

Article

Multilingual Multi-Target Stance Recognition in Online Public Consultations

Valentin Barriere ^{1,*}  and Alexandra Balahur ² ¹ Centro Nacional de Inteligencia Artificial, Santiago 4860, Chile² European Commission, Joint Research Center, 1050 Bruxelles, Belgium

* Correspondence: valentin.barriere@cenia.cl

Abstract: Machine Learning is an interesting tool for stance recognition in a large-scale context, in terms of data size, but also regarding the topics and themes addressed or the languages employed by the participants. Public consultations of citizens using online participatory democracy platforms offer this kind of setting and are good use cases for automatic stance recognition systems. In this paper, we propose to use three datasets of public consultations, in order to train a model able to classify the stance of a citizen within a text, towards a proposal or a debate question. We studied stance detection in several contexts: using data from an online platform without interactions between users, using multilingual data from online debates that are in one language, and using data from online intra-multilingual debates, which can contain several languages inside the same unique debate discussion. We propose several baselines and methods in order to take advantage of the different available data, by comparing the results of models using out-of-dataset annotations, and binary or ternary annotations from the target dataset. We finally proposed a self-supervised learning method to take advantage of unlabelled data. We annotated both the datasets with ternary stance labels and made them available.

Keywords: stancerecognition; multilingual models; online debates; public consultations; natural language processing; transformers

MSC: 68T50



Citation: Barriere, V.; Balahur, A. Multilingual Multi-Target Stance Recognition in Online Public Consultations *Mathematics* **2023**, *11*, 2161. <https://doi.org/10.3390/math11092161>

Academic Editor: Florentina Hristea

Received: 21 February 2023

Revised: 29 March 2023

Accepted: 14 April 2023

Published: 4 May 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Stance recognition is a Natural Language Processing (NLP) task that has as its objective the automatic detection and classification of the opinions and attitudes expressed by users in different languages on a wide range of topics. The task has gained momentum in Natural Language Processing during the past few years. As such, different methods have been proposed to tackle it, and various corpora have been developed and employed to train and test stance recognition classification models. Additionally, the increasing availability of multilingual online platforms and social media platforms such as the “*Conference for the Future of Europe*” and other large-scale citizen consultation projects such as *Decidim* (<https://decidim.org/>, accessed on 20 February 2023) or *Make.org* (<https://make.org/>, accessed on 20 February 2023) have led to a growing need for methods to analyse and understand the attitudes and opinions expressed in multiple languages. These platforms provide a unique opportunity to study public opinion on political, societal, and economic issues, which are becoming increasingly important in the context of participatory democracy.

The analysis of multilingual online debates has the potential to provide valuable insights into the attitudes and opinions of citizens from different backgrounds, as well as to identify commonalities and differences among these attitudes across languages and cultures. Furthermore, it can also provide a means to identify and address potential language barriers in democratic processes, by fostering more significant participation in these processes with citizens from different geographical, sociological, and cultural backgrounds. Nevertheless,

there remains a series of important aspects related to this task that have not yet been tackled in the literature. The work presented in this article and the corresponding contributions is motivated by the need to fill existing gaps in research on stance classification, in view of a real-life application of this task in a large-scale citizen consultation project.

Stance recognition algorithms are of interest for multiple reasons. They can be easily deployed on social media or debating platforms [1]. They are heavily used for misinformation and disinformation detection [2–4], but also to predict poll results [5], users polarisation [6], or in order to analyse citizen contributions in a consultation project [7]. Another essential aspect of stance detection is the use of rhetorical strategies, such as hedging, attributions, or denials, that can be used to display varying degrees of certainty or uncertainty [8]. These strategies are particularly relevant in the context of political discourse, where participants may use them to express their stance in a nuanced manner.

Many of the works have focused on data from Twitter, incorporating conversational and interactional context [9,10] in order to better classify the stances of the users in a thread of tweets or simply by taking the tweets in an independent way [11–14]. The SemEval-2017 task 8 [15] proposes to use the interactional context of Twitter threads, focusing on rumour-oriented stance classification, where the objective is to identify support towards a rumour and an entire statement, rather than individual target concepts.

Foundation works were made before Twitter, and hence based on online debate websites [16–18] or more rarely on Congress [19]. The authors proposed to model the text using linguistics features, crafted regarding the targets of the stances, that were fixed, pre-defined, and opposed, such as “*Windows vs. Mac*”. Typical debates were also created around hot social topics in the form of ideological debates on subjects such as “*public healthcare*” or “*gun control*” [17], but also on playful ones such as “*cats*” vs. “*dogs*” [20]. Most of those papers were about hybrid methods mixing statistical learning tools with high-level linguistic features [21,22]. Since then, the conversation has been integrated with graphical models that allow taking into account its dynamics [22–24] through the different successive speech turns of the participants. Neural networks [12,25–27] fall into this type of model and can even be pre-trained for the conversation setting [10] to understand better the conversational context to analyse stances in Twitter threads.

Recently, there have been a few efforts to tackle multilingual stance recognition. Lai et al. [28] presented a model for stance analysis over tweets using mainly high-level linguistic features such as stylistic, structural, affective, or contextual knowledge, but no dense contextual vectors. Hardalov et al. [29] proposed a few-shot cross-lingual neural model by aggregating different language datasets together. The TW-10 Referendum Dataset [30] contains tweets in Catalan and Spanish with stance annotation towards the independence of Catalonia. All of them are tweet datasets.

Stance-annotated datasets containing highly varying targets are rare. They usually focus on a set of defined targets or concepts [14,29]. Building on seminal work on stance, the SemEval 2016 task [11] was capable of targeting abstract concepts (e.g., “*atheism*” or “*abortion*”), as well as persons (e.g., “*Hillary Clinton*” or “*Donald Trump*”). One example of a dataset with highly varying targets is Stanceosaurus [31], which contains tweets in English, Hindi, and Arabic annotated with stance towards 251 misinformation claims over a diverse set of global and regional topics. Sobhani et al. [32] proposed multi-target stance classification that includes two targets per instance, such as classifying the stances of a tweet in relation to both Sanders and Trump. While the framework permits more than two targets, it is still limited to a specific and defined set of targets. It has been extended when the targets can be a written concept or proposal [12,33]. Finally, Deng et al. [34] proposed complex models for cross-target stance recognition, applied to a small set of specific targets.

In Vamvas and Sennrich [33], the authors proposed the X-stance dataset, containing 67k comments over 150 political issues in three languages. Their approach was to reformulate the target in a natural question in order to train one multilingual multi-target model on the entire dataset easily. Similarly, in the *procon* dataset, containing 6019 comments over 419 controversial issues, each target was also reformulated as a question [35]. This

allows using the semantic knowledge encoded inside a pre-trained language model [36] and implicitly captures relationships between topics [25]. Hardalov et al. [37] combined this technique with label embedding [38] in order to train on 16 English datasets from various domains. However, none of these datasets contain interactional data from multilingual online political debates. On the contrary, Barriere et al. [39] presented the **Debating Europe (DE)** dataset, a multi-target, multi-lingual stance classification over online debates, integrating the interactional context inside a model. In all the presented works, the language of the comments and propositions is the same, which can be seen as *intra-monolingual*. Finally, Barriere and Jacquet [7] presented the **CoFE** dataset, which was collected from an online debating platform that contains 4.2k proposals and 20k comments in various languages. A particularity of this dataset is that the comments and the propositions in the same discussion can be written in different languages because of the use of a machine translation system on the online platform, making it *intra-multilingual*.

Another classical issue is that, when the labels are scarce because of the difficulty or time needed to annotate, it is possible to use several techniques to take advantage of the available resources [40]. Hardalov et al. [29] proposed a novel noisy sentiment-based stance detection pre-training leveraging Wikipedia data, for cross-language few-shot learning. Semi-supervised learning methods such as self-training [41], label propagation, or label spreading [42,43] are also profitable options. Giasemidis et al. [44] used the latter methods for rumour-related stance recognition over Twitter data. A recent work on the domain is that of [45], which used knowledge distillation on COVID tweets for the same type of task. On other types of data, Wei et al. [46] proposed an interesting self-training method for imbalanced images on CIFAR, but they assumed the distributions of the unlabelled and labelled datasets were the same, which is a strong assumption that is not true every time.

In this work, we focused on studying models that analyse the stances of a comment on a target formulated in natural language, not necessarily with the same language. This setting makes the task more difficult due to the high variability in terms of topics and in terms of languages. It is also important to note that restricting a dataset to one language could induce nationality or cultural bias. To the best of the authors' knowledge, having several different languages inside the same online debate is specific and could only be found in the literature in [7]. Here, we address the problem of ternary stance classification, i.e., whether a comment is *pro*, *against*, or *other* towards the proposal it is commenting on. Moreover, we propose to use two approaches to learn even with limited labels: a pre-training over other similar datasets [40] and a semi-supervised learning self-training method [41] to take advantage of large available datasets that are not annotated.

This research aimed to contribute to the field of multilingual stance recognition by addressing the challenges and opportunities presented by analysing online multilingual debates. In particular, the paper focuses on developing models and methods for recognising the stance of users in different languages on a given topic and how to make use of the cross-lingual information present in the debates. Section 2 refers to the three stance datasets mainly used in this work and especially the collection and annotation of two of them. Section 3 refers to the Machine Learning experiments and Section 4 to the results and discussions of these experiments.

2. Materials and Methods

In this section, we describe all the datasets we used in our experiments, the methods employed to collect and annotate them when applicable, as well as the details of the training models. The datasets used were the X-stance dataset [33], the Debating Europe dataset [39], and the CoFE dataset [7].

2.1. Debating Europe Dataset

We released the Debating Europe (DE) dataset, composed of online debates annotated with stance annotations at the comment level.

2.1.1. Data Extraction

The DE dataset consists of debates collected in September 2020 from the “*Debating Europe*” platform (<https://www.debatingeurope.eu/>, accessed on 20 February 2023). Most of the debates revolve around questions such as “*Should we have a European healthcare system?*” or “*Do the benefits of nuclear power outweigh the risks?*”, which can typically be rephrased as yes/no questions. Each debate includes a topic tag, a text paragraph providing the context of the debate, and comments about either the main context or previous comments.

We used a scraped version of the “*Debating Europe*” website containing all the debate questions with their associated presentation texts, comments, and replies to comments. Examples of conversations can be seen in Figure 1.

The dataset contains 125,798 comments for 1406 debates. Additional statistics are provided in Table 1. More information about the general distribution of the words is available in Appendix B, Table A1.

Table 1. Low-level statistics on the DE dataset, regarding the presence or absence of label annotation. μ_{com}/μ_{deb} is the average mean of the respective units (comments or words) at the comment/debate level.

Label	% DE	Unit	μ_{com}	μ_{deb}	Σ
✗	100%	Comments	∅	89.5	125,798
		Words	51.7	4623	6,499,625
✓	2.0%	Comments	∅	140	2523
		Words	33.4	4683	84,289

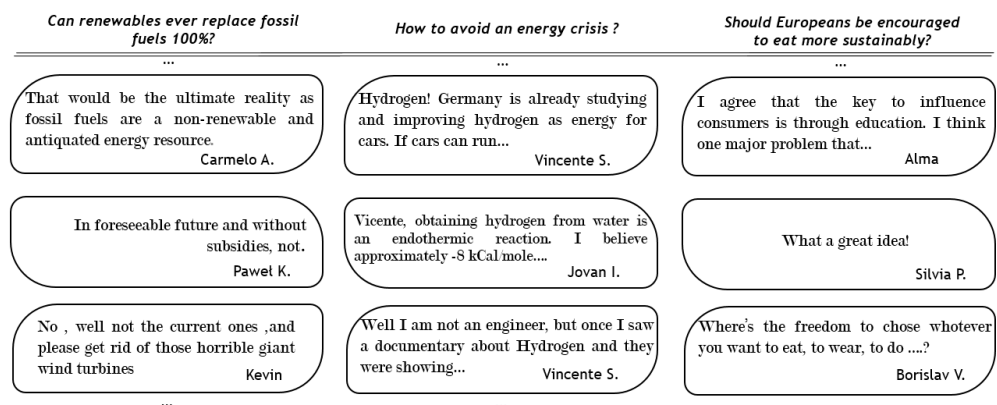


Figure 1. Examples of comments from 3 debates of the Debating Europe Dataset.

2.1.2. Annotation

Subset Selection

We annotated 18 debates from the entire dataset scraped from Debating Europe. The criteria for selecting these debates were the number of comments associated with each debate and their relevance to one or more of the new “*European Green Deal*” policy areas. These policy areas are biodiversity, from farm to fork, sustainable agriculture, clean energy, sustainable industry, building and renovating, sustainable mobility, eliminating pollution, and climate action. More information is available online (<https://tinyurl.com/GreenDealEC>, accessed on 20 February 2023).

To filter the debates, we used the metadata from the “*Debating Europe*” website, keeping only those with the “*Greener*” tag, resulting in 150 debates. Finally, we selected the ones with at least 25 comments. When necessary, the debate question was reformulated into a closed question to make it compatible with our framework. Additional information about the debates and policy areas can be found in the Appendix A.

Annotation Scheme

The annotation scheme and guidelines were developed to identify citizens' stance toward the debate question at the comment level. To achieve this, four labels were defined: *Yes*, *No*, *Neutral*, and *Not answering*. Each comment was annotated to indicate whether the user responded to the answer and, if so, whether they were in favour of, against, or neutral with respect to the original question. The questions of the annotated debates can be found in Appendix A. The annotation task was carried out by a single expert using the INCEpTION software [47].

Final Annotations

We obtained 2523 labels for the 18 debates, with four classes: *Yes* (40.1%), *No* (19.4%), *Neutral* (11.2%), and *Not answering* (29.3%). We included the last category to determine whether the commenter was interested in answering the debate question. In the following experiments, we merged the *Neutral* and *Not answering* classes into a single class to simplify the work [11,48]. Since a single expert performed the annotation, validating the dataset using classical inter-annotator agreement metrics was impossible. Instead, we validated the dataset by demonstrating its usefulness for cross-datasets, cross-topics, and cross-lingual transfer learning. The results are presented in Section 3.1.1. The annotated dataset consisted of 2523 comments, totalling 84,289 words. Additional information about the overall distribution of words can be found in Table 1 and in Appendix B, Table A1.

2.2. CoFE

We released the CoFE dataset, which is composed of multilingual online debates over contemporary hot topics. It has been partially annotated in stance at the comment level by the commenters themselves when they were posting their comments or by external coders afterwards. The text of the proposals and comments have been automatically translated so that participants can interact with each other in their native languages. Here, we present the data collection process and the annotation plus the validation of the annotation, used to create the several subdatasets used in this study: CF_S , CF_U , CF_{E-D} , and CF_{E-T} .

2.2.1. CoFE Participatory Democracy Platform

The raw data used in this study consisted of current questions being debated at the Conference on the Future of Europe (CoFE) (<https://futureu.europa.eu/?locale=en>, accessed on 20 February 2023). The CoFE is an online platform where users can write proposals in any of the EU24 languages (and more: Catalan and Esperanto have been observed to be used on the platform). Users can also comment on and endorse proposals or like other comments. All texts are automatically translated into any of the EU24 languages.

The dataset includes more than 20,000 comments on 4200 proposals in 26 languages, with English, German, and French being the most-commonly used languages on the platform. The language distribution is shown in Figure 2.

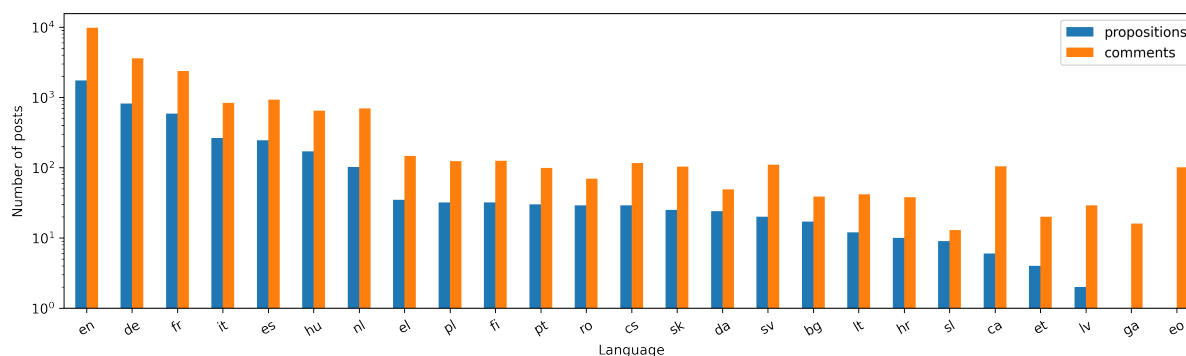







Figure 2. Number of posts and comments per language, using ISO 3166-1 (https://en.wikipedia.org/wiki/ISO_3166-1, accessed on 20 February 2023) alpha-2 country codes.

The proposals in the dataset have been dispatched into one of ten topics by the participants, as shown in Figure 3. As can be seen in the figure, some topics, such as “European Democracy” and “Values, Rights, and Security” have generated more discussion than others. The topic with the largest number of proposals is “Climate Change and the Environment”. Table 2 provides examples of the proposals, comments, and stance labels.

Table 2. Examples of comments and proposals with the associated stance (url links in the appendix).

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	

2.2.2. Online Debates with Intra-Multilingual Interactions

The CoFE dataset includes long debates with comments organised into threads, allowing for the study of interactions between users responding to each other in different languages. The full dataset consists of 4247 debates with a total of over 15,961 threads, including 1 to 4 comments in response to each other and 5085 threads with 2 or more comments. The distribution of threads by length is shown in Table 3. The debates have generated a range of interests among the participants, with 3576 debates containing 5 comments or fewer and 382 debates having 10 or more comments, reaching a total of 11,942 comments.

In terms of multilingual aspects, more than 40% of the proposal/comment pairs, as well as 46% of the threads include at least two languages, and 684 debates contain three or more distinct languages. Finally, the dataset also includes the number of likes and dislikes for each comment and the number of endorsements for each proposal (a user can endorse a proposal without commenting).

Table 3. Number of threads regarding their length in term of comments.

Length	1	2	3	4	All
Number	10,876	2365	1920	800	15,961

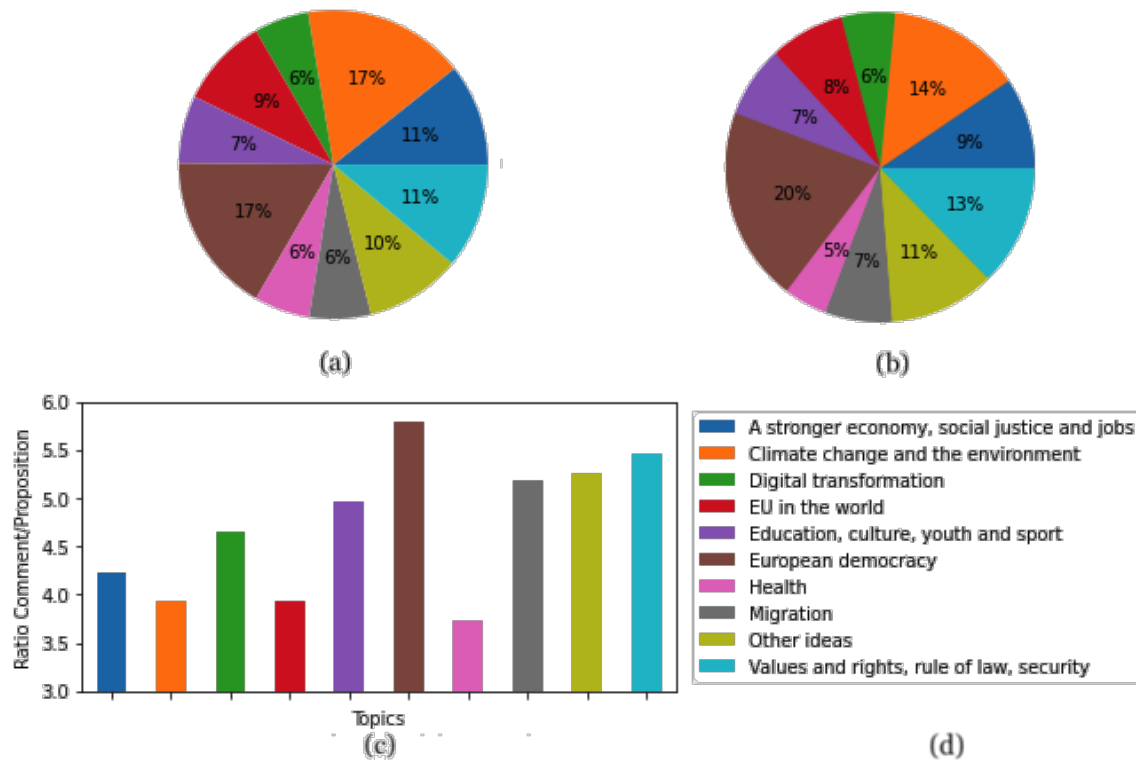


Figure 3. Topics' distribution in the propositions (a), comments (b), and the ratio of comments over propositions (c) regarding the different topics (d).

2.2.3. Annotation

A portion of the data (more than 7 k comments, in 24 languages) has already been annotated by the commenters with a self-tag assessing whether they are in favour of or against the proposal. We refer to this set as CF_S (as Self-annotated). Two other subparts of the data (without a self-tag) have been manually annotated: one to be used for testing purposes and another one to be used for training purposes. The subset of 1283 comments in six morphologically different languages (fr, de, en, el, it, hu) was tagged using the Inception platform [47]. We refer to this set as CF_{E-T} (as Externally annotated-Test). Another subset of 1500 comments in the most-common languages of the platform (fr, de, en, es) was tagged by using the Inception platform [47]. We refer to this set as CF_{E-D} (as Externally annotated-Test).

Annotation Scheme

Annotating the stance of a comment on an entire proposition can be challenging, in particular when the participant expresses multiple stances within his/her comment. To address this issue, we asked the coders to label not only the prominent stance of the comment, but also any secondary stance they believed to be present. This allows for the consideration of cases where multiple contradictory stances are present within the same comment in order to determine the stance that is most-commonly agreed upon by the coders. In the end, the secondary stances were used to aggregate in 1.0% of the cases.

For CF_{E-T} , we collected a total of 3814 annotations that were distributed among 15 different people. More than 95% of the examples were tagged three times, and the others were tagged two times only. For CF_{E-D} , we collected a total of 3500 annotations that were dis-

tributed among four different people. The French and English comments (1000 comments) were tagged three times, and the German and Spanish (300 + 200 comments) were tagged by only one annotator, which is reliable since they were not used for testing purposes. We manually removed four proposals that were not real debates and 61 of the 200 annotated Spanish comments that were judged of bad quality.

Annotation Validation and Aggregation

The inter-annotator agreement for a three-class stance annotation task was evaluated using Krippendorff's α [49] using only the prominent stance annotations. It yielded a value of 0.68 for CF_{E-T} , which is considered satisfactory for this type of task. It should be noted that the level of agreement among annotators can vary greatly depending on the specific target of the stance detection task and the annotators' confidence in their annotations [50]. For CF_{E-D} , the obtained Krippendorff's α was 0.61, which is less good. This is not a problem since these data are considered and used as the silver standard and not the gold standard.

The stances were aggregated through a majority vote using the primary stances. The secondary stances were included in cases where there was no consensus using the primary stance (3.4% of the time), and they helped to reach a consensus in 1.0% of cases. Comments without consensus in the annotations were discarded for both cases. A total of 1228 annotated comments were obtained for CF_{E-T} : 600 English, 241 French, 230 German, 88 Italian, 37 Hungarian, and 32 Greek. A total of 1414 comments were obtained for CF_{E-D} : 675 English, 300 French, 300 German, and 139 Spanish.

Final Datasets

We obtained three labelled datasets and one unlabelled dataset. The first labelled dataset, called CFS, consists of 6985 stances with binary annotations that were self-annotated. The second and third labelled datasets, called CF_{E-T} (this version is slightly bigger than the one from [7]) and CF_{E-D} (as Test and as Development, respectively), consist of, respectively, 1226 and 1414 multilingual comments with ternary annotations that were externally annotated. The fourth dataset, called CF_U , is the remaining unlabelled comments.

2.3. Dataset Generalities

All the datasets used in this paper have common properties: they contain short texts written in the context of, or answering to, a controversial question of political range. In this layout, the targets of the stance are not a defined person or subject. They vary greatly and are expressed in the form of text in natural language. Table 4 compares the three 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.

Table 4. Comparison with other annotated datasets.

Dataset	X-Stance	DE	CF_S	CF_E	CF_U
Classes	2	3	2	3	\emptyset
Languages	3	2	25	22	26
Targets	150	18	2724	757	4274
Comments	67,271	2523	6985	1206	12,024
Debate	✗	✓	✓	✓	✓
Intra Mult.	✗	✗	✓	✓	✓

2.4. X-Stance Dataset

The X-stance (XS) dataset [33] is a collection of 67,271 comments in French, German, and Italian on more than 150 political issues (referred to as *targets*) extracted from the Swiss application *Smartvote*. Each of the comments is associated with a label. To leverage the semantics information contained in a pre-trained model [51], the authors proposed incor-

porating the target into a natural language question, such as “*La Suisse devrait-elle conclure un accord de libre-échange avec les Etats-Unis?*”, which can be interpreted as a debate title. This allows the model to learn across targets and perform effectively in a zero-shot learning setting. Indeed, this approach in which the target can be viewed as a debate title enables the model to learn across targets, maintains efficiency in a zero-shot learning scenario, and leverages the knowledge transfer capability of transformer-based language models [51] (this method has also been used by others [25,37] for zero-shot stance classification).

The annotations from the annotators were consolidated into two classes: *in favour* of and *against* the proposition, which can be represented as *yes* or *no* when the proposition is phrased as a question.

3. Experiments

We ran two different sets of experiments of very similar model types, detecting the stances of comments toward a proposal formulated in natural language. The first set presented in Section 3.1 targeted the Debating Europe as the test dataset, focusing on cross-lingual transfer learning, integrating context and semi-supervised learning. The second set of experiments presented in Section 3.2 targeted the CoFE dataset as the test dataset, proposing several baselines on multilingual data.

3.1. Debating Europe

The three experiments below are complementary. The first experiment focused on transfer learning across topics, targets, and languages. The second one focused on the interactive aspect of online debates. The last experiment highlights the value of the unlabelled DE dataset, by presenting a self-training method handling unlabelled and imbalanced data.

3.1.1. Multilingual Stance Detection Using Transfer Learning

It is well known that when the source and target domains are dissimilar, standard transfer learning may fail and result in negative transfer [52]. Therefore, demonstrating that the small DE dataset can improve the performance on the 25-times larger non-English XS dataset through transfer learning across topics and languages is a way to validate the annotations. The XS dataset, which consists of multilingual comments responding to political debate questions from various topics, is an ideal candidate for transfer learning. The DE dataset consists of comments from the online debate forum Debating Europe, so the targets of the stances are closely related to those in the multilingual XS dataset. For these reasons, we first investigated the potential of using our labels to enhance performance across different topics and languages.

3.1.2. Data Augmentation with Semi-Supervised Learning

As mentioned in Section 2.1, we annotated only a small portion of the available DE dataset, leaving a large amount of data unlabelled, which could potentially be useful in improving model performance. To maximise the potential of this unlabelled data, we propose to use a self-training method [41]. The general principle we followed was to leverage some of the model’s own predictions on unlabelled data by adding pseudo-examples to the training set in an iterative way. Typically, new unlabelled examples were selected regarding how confident the prediction of their label is, and they were added to the training set for the supervised step of the next iteration. We compared two classical methods: using a threshold on the model’s class probability and selecting the k predictions with the highest probability (respectively referred to as *thresh* and *k-best* in Section 4.1.2). We are aware of the potential drawbacks of self-training, such as the inability of the model to correct its own mistakes and the amplification of errors [53]. Thus, if the unlabelled dataset is imbalanced, the classifier bias may be amplified by the pseudo-labels, exacerbating the class imbalance issue [46].

To mitigate this risk, we propose a technique that combines both methods by adding a definite and balanced number of k_{max} examples chosen randomly from those with a

probability above the threshold, at each iteration of the ST algorithm. Our technique makes no assumptions about the label distribution of the unlabelled dataset and, at the same time, can help to prevent the training set from being flooded with pseudo-examples from outer domains.

3.2. Experiences on CoFE

3.2.1. Multilingual Stance Detection Using Transfer Learning

A set of several baselines is proposed over the CF_{E-T} dataset, which is the subpart that had been externally annotated to be used as a test set. X -stance 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 DE dataset makes it difficult to naively aggregate the datasets altogether (the model called *All-1 training*).

Several configurations were compared. First, we compared the models that do not use any comments from the CoFE dataset. Subsequently, we compared the models that use only binary annotation from the CoFE dataset and, finally, the models that use ternary annotations during the training. First, we trained a *cross-datasets* model that does not use any of the CoFE data during the training, and we compared it to two strong baselines trained on stance recognition from various domains: an English model [37] trained over 16 English stance datasets from various domains and a multilingual model [29] pre-trained over the same 16 English datasets and fine-tuned over 14 non-English datasets. Second, we present a *cross-debates* model trained on X -stance and the subpart of CF_S not containing debates from the test and two models that use the three datasets (*All-2 trainings* and *All-1 training*). Third, we present models trained with the CF_{E-D} dataset of ternary stance annotations from the CoFE, alone (*CF_{E-D-1} training*) or with other data using a one-step (*All-1 training*) or a two-step (*All-2 trainings*) training process.

If not specified, all of our models were trained using a two-step training process: trained over binary data, then fine-tuned over ternary data. *Cross-datasets* was pre-trained over X -stances and fine-tuned with Debating Europe. *Cross-debates* was trained with X -stances and Debating Europe, plus CF_S minus all debates included in CF_E . *All-2 trainings* was trained over X -stances and CF_S , then Debating Europe (and CF_{E-D} when the case is warranted). *All-1 training* was trained over X -stances and CF_S and Debating Europe (and CF_{E-D} when the case is warranted).

3.2.2. Data Augmentation with Semi-Supervised Learning

As in Section 3.1.2, we ran experiments with self-supervised learning. We used the model that gave the best results of the transfer learning experiments, by adding the unlabelled CF_U dataset during the second step of the learning phase. We followed the same protocol as specified before.

3.3. Methodological Protocol

In our study, the protocol of [36,54] was followed for training transformers, which had previously been used for multilingual sentiment analysis and text classification. The `transformers` library [55] was used to access pre-trained models and to train our models. XLM-R [56] was employed as a multilingual learning model, referred to as $XLM-R_{ft}$ when it had been previously trained on a dataset (as described in Section 3.1).

For optimisation, the Adam algorithm [57] with early stopping based on the training loss was used. The learning rate was set to 2×10^{-6} for the first training of the model on a stance task and to 5×10^{-7} when fine-tuning on another dataset for transfer learning. Performance on the development set was evaluated after each training epoch, and the model that achieved the best performance was kept. The batch size was set to 32. Unlike [33], no hyperparameter optimisation was performed on the development set, and a shorter maximum sequence length (128 instead of 512) was used to speed up training and evaluation.

The transformer encoding of the debate and comments was carried out according to the protocol of [33], in which each transformer was used as follows:

[CLS] Target [SEP] Comment [SEP]

For X-stance and Debating Europe, closed questions were used as the target text. For the CoFE, the debate title was simply used.

For the transfer learning, a multilingual pre-trained transformer XLM-R [56] was 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. For the ST, a maximum of five iterations was set out, with a probability threshold of 0.99 and 600 and 2000 as the maximum number of examples added at each iteration when applicable.

Metrics widely employed for this kind of task were computed in order to compare our models: the accuracy, precision, recall, as well as macro-F1 score, in order to reflect both the global and per-class model's performances and take into account class imbalance. The DE dataset was divided into three training/validation/test sets in a stratified way with a ratio of 75/5/20. To compare the results, the same partition as [33] was carried out for the XS dataset. CF_{E_T} was used as the test set for the CoFE. Experiments were run using Tensorflow 2.4.1 [58], transformers 3.5.1 [55], a GPU Nvidia RTX-8000, and CUDA 12.0.

4. Results and Discussion

This section presents the results of the experiments over the Debating Europe (Section 4.1) and CoFE datasets (Section 4.2). It highlights how models can take advantage of datasets, even though the regimes are cross-lingual, cross-topics, and even cross-tasks in the case of binary labelled data. It also shows the efficiency of multilingual self-supervised learning for this kind of data and task.

4.1. Results on Debating Europe

The experiments were complementary. The first one gave an insight into the effect of pre-training a classification model over a non-English multilingual dataset from another domain. The second experiment used a self-training method applicable to a dataset of unlabelled and imbalanced data.

4.1.1. Cross-Datasets Transfer Learning

Here, we investigated the effects of pre-training over one dataset before fine-tuning over another one. Table 5 shows the results of applying transfer learning from Debating Europe to X-stance, while Table 6 shows the results of applying transfer learning from X-stance to Debating Europe. The former gave an insight into the effect of pre-training over a non-English multilingual dataset from another domain. The latter gave an insight into the effect of pre-training on English and specialised data from an online debate.

Table 5. Results over X-stance dataset for a binary classification, best result in bold. The M-BERT results came from [33].

	Intra-Target			X-Question			X-Topic			X-Lingual
	DE	FR	Mean	DE	FR	Mean	DE	FR	Mean	IT
M-BERT [33]	76.8	76.6	76.6	68.5	68.4	68.4	68.9	70.9	69.9	70.2
XLM-R	76.3	78.0	77.1	71.5	72.9	72.2	71.2	73.7	72.4	73.0
XLM-R _{ft}	77.3	79.0	78.1	71.5	74.8	73.1	72.2	74.7	73.4	73.9

As can be seen in Tables 5 and 6, the transfer learning approach was efficient for both datasets, even though they had different languages, topics, and targets. Pre-training over Debating Europe allowed for reaching higher results on the X-stance dataset. It is important to note that this worked even if the DE dataset is very small compared to X-

stance. Moreover, it is a way to validate the annotation that has been made by one expert only, without the possibility of calculating an inter-annotator agreement.

4.1.2. Self-Training Setting

The results of the ST setups are presented in Table 6. Analysing the results, we can see that not all settings led to satisfactory results. In order to understand the causes of this failure, we analysed the distribution of the pseudo-labels (see Figure 4), along with the number of pseudo-labels. By analysing the distribution, it is possible to gain an understanding of the weaknesses of each method and to conclude on the reason why our method performed well: it did not overwhelm the gold labels with weak labels and offered a balanced distribution.

Table 6. Results over the Debating Europe dataset for a 3-class classification using ST. k_{max} is the number of examples added at each iteration.

Unsupervised Method	Threshold	k_{max}	Balanced	Model	Prec.	Rec.	F1	Acc
✗	✗	✗	✗	XLM-R	68.6	69.3	68.9	70.1
				XLM-R _{ft}	70.7	69.9	70.2	72.1
thresh-0.99	0.99	✗	✗	XLM-R	68.6	69.8	69.1	70.7
				XLM-R _{ft}	68.9	69.6	69.0	70.9
k-best-2000	✗	2000	✗	XLM-R	67.5	68.3	67.8	69.3
				XLM-R _{ft}	70.4	69.9	69.8	71.9
k-best-600	✗	600	✗	XLM-R	69.4	68.5	68.0	69.5
				XLM-R _{ft}	72.5	70.3	71.1	73.3
our-2000	0.99	2000	✓	XLM-R	69.5	69.4	69.4	71.3
				XLM-R _{ft}	70.5	69.9	69.3	71.7
our-600	0.99	600	✓	XLM-R	70.9	71.6	71.1	72.7
				XLM-R _{ft}	71.5	71.5	71.4	73.5

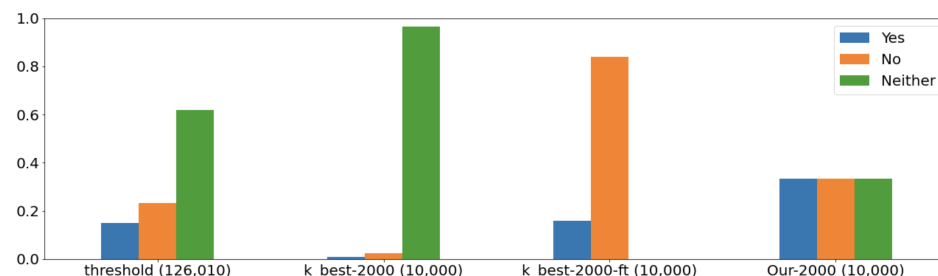


Figure 4. Distribution of the pseudo-labels.

The threshold method did not improve the model’s performance due to the small size of our dataset and the lack of model calibration. Specifically, the non-calibration of the model led to the addition of too many pseudo-examples at each iteration (more than 30-times the number of labels in pseudo-labels in the end), which significantly decreased the model’s performance. On the other hand, the k-best method was able to reduce the number of examples added at each iteration and performed well with the XLM-R_{ft}, as this latter model was trained on a larger number of examples and appeared to be more robust.

4.2. Results on CoFE

4.2.1. Baselines for Scarce Annotation Regimes

We present the results in a cross-datasets setting and without or with manually annotating data from the target dataset for the target task.

The results of the models trained without access to the three-class labelled data can be found in Table 7. We evaluated the performance of our proposed configurations using the F1, macro-F1, and accuracy metrics over the externally annotated dataset CF_{E-T}. The first column shows the different configurations we used, and the next columns show the

annotation used during the training: OODataset means annotations Out-Of-Dataset; CoFE-2 means binary annotations from the CoFE; CoFE-3 means ternary annotations from the CoFE. The following columns show the results for each class (– for the negative class, ~ for the neutral class, and + for the positive class), and the last columns show the accuracy and macro-F1 of the configurations.

Table 7. F1, macro-F1 and accuracy of the different baselines over the externally annotated dataset CF_{E-T} .

Model	Annotations Used			–	~	+	Acc.	M-F1
	CoFE-3	CoFE-2	OODataset					
Hardalov et al. [37] + MT	✗	✗	✓	7.7	29.5	61.4	46.3	32.8
Hardalov et al. [29]	✗	✗	✓	20.7	19.1	58.9	43.2	32.9
Cross-datasets	✗	✗	✓	45.3	44.0	62.6	52.7	50.6
All-1 training	✗	✓	✓	56.8	00.6	77.9	62.9	45.1
Cross-debates	✗	✓	✓	54.3	41.4	77.3	63.0	57.6
All-2 trainings	✗	✓	✓	52.9	45.0	76.3	63.1	58.1
CF_{E-D} -1 training	✓	✓	✗	42.1	39.9	75.6	62.3	52.5
All-1 training	✓	✓	✓	57.9	30.0	78.5	65.4	55.5
All-2 trainings	✓	✓	✓	57.3	40.2	80.5	67.3	59.3

The first section of the table lists models that used only annotations from the Out-Of-Domain (OODataset) dataset. The first two models, Hardalov et al. [37] + MT and Hardalov et al. [29], both use only OODataset annotations from 16 English stance datasets and 10 multilingual stance datasets from various domains and had an accuracy of 46.3 and 43.2, respectively, with a macro-F1 of 32.8 and 32.9, respectively. The third model in this section, cross-datasets, also uses only OODataset annotations from X-stance and Debating Europe and had an accuracy of 52.7 and a macro-F1 of 50.6.

The table's second section lists models that add binary annotations from the target dataset (CoFE-2) in the training set. The first three models in this section, "All-1 training", cross-debates, and "All-2 trainings", use binary annotations from the target dataset. The first configuration, "All-1 training", showed an F1 of 59.7 for the negative class, 00.7 for the neutral class, and 79.5 for the positive class. The accuracy of this configuration was 65.5, and the macro-F1 was 46.6. The second configuration, "cross-datasets", showed an F1 of 54.3 for the negative class, 30.5 for the neutral class, and 73.9 for the positive class. The accuracy of this configuration was 59.6, and the macro-F1 was 52.9. The third configuration, "cross-debates", showed an F1 of 55.3 for the negative class, 40.4 for the neutral class, and 76.6 for the positive class. The accuracy of this configuration was 63.2, and the macro-F1 was 57.4. Finally, the fourth configuration, "All-2 trainings" showed an F1 of 55.4 for the negative class, 44.6 for the neutral class, and 77.3 for the positive class. The accuracy of this configuration was 64.3, and the macro-F1 was 59.1.

The last section of the table lists models that use ternary annotations from the target dataset (CoFE-3). The first two models in this section, CF_{E-D} -1 training and All-1 training, use ternary annotations from the target dataset and had accuracy scores of 62.3 and 65.4, respectively, with macro-F1 scores of 52.5 and 55.5, respectively. The last model in the table, "All-2 trainings", had the best macro-F1 score of 59.3 and the highest accuracy score of 67.3 among all models in the table.

4.2.2. Self-Training Setting

The results of the three-class classification using self-training with the CF_U dataset on the CF_{E-T} dataset are presented in Table 8. The model uses an unlabelled dataset, CF_U , to augment the training data through the ST process. The columns in the table represent the unsupervised method used, the threshold applied during the ST process, the maximum

number of examples (k_{max}) added at each iteration, whether the distribution of pseudo-labels was balanced, and the precision results for the negative (−), neutral (~), and positive (+) classes, as well as the overall accuracy (Acc) and the macro-weighted F1-score (m-F1). The results showed that using a balanced distribution of pseudo-labels led to better performance compared to the models without this balance. Specifically, the best results in terms of the macro-F1 were obtained by the unsupervised method with a threshold of 0.99, a maximum number of examples added of 2000, and a balanced distribution of pseudo-labels. This model achieved a macro-weighted F1-score of 63.2, which was the highest among the models compared.

Table 8. Results of the best model over the CF_{E-T} dataset for a 3-class classification using ST with the unlabelled CF_U dataset. k_{max} is the number of examples added at each iteration.

Unsupervised Method	Threshold	k_{max}	Balanced	−	~	+	Acc	M-F1
\times	\times	\times	\times	57.3	40.2	80.5	67.3	59.3
thresh-0.99	0.99	\times	\times	43.6	55.8	77.3	65.2	58.9
k-best-2000	\times	2000	\times	59.6	42.6	79.9	66.2	60.4
k-best-600	\times	600	\times	51.8	50.4	78.8	66.4	60.3
our-2000	0.99	2000	✓	57.6	52.7	79.2	67.8	63.2
our-600	0.99	600	✓	56.8	51.5	76.4	65.1	61.6

4.3. Analysis of the Results

In this part, we focused our analysis on the experiments using the CoFE data. From the results, we can draw different conclusions regarding the three different parts of Table 7: using only out-of-dataset annotations, using binary annotations from the target dataset, or using ternary annotations from the target dataset. We also discuss the results of the self-training experiments briefly.

Cross-Datasets Data

Our cross-datasets model trained over X-stance and Debating Europe allowed results that were better than two strong cross-datasets baselines. The first baseline is a model trained on English data, using English as the pivot language and machine translation. It gave poor performances on the negative class. The second baseline is a multilingual model, also using the X-stance dataset during its learning phase, making the low results surprising. The gain in performance of our model must come from the training data, which are online debates on political topics.

Binary Labels' Annotations from CoFE

The first conclusion came from the poor performances of the “All-1 training” configuration on the neutral class (0.06): the two-step learning process is mandatory to obtain proper results on the neutral class when tackling ternary stance classification and using only the large binary labelled datasets available. The second conclusion is that it was possible to achieve better results with our method even if we completely dropped the examples from the target dataset (macro-F1 rising to 50.6). Third, the “cross-debates” configuration obtained far better results than the “cross-datasets”; hence, the adaptation towards the domain and languages, which are contained in the target dataset, seems to be important (50.6 to 57.6). Fourth, the results of the “cross-debates” configuration, which is zero-shot regarding the target, were still good compared to the model that had seen examples from the test debates (57.6 vs. 58.1). Finally, we can see that our last proposed configuration, “All-2 trainings”, achieved the best performance, with the highest macro-F1 of 58.1. This suggested that the use of both debates and languages from the target dataset during the training improved the performance of the overall stance classification. Interestingly, it also

improved the performance on the neutral class, even though the labels used during training were only binary.

Ternary Labels' Annotations from CoFE

We can draw two main conclusions from the last section of Table 7. The first one would be that the best results came from the model using the more annotated data ("All-2 trainings" with CoFE-3 reaching the highest macro-F1). The second conclusion came from looking at the performance of CF_{E-D-1} training being a bit higher than the cross-datasets one (52.5 vs. 50.6). This gap between the two results means that, even if costly, annotating data from the target dataset in a ternary way is not enough to reach high performances.

The results of our model on the dataset used by [29,37] can be seen in Appendix C.

Self-Training

All the self-training methods allowed for the improvement of the results, contrary to the experiments on the Debating Europe dataset (Section 3.1.2). The threshold method was the only one that was harmful to the performances. As the model was not calibrated, the first iteration already pseudo-labelled almost all of the unlabelled data (15% of negative, 39% of neutral, and 46% of positive). Hence, the pseudo-labelling only depended on the network trained at the first iteration: all the biases were inserted in the pseudo-labelled data, which overflowed the real training data.

5. Conclusions and Future Work

In this work, we focused on the task of ternary stance recognition, using data from public consultations and digital democracy platforms. We addressed the issue of multi-target stance recognition as defined in [33], where the target can also be expressed like a comment, in natural language. We can point out several contributions. We define the concept of intra-multilinguality, where the target and the comment can come in different languages, by using a platform that automatically translates the textual content so that the users can interact in their native languages. We collected and annotated parts of the dataset presented and made them available online for two shared tasks [59,60]. Finally, we proposed a series of methods to learn with a limited amount of labels, by pre-training over similar datasets and leveraging information from non-annotated data with the help of self-training methods.

Future work in this context will include studying the interactions between the participants of the debates, firstly within the different debates, by studying conversation dynamics [6] in the form of the threads that are available in the CoFE dataset and, secondly, within the platform, by looking at the group of topics each user is interested in to cluster political views at the user level. Another interesting way to study political debates would be to use multimodal content in several forms. Within the CoFE dataset, some descriptions contain multimodal data such as photos or videos, making this integration possible. A step further would be to use virtual video conference meetings to add real-time multimodal content and interactions between the participants to study the dynamics of a real-time debate. Ultimately, an embodied conversational agent [61] could be used as a moderator of the multimodal debates [62]. Finally, it would be interesting to look at the cultural and national biases that we can find in this dataset, by analysing the data separately in a monolingual way both at the semantic and linguistic levels, to understand how these biases influence the quality of the data and the classification performances.

Author Contributions: Conceptualisation—DE, A.B. and V.B.; conceptualisation—CoFE, V.B.; methodology, V.B.; software, V.B.; validation, V.B.; formal analysis, V.B.; investigation, V.B.; writing—original draft preparation, V.B.; writing—review and editing, V.B. and A.B.; visualisation, V.B. All authors have read and agreed to the published version of the manuscript.

Funding: V.B. research was funded by the National Center for Artificial Intelligence CENIA FB210017, Basal ANID.

Data Availability Statement: The CoFE datasets CF_U , CF_S , and CF_{E-D} are already available in the context of the Touché Lab @ CLEF 2023 (<https://touche.webis.de/clef23/touche23-web/multilingual-stance-classification.html> (accessed on 20 February 2023)). The Debating Europe dataset will be available online after publication.

Acknowledgments: We would like to thank Brian Ravenet, Léo Hemamou, and Simon Luck for helping to annotate CF_{E-D} and Guillaume Jacquet for helping in managing the annotation phase performed at the Joint Research Center during the annotation process of a subpart of CF_{E-T} . We thank the Big Data Analytics Platform of the JRC.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ML	Machine Learning
NLP	Natural Language Processing
DA	Data Augmentation
SSL	Self-Supervised Learning
ST	Self-Training
CoFE	Conference on the Future of Europe
DE	Debating Europe

Appendix A. Targets of the Annotated Debates from Debating Europe

The debates chosen for the annotation are the ones below: *Should we consume less energy?*, *Should we make the cities greener?*, *Can renewables ever replace fossil fuels 100?*, *Should we invest more in clean energies to avoid an energy crisis?*, *Should we cut CO2 emission and invest into clean energies?*, *Should we think about the real cost of the food we eat?*, *Should all cars be electric by 2025?*, *Does organic food really make a difference?*, *Should Europeans be encouraged to eat more sustainably?*, *Sustainable agriculture: With or without pesticides?*, *Should all EU countries abandon nuclear power?*, *Should we stop flying to help the environment?*, *Should plastic packaging be banned?*, *Should we all eat less meat?*, *Should we invest in cheap and clean energies?*, *Should we move towards a low-carbon economy or invest into clean energies?*, *Should the European Union ban plastic bags?*, and *Should plastic water bottles be banned?*.

Appendix B. Statistics on Debating Europe Annotated Dataset

Statistics on the Debating Europe annotated dataset can be found in Table A1.

Table A1. Low-level statistics on the Debating Europe dataset. Here, μ represents the average mean, σ the standard deviation, med the median, and Σ the sum.

Aggregation-Level		Debate			Comment			All
Units	Label	μ	σ	Med	μ	σ	Med	Σ
Comments	All	140	99	101	1	0	1	2523
	Yes	56	37	39	1	0	1	1012
	No	29	39	14	1	0	1	489
	Neutral	18	18	11	1	0	1	282
	Not answering	41	23	35	1	0	1	740
Words	All	4683	2721	3794	33	60	16	84,289
	Yes	1933	1221	1772	34	74	13	34,790
	No	942	1157	554	33	43	19	16,012
	Neutral	814	808	478	46	73	23	13,023
	Not answering	1137	627	972	28	39	16	20,464

Appendix C. Results of the Stance Models over Other Datasets

This section contains the results of the cross-datasets model we trained over data related to political topics: pre-trained over X-stance and fine-tuned over Debating Europe. We applied them on the stance datasets used in [37]. We only used the datasets with three or two labels, so we could achieve hard mapping using our model, and we removed the scd dataset, which has no target.

Table A2. Results of our cross-datasets model over binary annotated English datasets from [37].

Model	Perspectrum	Poldeb	Snopes	Argmin	Ibmcs	All
Hardalov et al. [37]	29.6	22.8	29.28	34.16	72.93	37.8
Cross-dataset	63.8	46.3	52.3	61.6	20.3	48.9

Table A3. Results of our cross-datasets model over ternary annotated English datasets from [37].

Model	Iac1	Emergent	Mtsd	Semeval16	Vast	All
Hardalov et al. [37]	35.2	58.49	23.34	37.01	22.89	35.4
Cross-dataset	15.5	21.6	16.7	13.0	29.1	19.2

When analysing the results from the ternary annotated stance datasets, we noticed that our network was struggling to predict things other than the neutral class for all the ternary datasets, leading to very poor results (see Table A3). Nevertheless, in the binary setting, where we discarded the neutral class to keep only the positive and negative, we could obtain higher competitive results (see Table A2). This result is interesting since our network was trained on online debates, but tested on data not only from debates, but also from news (Snopes) or other sources (IBMCS and Argmin).

References

- ALDayel, A.; Magdy, W. Stance detection on social media: State of the art and trends. *Inf. Process. Manag.* **2021**, *58*, 102597. [CrossRef]
- Hardalov, M.; Arora, A.; Nakov, P.; Augenstein, I. A Survey on Stance Detection for Mis- and Disinformation Identification. *arXiv* **2021**, arXiv:2103.00242.
- De Magistris, G.; Russo, S.; Roma, P.; Starczewski, J.T.; Napoli, C. An Explainable Fake News Detector Based on Named Entity Recognition and Stance Classification Applied to COVID-19. *Information* **2022**, *13*, 137. [CrossRef]
- Yang, R.; Ma, J.; Lin, H.; Gao, W. *A Weakly Supervised Propagation Model for Rumor Verification and Stance Detection with Multiple Instance Learning*; Association for Computing Machinery: New York, NY, USA, 2022; Volume 1, pp. 1761–1772. [CrossRef]
- Beauchamp, N. Predicting and Interpolating State-Level Polls Using Twitter Textual Data. *Am. J. Political Sci.* **2017**, *61*, 490–503. [CrossRef]
- Sakketou, F.; Lahnala, A.; Vogel, L.; Flek, L. Investigating User Radicalization: A Novel Dataset for Identifying Fine-Grained Temporal Shifts in Opinion. In Proceedings of the LREC, Marseille, France, 20–25 June 2022; pp. 3798–3808.
- Barriere, V.; Jacquet, G. CoFE: A New Dataset of Intra-Multilingual Multi-target Stance Classification from an Online European Participatory Democracy Platform. In Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, Online, 21–24 November 2022.
- Gupta, A.; Blodgett, S.L.; Gross, J.H.; O'Connor, B. EXPRES: Examining Political Rhetoric with Epistemic Stance Detection. *arXiv* **2022**, arXiv:2212.14486v2.
- Gorrell, G.; Bontcheva, K.; Derczynski, L.; Kochkina, E.; Liakata, M.; Zubiaga, A. RumourEval 2019: Determining rumour veracity and support for rumours. In Proceedings of the SemEval 2019, Minneapolis, MN, USA, 6–7 June 2019; pp. 845–854.
- Matero, M.; Soni, N.; Balasubramanian, N.; Schwartz, H.A. MeLT: Message-Level Transformer with Masked Document Representations as Pre-Training for Stance Detection. In Proceedings of the Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 2959–2966. [CrossRef]
- Mohammad, S.M.; Kiritchenko, S.; Sobhani, P.; Zhu, X.; Cherry, C. A Dataset for Detecting Stance in Tweets. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia, 23–28 May 2016. [CrossRef]

12. Augenstein, I.; Rocktäschel, T.; Vlachos, A.; Bontcheva, K. Stance detection with bidirectional conditional encoding. In Proceedings of the EMNLP 2016—Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 876–885. [\[CrossRef\]](#)
13. Dos Santos, W.R.; Paraboni, I. Moral stance recognition and polarity classification from twitter and elicited text. In Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP, Varna, Bulgaria, 2–4 September 2019; pp. 1069–1075. [\[CrossRef\]](#)
14. Li, Y.; Sosea, T.; Sawant, A.; Nair, A.J.; Inkpen, D.; Caragea, C. P-Stance: A Large Dataset for Stance Detection in Political Domain. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Punta Cana, Dominican Republic, 1–6 August 2021; pp. 2355–2365. [\[CrossRef\]](#)
15. Derczynski, L.; Bontcheva, K.; Liakata, M.; Procter, R.; Hoi, G.W.S.; Zubiaga, A. SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. *arXiv* **2017**, arXiv:1704.05972.
16. Somasundaran, S.; Wiebe, J. Recognizing stances in online debates. In Proceedings of the ACL-IJCNLP 2009—Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing of the AFNLP, Singapore, 2–7 August 2009; pp. 226–234. [\[CrossRef\]](#)
17. Somasundaran, S.; Wiebe, J. Recognizing Stances in Ideological On-Line Debates. In Proceedings of the NAACL Workshop, Los Angeles, CA, USA, 2 June 2010.
18. Walker, M.A.; Anand, P.; Tree, J.E.; Abbott, R.; King, J. A corpus for research on deliberation and debate. In Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, 21–27 May 2012; pp. 812–817.
19. Thomas, M.; Pang, B.; Lee, L. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In Proceedings of the COLING/ACL 2006—EMNLP 2006: 2006 Conference on Empirical Methods in Natural Language Processing, Sydney, Australia, 22–23 July 2006; pp. 327–335.
20. Anand, P.; Walker, M.; Abbott, R.; Tree, J.E.F.; Bowmani, R.; Minor, M. Cats Rule and Dogs Drool!: Classifying Stance in Online Debate. In Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2011), Portland, OR, USA, 24 June 2011; pp. 1–9.
21. Abbott, R.; Walker, M.; Anand, P.; Fox Tree, J.E.; Bowmani, R.; King, J. How can you say such things?!?: Recognizing disagreement in informal political argument. In Proceedings of the Workshop on Languages in Social Media, Portland, OR, USA, 23 June 2011; pp. 2–11.
22. Walker, M.A.; Anand, P.; Abbott, R.; Grant, R. Stance classification using dialogic properties of persuasion. In Proceedings of the NAACL HLT 2012—2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies—Proceedings, Montreal, QC, Canada, 3–8 June 2012; pp. 592–596.
23. Sridhar, D.; Foulds, J.; Huang, B.; Getoor, L.; Walker, M. Joint Models of Disagreement and Stance in Online Debate. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, 27–31 July 2015; pp. 116–125.
24. Barriere, V. Hybrid Models for Opinion Analysis in Speech Interactions. In Proceedings of the ICMI, Glasgow, UK, 13–17 November 2017; pp. 647–651.
25. Allaway, E.; McKeown, K. Zero-Shot Stance Detection: A Dataset and Model Using Generalized Topic Representations. *arXiv* **2020**, arXiv:2010.03640.
26. Villa-Cox, R.; Kumar, S.; Babcock, M.; Carley, K.M. Stance in Replies and Quotes (SRQ): A New Dataset For Learning Stance in Twitter Conversations. In Proceedings of the AACL, New York, NY, USA, 7–12 February 2020.
27. Hazarika, D.; Poria, S.; Zimmermann, R.; Mihalcea, R. Emotion Recognition in Conversations with Transfer Learning from Generative Conversation Modeling. *arXiv* **2019**, arXiv:1910.04980.
28. Lai, M.; Cignarella, A.T.; Hernández Fariás, D.I.; Bosco, C.; Patti, V.; Rosso, P. Multilingual stance detection in social media political debates. *Comput. Speech Lang.* **2020**, *63*, 101075. [\[CrossRef\]](#)
29. Hardalov, M.; Arora, A.; Nakov, P.; Augenstein, I. Few-Shot Cross-Lingual Stance Detection with Sentiment-Based Pre-Training. *arXiv* **2022**, arXiv:2109.06050.
30. Zotova, E.; Agerri, R.; Nuñez, M.; Rigau, G. Multilingual stance detection: The catalonia independence corpus. In Proceedings of the LREC 2020—12th International Conference on Language Resources and Evaluation, Marseille, France, 11–16 May 2020; pp. 1368–1375.
31. Zheng, J.; Baheti, A.; Naous, T.; Xu, W.; Ritter, A. STANCEOSAURUS: Classifying Stance Towards Multicultural Misinformation. In Proceedings of the EMNLP, Abu Dhabi, United Arab Emirates, 7–11 December 2022.
32. Sobhani, P.; Inkpen, D.; Zhu, X. A Dataset for Multi-Target Stance Classification. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain, 3–7 April 2017; Volume 2, pp. 551–557.
33. Vamvas, J.; Sennrich, R. X-stance: A Multilingual Multi-Target Dataset for Stance Detection. In Proceedings of the SwissText, Zurich, Switzerland, 23–25 June 2020.
34. Deng, R.; Panl, L.; Clavel, C. Domain Adaptation for Stance Detection towards Unseen Target on Social Media. In Proceedings of the 2022 10th International Conference on Affective Computing and Intelligent Interaction, ACII 2022, Nara, Japan, 18–21 October 2022. [\[CrossRef\]](#)
35. Hosseini, M.; Dragut, E.; Mukherjee, A. Stance Prediction for Contemporary Issues: Data and Experiments. *arXiv* **2020**, arXiv:2006.00052.

36. Barriere, V.; Jacquet, G. How does a pre-trained transformer integrate contextual keywords? Application to humanitarian computing. In Proceedings of the International ISCRAM Conference, Blacksburg, VA, USA, May 2019 2021; pp. 766–771.
37. Hardalov, M.; Arora, A.; Nakov, P.; Augenstein, I. Cross-Domain Label-Adaptive Stance Detection. In Proceedings of the EMNLP, Virtual, 7–11 November 2021; Volume 19.
38. Augenstein, I.; Ruder, S.; Søgaard, A. Multi-Task learning of pairwise sequence classification tasks over disparate label spaces. In Proceedings of the NAACL HLT 2018—2018 Conference North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA, 1–6 June 2018; Volume 1, pp. 1896–1906.
39. Barriere, V.; Balahur, A.; Ravenet, B. 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, Marseille, France, 20–25 June 2022; European Language Resources Association: Marseille, France, 2022; pp. 16–21.
40. Bai, F.; Ritter, A.; Xu, W. Pre-train or Annotate? Domain Adaptation with a Constrained Budget. In Proceedings of the EMNLP 2021—2021 Conference on Empirical Methods in Natural Language Processing, Virtual, 7–11 November 2021; pp. 5002–5015.
41. Yarowsky, D. Unsupervised word sense disambiguation rivaling supervised methods. In Proceedings of the ACL, Cambridge, MA, USA, 26–30 June 1995; pp. 189–196. [\[CrossRef\]](#)
42. Zhu, X.; Ghahramani, Z. *Learning from Labeled and Unlabelled Data with Label Propagation*; Technical Report; Technical Report CMU-CALD-02-107; Carnegie Mellon University: Pittsburgh, PA, USA, 2002.
43. Zhou, D.; Bousquet, O.; Navin Lal, T.; Weston, J.; Schölkopf, B. Learning with Local and Global Consistency. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, Canada, 8–13 December 2003. [\[CrossRef\]](#)
44. Giasemidis, G.; Kaplis, N.; Agraftiotis, I.; Nurse, J.R. A Semi-Supervised Approach to Message Stance Classification. *IEEE Trans. Knowl. Data Eng.* **2020**, *32*, 1–11. [\[CrossRef\]](#)
45. Glandt, K.; Khanal, S.; Li, Y.; Caragea, D.; Caragea, C. Stance Detection in COVID-19 Tweets. In Proceedings of the ACL-IJCNLP, Virtual, 1–6 August 2021; pp. 1596–1611. [\[CrossRef\]](#)
46. Wei, C.; Sohn, K.; Mellina, C.; Yuille, A.; Yang, F. CRST: A Class-Rebalancing Self-Training Framework for Imbalanced Semi-Supervised Learning. In Proceedings of the CVPR, Nashville, TN, USA, 20–25 June 2021.
47. Klie, J.C.; Bugert, M.; Boullosa, B.; de Castilho, R.E.; Gurevych, I. The INCEPTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In Proceedings of the International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 5–9.
48. Küçük, D.; Fazli, C.A. Stance detection: A survey. *ACM Comput. Surv.* **2020**, *53*, 1–37. [\[CrossRef\]](#)
49. Krippendorff, K. *Content Analysis: An Introduction to Its Methodology*; SAGE Publications: Los Angeles, CA, USA, 2013. [\[CrossRef\]](#)
50. Joseph, K.; Shugars, S.; Gallagher, R.; Green, J.; Mathé, A.Q.; An, Z.; Lazer, D. (Mis)alignment Between Stance Expressed in Social Media Data and Public Opinion Surveys. In Proceedings of the EMNLP 2021—2021 Conference on Empirical Methods in Natural Language Processing, Virtual, 7–11 November 2021; pp. 312–324. [\[CrossRef\]](#)
51. Yin, W.; Hay, J.; Roth, D. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In Proceedings of the EMNLP-IJCNLP 2019, Hong Kong, China, 3–7 November 2019; pp. 3914–3923. [\[CrossRef\]](#)
52. Rosenstein, M.T.; Marx, Z.; Kaelbling, L.P.; Dietterich, T.G. To transfer or not to transfer. In Proceedings of the NIPS 2005 Workshop Transfer Learning, Vancouver, BC, Canada, 5–8 December 2005; Volume 898, p. 3.
53. Ruder, S. Neural Transfer Learning for Natural Language Processing. Ph.D. Thesis, University of Galway, Galway, Ireland, 2019.
54. Barriere, V.; Balahur, A. Improving Sentiment Analysis over non-English Tweets using Multilingual Transformers and Automatic Translation for Data-Augmentation. In Proceedings of the COLING, Barcelona, Spain, 12 December 2020.
55. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. HuggingFace’s Transformers: State-of-the-art Natural Language Processing. *arXiv* **2019**, arXiv:1910.03771.
56. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised Cross-Lingual Representation Learning at Scale. *arXiv* **2020**, arXiv:1911.02116. [\[CrossRef\]](#)
57. Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations, Banff, AB, Canada, 14–16 April 2014; pp. 1–13. <http://arxiv.org/abs/1412.6980>.
58. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. TensorFlow: A system for large-scale machine learning. In Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016, Savannah, GA, USA, 2–4 November 2016; pp. 265–283.
59. Bondarenko, A.; Fröbe, M.; Kiesel, J.; Schlatt, F.; Barriere, V.; Ravenet, B.; Hemamou, L.; Luck, S.; Reimer, J.H.; Stein, B.; et al. Overview of Touché, 2023: Argument and Causal Retrieval. In Proceedings of the ECIR, Dublin, Ireland, 2–6 April 2023.
60. Mirzakhmedova, N.; Kiesel, J.; Alshomary, M.; Heinrich, M.; Handke, N.; Cai, X.; Barriere, V.; Dastgheib, D.; Ghahroodi, O.; Sadraei, M.A.; et al. The Touché23-ValueEval Dataset for Identifying Human Values behind Arguments. *arXiv* **2023**, arXiv:2301.13771.

61. Pelachaud, C. Multimodal Expressive Embodied Conversational Agents. In Proceedings of the 13th annual ACM International Conference on Multimedia, Singapore, 6–11 November 2005; pp. 683–689. [[CrossRef](#)]
62. Argyle, L.P.; Busby, E.; Gubler, J.; Bail, C.; Howe, T.; Rytting, C.; Wingate, D. AI Chat Assistants can Improve Conversations about Divisive Topics. *arXiv* **2023**, arXiv:2302.07268v1.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.