automatically translated in any of the EU24 languages.

It contains more than 20k comments on 4.2k proposals in 26 languages. English, German and French are the main languages of the platform. The language distribution can be seen in Figure 1.

Each proposal has been dispatched in one of ten topics by the participants (see Figure 2). As it is shown in Figure 2, some topics are more prone to discussions than others, like *European Democracy* or *Values, Rights and Security*. The topic with the biggest number of propositions is *Climate Change and the Environment*. Examples of proposals, comments and stance labels are shown in Table 1.

2.2 Online Debates with Intra-multilingual Interactions

The CoFE dataset contains long debates with comments answering to each other in the form of

¹<https://futureu.europa.eu/?locale=en>

²And more: we saw people used Catalan and Esperanto






Title	Topic	Proposal	Comment	Stance	url
Focus on Anti-Aging and Longevity research	Health	The EU has presented their green paper on ageing, and correctly named the aging...	The idea of prevention being better than a cure is nothing new or revolutionary. Rejuvenation...	Pro	
Set up a program for returnable food packaging...	Climate change and the environment	The European Union could set up a program for returnable food packaging made from...	Bringing our own packaging to stores could also be a very good option. People would be...	Pro	
Impose an IQ or arithmetic-logic test to immigrants	Migration	We should impose an IQ test or at least several cognitive tests making sure immigrants have...	On ne peut pas trier les migrants par un simple score sur les capacités cognitives. Certains furent la guerre et vous...	Against	
Un Président de la Commission directement élu...	European democracy	Les élections, qu’elles soient présidentielles ou législatives, sont au coeur du processus...	I prefer sticking with a representative system and have the President of the...	Against	
Europa sí, pero no así	Values and rights, rule of law, security	En los últimos años, las naciones que forman parte de la UE han visto como su soberanía ha sido...	Zdecydowanie nie zgadzam się z pomysłem, aby interesy indywidualnych Państw miały...	Against	

Table 1: Examples of comments and proposals with the associated stance

threads, making it possible to study interactions between the users answering each other in different languages. The full dataset is composed of 4,247 debates for a total of more than 15,961 threads of 1 to 4 comments answering to each other, including 5,085 threads of 2 or more comments. The debates rose different interests for the participants: it contains 3,576 debates with five comments or less, but also 382 debates (11,942 comments) with 10 or more comments. Concerning the multilingual aspects: more than 40% of the proposition/comments pairs, as well as 46% of the threads have at least two languages, and 684 debates contain three or more distinct languages. Finally, we also release the number of likes and dislikes of every comment, and the number of endorsements per proposal.³

2.3 Annotation

A portion of the data (more than 7k comments, in 24 languages) has already been annotated by the commenters with a self-tag assessing whether they are in favor or against the proposal. We refer to this set of CF_S . Another part of the data (with no self-tag) has been manually annotated: 1206 comments in 6 morphologically different languages⁴ were tagged by using the Inception platform (Klie et al., 2018). We refer to this set of CF_E .

Annotation Scheme Annotating the stance of a comment over a full proposition is a difficult task, especially when the participant can express several stances inside its comment. For this reason we asked the coders to label not only the prominent stance of the comment but also the secondary

stance if they think there would be a second one. This allows taking care of the cases where there would be several contradictory stances in the same comment in order to consider the mostly agreed stance amongst the coders. In the end, the secondary stances were used to aggregate in 2.2% of the cases.

We collected a total of 3,614 annotated comments that were distributed among 15 different people. More than 80% of the examples were tagged 3 times, the others were tagged 2 times only.

Annotation validation and aggregation The Inter Annotator Agreement was estimated through the use of Krippendorff’s α (Krippendorff, 2013) using only the prominent stance annotations for a 3-classes stance annotation task. We obtained a value of 0.69, which is far more than correct.

The stances were aggregated with a majority vote using the primary stances. The secondary stances were added when there was no consensus using the primary stance (7.8% of the time), and they helped finding a consensus in order to aggregate in 2.2% of the cases.

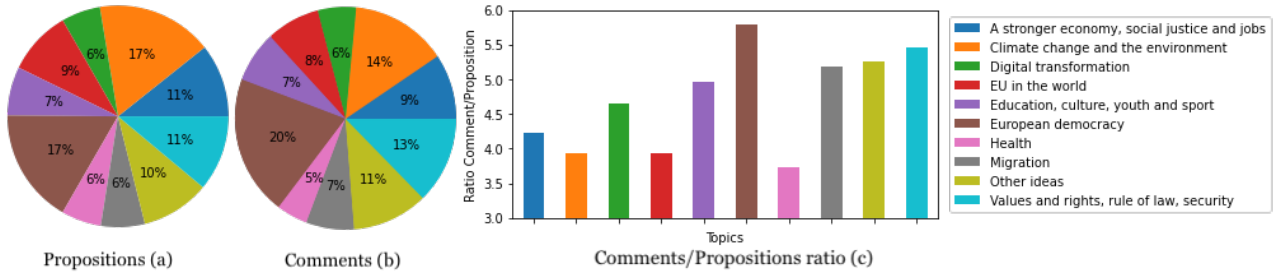
The comments without any consensus in the annotations were discarded, obtaining a total of 1206 annotated comments: 598 English, 241 French, 193 German, 88 Italian, 49 Greek and 37 Hungarian.

Final Datasets We obtained two labeled and one unlabeled datasets. The first one is the self-annotated dataset composed of 6,985 stances with binary annotations, it is called CF_S . The second one is the externally-annotated dataset composed of 1,206 annotated stances with ternary annotations, called CF_E . The last one is the remaining 12,024 unlabeled comments, called CF_U . Table

³A user can endorse a proposal without commenting

⁴fr, de, en, el, it, hu

Figure 2: Topics distribution in the propositions (a), comments (b), and the ratio of comments over propositions (c)



Dataset	XStance	DE	CF _S	CF _E	CF _U
Classes	2	3	2	3	∅
Languages	3	2	25	22	26
Targets	150	18	2,724	757	4,274
Comments	67,271	2,523	6,985	1,206	12,024
Debate	✗	✓	✓	✓	✓
Intra Mult.	✗	✗	✓	✓	✓

Table 2: Comparison with other annotated datasets

2 compares the datasets proposed with two other datasets of stance recognition where the targets are political proposals or questions formulated as text. The CF datasets have the most targets, are intra-multilingual with many languages and contain interactions between users in the form of threads.

3 Baselines

A set of several baselines are proposed over the CF_E dataset. XStance and CF_S are big datasets annotated in a binary way. However, they cannot be used to train a model for a ternary classification. Moreover, the small size of the tri-class dataset makes it difficult to naively aggregate the datasets altogether (model called *All - 1 training*). The protocol of Barriere et al. (2022) has been followed for the training phases. A multilingual pre-trained transformer XLM-R (Conneau et al., 2020) is pre-trained on a 2-class dataset, then fine-tuned over a 3-class dataset with a different classification head in order to obtain a ternary classifier. Each transformer encodes the debate and comments as follows: [CLS] Target [SEP] Comment [SEP]. As Target text, closed questions have been used for XStance and Debating Europe. For CoFE, we simply used the debate title.

Several configurations are compared. A *cross-datasets* model that do not use any of the CoFE data during the training, a *cross-debates* model that trains on XStance and the subpart of CF_S not containing debates that are in the test, and a model that uses the three datasets (*All - 2 trainings*). *Cross-datasets* is pre-trained over XStances and fine-tuned with Debating Europe, *cross-debates* is

Model	-	~	+	Acc.	m-F1
All - 1 training	59.7	00.7	79.5	65.5	46.6
Cross-datasets	54.3	30.5	73.9	59.6	52.9
Cross-debates	55.3	40.4	76.6	63.2	57.4
All - 2 trainings	55.4	44.6	77.3	64.3	59.1

Table 3: F1, macro-F1 and Accuracy of the different baselines over the externally annotated dataset CF_E

trained with XStances and Debating Europe, plus CF_S minus all debates included in CF_E, and *All - 2 trainings* is trained over XStances and CF_S, then Debating Europe. The reader is referred to Barriere et al. (2022) for other details on the training protocol. Accuracy and macro-F1 have been used to reflect both the global and per-class model’s performances. Results can be found in Table 3.

It’s worth noting that the results of the model that is zero-shot regarding the target are still good (57.1 vs 59.1), and that the adaptation towards the domain and languages seems being important (52.9).

4 Conclusion

We presented a new dataset for stance recognition in online debates on contemporary issues related to the future of the European Union, containing 20k comments for 4.2k propositions in 26 languages. This dataset is rich in intra-multilingual interactions between participants, meaning that users can interact with each others using different languages. 46% of the threads have at least two languages. On top of the 7k binary pro/against self-annotations in 25 languages contained in the dataset, a set of 1206 comments from morphologically different languages has been labeled in a 3-class fashion by external annotators. Finally, a few baselines have been tested over the externally annotated dataset CF_E. Future work could embrace using target-based data-augmentation (Li and Caragea, 2021) over our dataset which has a very versatile target space, or integrating the available metadata present in the release, like the number of dis/likes per comment and the number of endorsements per proposal.

References

- Rob Abbott, Marilyn Walker, Pranav Anand, Jean E Fox Tree, Robeson Bowmani, and Joseph King. 2011. [How can you say such things?!?: recognizing disagreement in informal political argument](#). *Proceedings of the Workshop on Languages in Social Media*, pages 2–11.
- Emily Allaway and Kathleen McKeown. 2020. [Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations](#).
- Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. [Stance detection with bidirectional conditional encoding](#). *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 876–885.
- Valentin Barriere. 2017. [Hybrid Models for Opinion Analysis in Speech Interactions](#). In *ICMI*, pages 647–651.
- Valentin Barriere, Alexandra Balahur, and Brian Ravenet. 2022. [Debating Europe : A Multilingual Multi-Target Stance Classification Dataset of Online Debates](#). In *Proceedings of the First Workshop on Natural Language Processing for Political Sciences (PoliticalNLP), LREC*, June, pages 16–21, Marseille, France. European Language Resources Association.
- Valentin Barriere, Chloe Clavel, and Slim Essid. 2018. [Attitude Classification in Adjacency Pairs of a Human-Agent Interaction with Hidden Conditional Random Fields](#). In *ICASSP*.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised Cross-Lingual Representation Learning at Scale](#). pages 31–38.
- Genevieve Gorrell, Kalina Bontcheva, Leon Derczynski, Elena Kochkina, Maria Liakata, and Arkaitz Zubiaga. 2019. [RumourEval 2019: Determining rumour veracity and support for rumours](#). In *SemEval 2019*, pages 845–854.
- Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021a. [A Survey on Stance Detection for Mis- and Disinformation Identification](#).
- Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021b. [Few-Shot Cross-Lingual Stance Detection with Sentiment-Based Pre-Training](#).
- Marjan Hosseinia, Eduard Dragut, and Arjun Mukherjee. 2020. [Stance Prediction for Contemporary Issues: Data and Experiments](#).
- Kenneth Joseph, Sarah Shugars, Ryan Gallagher, Jon Green, Alexi Quintana Mathé, Zijian An, and David Lazer. 2021. [\(Mis\)alignment Between Stance Expressed in Social Media Data and Public Opinion Surveys](#). *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 312–324.
- Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. [The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation](#). *Proceedings of the International Conference on Computational Linguistics*, pages 5–9.
- Klaus Krippendorff. 2013. [Content Analysis: An Introduction to Its Methodology](#). In *Content Analysis: An Introduction to Its Methodology*.
- Mirko Lai, Alessandra Teresa Cignarella, Delia Irazú Hernández Farías, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2020. [Multilingual stance detection in social media political debates](#). *Computer Speech and Language*, 63.
- Yingjie Li and Cornelia Caragea. 2021. [Target-Aware Data Augmentation for Stance Detection](#). In *NAACL*, pages 1850–1860.
- Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021. [P-Stance: A Large Dataset for Stance Detection in Political Domain](#). *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2355–2365.
- Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. [A Dataset for Detecting Stance in Tweets](#).
- Flora Sakketou, Allison Lahnama, Liane Vogel, and Lucie Flek. 2022. [Investigating User Radicalization: A Novel Dataset for Identifying Fine-Grained Temporal Shifts in Opinion](#). In *LREC*, June, pages 3798–3808.
- Swapna Somasundaran and Janyce Wiebe. 2009. [Recognizing stances in online debates](#). *ACL-IJCNLP 2009 - Proceedings of the Conf.*, pages 226–234.
- Swapna Somasundaran and Janyce Wiebe. 2010. [Recognizing Stances in Ideological On-Line Debates](#). In *NAACL Workshop*.
- Dhanya Sridhar, James Foulds, Bert Huang, Lise Getoor, and Marilyn Walker. 2015. [Joint Models of Disagreement and Stance in Online Debate](#). *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 116–125.
- Jannis Vamvas and Rico Sennrich. 2020. [X-stance: A Multilingual Multi-Target Dataset for Stance Detection](#). In *SwissText*.
- Marilyn A. Walker, Pranav Anand, Robert Abbott, and Ricky Grant. 2012. [Stance classification using dialogic properties of persuasion](#). *NAACL HLT 2012 - Proceedings*, pages 592–596.