Few-Shot Cross-Lingual Stance Detection with Sentiment-Based Pre-Training

Momchil Hardalov,[1,2] Arnav Arora,[1,3] Preslav Nakov,[1,4] Isabelle Augenstein[1,3]

[1] Checkstep Research
[2] Sofia University "St. Kliment Ohridski", Bulgaria
[3] University of Copenhagen, Denmark
[4] Qatar Computing Research Institute, HBKU, Doha, Qatar
{momchil, arnav, preslav.nakov, isabelle}@checkstep.com

Abstract

The goal of stance detection is to determine the viewpoint expressed in a piece of text towards a target. These viewpoints or contexts are often expressed in many different languages depending on the user and the platform, which can be a local news outlet, a social media platform, a news forum, etc. Most research in stance detection, however, has been limited to working with a single language and on a few limited targets, with little work on cross-lingual stance detection. Moreover, non-English sources of labelled data are often scarce and present additional challenges. Recently, large multilingual language models have substantially improved the performance on many non-English tasks, especially such with limited numbers of examples. This highlights the importance of model pre-training and its ability to learn from few examples. In this paper, we present the most comprehensive study of cross-lingual stance detection to date: we experiment with 15 diverse datasets in 12 languages from 6 language families, and with 6 low-resource evaluation settings each. For our experiments, we build on pattern-exploiting training, proposing the addition of a novel label encoder to simplify the verbalisation procedure. We further propose sentiment-based generation of stance data for pre-training, which shows sizeable improvement of more than 6% F_1 absolute in low-shot settings compared to several strong baselines.

1 Introduction

As online speech gets democratised, we see an ever-growing representation of non-English languages on online platforms. However, in stance detection multilingual resources are scarce (Joshi et al. 2020). While English datasets exist for various domains and in different sizes, non-English and multilingual datasets are often small (under a thousand examples (Lai et al. 2018, 2020; Lozhnikov, Derczynski, and Mazzara 2020)) and focus on narrow, potentially country- or culture-specific topics, such as a referendum (Taulé et al. 2017; Lai et al. 2018), a person (Hercig et al. 2017; Lai et al. 2020), or a notable event (Swami et al. 2018), with few exceptions (Vamvas and Sennrich 2020).

Recently, notable progress was made in zero- and few-shot learning for natural language processing (NLP) using pattern-based training (Brown et al. 2020; Schick and Schütze 2021a; Gao, Fisch, and Chen 2021). These approaches shed light on the ability of pre-trained models to perform in low-resource scenarios, making them an ideal option for modelling cross-lingual stance. Yet, previous work mostly focused on single-task and single-language scenarios. In contrast, here we study their multilingual performance, and their ability to transfer knowledge across tasks and datasets. Moreover, a limitation of these models, especially for pattern-exploiting training (Schick and Schütze 2021a), is the need for label verbalisation, i.e., to identify single words describing labels. This can be inconvenient for label-rich and nuanced tasks such as stance detection. We overcome this limitation by introducing a label encoder.

Another line of research is transfer learning from different tasks and domains. Recent studies have shown that multi-task and multi-dataset learning can increase both the accuracy and the robustness of stance detection models (Schiller, Daxenberger, and Gurevych 2021; Hardalov et al. 2021). Nonetheless, pre-training should not necessarily be performed on the same task; in fact, it is important to select the auxiliary task to pre-train on carefully (Poth et al. 2021). Additional or auxiliary data, albeit from a similar task, can also improve performance. An appealing candidate for stance detection is sentiment analysis, due to its semantic relationship to stance (Ebrahimi, Dou, and Lowd 2016; Sobhani, Mohammad, and Kiritchenko 2016).

Our work makes the following contributions:

- We present the largest study of cross-lingual stance detection, covering 15 datasets in 12 diverse languages from 6 language families.
- We explore the capabilities of pattern training both in a few-shot and in a full-resource cross-lingual setting.
- We introduce a novel label encoding mechanism to overcome the limitations of predicting multi-token labels and the need for verbalisation (single-token labels).
- We diverge from stance-to-stance transfer by proposing a novel semi-supervised approach to produce automatically labelled instances with a trained sentiment model, leading to sizeable improvements over strong baselines.
- We show that our newly introduced semi-supervised approach outperforms models fine-tuned on few shots from multiple cross-lingual datasets, while being competitive with pre-trained models on English stance datasets.

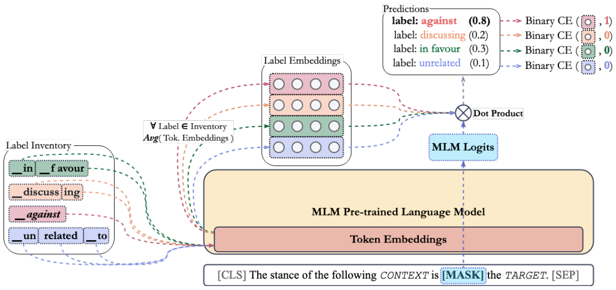

Figure 1: The architecture of the proposed method and the prompt used for prediction. The *CONTEXT* and the *TARGET* are replaced with the corresponding ones for each example. The label inventory comes from the training dataset.

2 Method

We propose an end-to-end few-shot learning, and a novel noisy sentiment-based stance detection pre-training.

2.1 Few-Shot Pattern-Exploiting Learning (PET)

PET and its variants (Schick and Schütze 2021a,b; Tam et al. 2021) have shown promising results when trained in a few-shot setting. They bridge the gap between downstream tasks like text classification and the pre-training of models by converting the dataset into a cloze-style question format that brings it closer to the masked language modelling objective. Using this technique, models with few hundred million parameters can outperform parameter-rich models such as GPT-3 (Brown et al. 2020) on various benchmark tasks (Wang et al. 2018) by fine-tuning on just 32 examples.[1] Our motivation for adopting this framework is threefold: *(i)* there has not been much prior work that puts these models under scrutiny in a cross-lingual setting, *(ii)* often, there is a data scarcity for many languages, which is also the case with stance datasets (only three of our datasets contain more than 2,000 training examples, see Section 3), *(iii)* the label inventories of different datasets are often shared or contain synonymous words such as *'pro', 'in favour', 'support'*, etc., which can be strong indicators for the model in both few-shot or full-resource setting.

2.2 Cross-lingual Stance Pattern Training

Figure 1 shows the architecture of our model. First, we use a simplified PET with a pre-trained language model to predict the likelihood of each label to fill a special mask token in a sentence-based template (see *Prompt* below). To obtain a suitable representation (*label embeddings*) for the labels, we use a *'label encoder'* that averages the pooled vectors from the model's token embeddings for each sub-word. Finally, we take the dot product of the *label embeddings* and the contextualised word embedding for the masked position to obtain the likelihood for each label to fit in.

[1]This comparison is not entirely fair as the GPT model uses priming and is not fine-tuned on any task-specific data.

Prompt The prompt design is an important aspect of the pattern-exploiting training procedure. In our work, we select a prompt that describes the stance task, rather than a punctuation-based one used in previous work (Schick and Schütze 2021a). In particular, our prompt is shown below, where the special token changes based on the model choice:

[CLS] *The stance of the following* ____**CONTEXT** *is* **[MASK]** *the* ____**TARGET**.[SEP]

Prior work (Qin and Eisner 2021; Logan IV et al. 2021; Lester, Al-Rfou, and Constant 2021) has studied aspects of PET such as prompt design, tuning, and selection. Here, we focus on the training procedure, and we leave the exploration of these in a multilingual setting for future work.

Label Encoder A well-known challenge in PET is the need for a fixed number of positions for the label, e.g., a single mask is needed for words present in the dictionary such as *'Yes/No'*; however, we need multiple positions to predict more complex ones with multiple tokens such as *'Unrelated'*. Moreover, if different labels have different lengths, the model needs to ignore some of the positions, e.g., to predict a padding inside the sentence. The label inventory commonly contains words tokenised into multiple tokens. Schick and Schütze (2021a) propose a simple verbalisation technique where the original labels are replaced with words that can be represented with a single token from the vocabulary, e.g., *'Favour'* → *'Yes'*, *'Against'* → *'No'*. Another possibility is to automatically detect such words, but this yields notable drop in performance compared to manual verbalisation by a domain expert (Schick, Schmid, and Schütze 2020).

Here, we propose a simple, yet effective, approach to overcome this problem. Instead of using a single token representation per label, we take the original label inventory and we tokenise all words, as shown in Figure 1. In the *'Label inventory'* box, we see four labels common for stance tasks and their tokens (obtained by the XLM-R's tokenizer) – {*'_against'*}, {*'_discuss', 'ing'*}, {*'_in', '_favour'*}, and {*'_un', 'related', '_to'*}. For each token of a label, we extract the vector representation from the MLM pre-trained model's (e.g., XLM-R) token embeddings $v_{TE}^{L_t} = TokEmb(L_t)$. Afterwards, we obtain the final label representation (LE_L) using an element-wise averaging for all $v_{TE}^{L_t}$ (see Eq. 1).

$$LE_L = \frac{1}{N}\sum_{t=0}^{N} TokEmb(L_t); \forall_L \in \{Labels\} \quad (1)$$

Note that for single tokens, this method defaults to the original MLM task used in learning BERT-based models (Devlin et al. 2019; Liu et al. 2019). The technique of averaging the embedding is shown to be effective with non-contextualised language models such as word2vec (Mikolov et al. 2013) and GloVe (Pennington, Socher, and Manning 2014) for representing entire documents or for obtaining a token-level representation with fastText (Joulin et al. 2017).

Finally, to obtain the label for each example, we take the dot product between the MLM representation for the masked token position, and each of the LE_L vectors. There is no need for padding, as both representations are of the same dimensionality by design (Conneau et al. 2020). Here, we must note that we select the candidates only from the task-related labels; however, we treat the task as a multi-label one, as we describe in more detail below.

Training Objective We use a standard binary-cross entropy (BCE) loss for each label, where for positive examples, we propagate 1, and for negative ones, we propagate 0. We do not use the original MLM cross-entropy over the entire dictionary, as this will force the model to recognise only certain words as the correct labels, whereas their synonyms are also a valid choice. Moreover, such a loss will prevent further knowledge transfer between tasks and will degraded the model's ability to perform in a zero-shot setting.

$$\mathcal{L}_{LE} = \sum_{y' \in y^p} \mathrm{BCE}(p(y'|x), 1) + \sum_{y'' \in y^n} \mathrm{BCE}(p(y''|x), 0) \tag{2}$$

$$\mathcal{L} = \lambda \cdot \mathcal{L}_{LE} + (1 - \lambda) \cdot \mathcal{L}_{MLM} \tag{3}$$

Positive and Negative Sampling The label encoder allows for sampling of positive and negative examples at training. This can be useful for tasks such as stance detection, where label inventories can differ, but labels overlap semantically. Indeed, this holds for our datasets, as is apparent in Table 1 where we see semantically similar labels like *support*, *agree*, *favor* etc. across several datasets.

To obtain a set of synonyms for each label, we use two publicly available sources: *(i)* Google Dictionary suggestions[2]; and *(ii)* synsets of the English WordNet (Miller 1998). However, this is prone to noise, as a word can have multiple meanings, and building a high-quality lexicon would require a human annotator proficient in the target language. Thus, we use negative sampling, as unrelated words are also undesirable to predict by the model, rather than using these examples to enrich the positive labels lexicon.

2.3 Sentiment-Based Stance Pre-Training

We propose a novel semi-supervised method for pre-training stance detection models using annotations from a sentiment analysis model. This is motivated by the observation that these are two closely related tasks (the difference being that sentiment analysis does not have a target).[3] To illustrate this, consider the sentence *'I am so happy that Donald Trump lost the election.'*, which has a *positive* sentiment, but when expressed towards a specific target, e.g., *'Donald Trump'*, then the expected label should be the opposite – *negative*, or more precisely *against*. This requires for the introduction of targets that can change the sentiment label. For further details how we produce corresponding datasets see Section 3.3.

We hypothesise that such pre-training could help bootstrap the model's performance, especially in a low-resource setting, similarly to pre-training on cross-domain stance datasets. We use the same model and pattern as for fine-tuning the cross-lingual stance models, and we use a masked language modelling objective and negative sampling to improve the language model's performance on one hand, and, on the other hand, to allow the model to associate synonyms as the label inventories are very diverse (see Table 1). We do not do positive sampling as it requires high-quality synonyms, which can only be obtained by manual annotations, while our goal is to design an end-to-end pipeline without a need for human interaction.

[2]https://github.com/meetDeveloper/freeDictionaryAPI

[3]Note that we consider the basic, untargeted variant of sentiment analysis here, as more resources exist for it.

3 Datasets

We use three types of datasets: 15 cross-lingual stance datasets (see Table 1), English stance datasets, and raw Wikipedia data automatically annotated for stance. The cross-lingual ones are used for fine-tuning and evaluation, whereas the rest are only used for pre-training. Appendix B provides additional examples for the cross-lingual datasets shown in Table 7. Further quantitative analysis of the texts are also shown in the Appendix in Table 5 and Figure 2.

3.1 Cross-Lingual Stance Datasets

ans (Khouja 2020). The Arabic News Stance corpus has paraphrased or contradicting news titles from several major news sources in the Middle East.

arabicfc (Baly et al. 2018) consists of claim-document pairs with true and false claims extracted from a news outlet and a fact checking website respectively. Topics include the *Syrian War* and other related Middle Eastern issues.

conref-ita (Lai et al. 2018) contains tweet-retweet-reply triplets along with their stance annotation pertaining to a polarising *referendum* held in Italy in December 2016 to amend the constitution.

czech (Hercig et al. 2017) provides stance-annotated comments on a news server in Czech on a proposed *Smoking ban in restaurants* and the Czech president *Miloš Zeman*.

dast (Lillie, Middelboe, and Derczynski 2019) includes stance annotations towards submissions on Danish subreddits covering various political topics.

e-fra, r-ita (Lai et al. 2020) consists of French tweets about the 2017 French presidential election and Italian ones about the 2016 Italian constitutional referendum.

hindi (Swami et al. 2018) has Hindi-English code-mixed tweets and their stance towards *demonetisation* of the Indian currency that took place in 2016.

ibereval (Taulé et al. 2017) contains tweets in Spanish and Catalan about the *Independence of Catalonia*, collected as part of a shared task held at IberEval 2017.

nlpcc (Xu et al. 2016) contains posts from the Chinese micro-blogging site Sina Weibo about manually selected topics like the *iPhoneSE* or the *open second child policy*.

rustance (Lozhnikov, Derczynski, and Mazzara 2020) includes posts on Twitter and Russian-focused media outlets on topics relating to Russian politics. The extraction was done in 2017.

sardistance (Cignarella et al. 2020) includes textual and contextual information about tweets relating to the Sardines movementin Italy towards the end of 2019.

xstance (Vamvas and Sennrich 2020) contains questions about topics relating to Swiss politics, answered by Swiss political candidates in French, Swiss German, or Italian, during elections held between 2011 and 2020.

	Dataset	Language	Target	Context	#Targets	#Contexts	Labels
1	ans	Arabic	Headline	Headline	2,749	2,857	agree (34%), disagree (63%), other (2%)
2	arabicfc	Arabic	Claim	Article	421	2,897	unrelated (68%), agree (16%), discuss (13%), disagree (3%)
3	conref-ita	Italian	Tweet	Tweet	947	963	against (70%), favor (17%), none (12%)
4	czech	Czech	Smoke ban, Milos Zeman	Comment	2	1,455	against (29%), in favor (24%), none (48%)
5	dast	Danish	Claim or Topic	Post	33	2,997	commenting (78%), denying (10%), querying (3%), supporting (9%)
6	e-fra	French	Emmanuel Macron, Marine Le Pen	Tweet	2	1,112	against (69%), favour (14%), none (17%)
7	hindi*	Hindi-En	Notebandi	Tweet	1	3,545	none (55%), favor (27%), against (18%)
8	ibereval-ca	Catalan	Independència de Catalunya	Tweet	1	4,319	favor (61%), neutral (36%), against (3%)
9	ibereval-es	Spanish	Independencia de Cataluña	Tweet	1	4,319	neutral (59%), against (33%), favor (8%)
10	nlpcc‡	Chinese	Two Children, Firecrackers, IphoneSE, Russia in Syria, Motorcycles ban	Post	5	2,966	against (40%), favor (39%), none (20%)
11	r-ita	Italian	Referendum costituzionale	Tweet	1	833	against (58%), none (22%), favor (20%)
12	rustance	Russian	Claim or Tweet	Comment	17	956	comment (69%), query (20%), support (6%), deny (5%)
13	sardistance	Italian	Movimento delle sardine	Tweet	1	3,242	against (55%), favor (24%), none (21%)
14	xstance-de	German	Question	Answer	173	46,723	against (50%), favor (50%)
15	xstance-fr	French	Question	Answer	178	16,309	favor (53%), against (47%)

Table 1: The cross-lingual datasets included in our work and their characteristics. If a dataset contains a small number of targets, then we list them, as they are in the dataset. ‡The targets of the *nlpcc* are in Chinese, except *IphoneSE*, we show the respective English translations. *The texts in the *hindi* dataset are code-mixed (Hindi-English).

3.2 English Stance Datasets

We use 16 English stance datasets from two recent large-scale studies of multi-task/multi-dataset stance detection (Schiller, Daxenberger, and Gurevych 2021; Hardalov et al. 2021). We followed the data preparation and the data pre-processing described in the aforementioned papers as is. The combined dataset contains more than 250K examples, 154K of which are used for training. The data comes from social media, news websites, debating forums, political debates, encyclopedias, and Web search engines, etc. The label inventory includes 24 unique labels. We refer the interested reader to the respective papers for further detail.

3.3 Sentiment-Based Stance Datasets

We use Wikipedia as a source of candidate examples for constructing our sentiment-based stance dataset due to its size and diversity of topics covered. To study the impact language has on pre-training, we construct two datasets: English (*enWiki*) and multilingual (*mWiki*). The latter includes examples from each of the languages covered by some of our datasets. In particular, we use the Wikipedia Python API to sample random Wiki articles. For the multilingual setup, for each language, we sampled 1,000 unique articles[4] (non-overlapping among languages), a total of 11,000. For the English-only setup, we sampled the same number of articles.

Next, to obtain the contexts for the datasets, we split the articles (with headings removed) at the sentence-level using a language-specific sentence splitting model from Stanza (Qi et al. 2020). Each context is then annotated with sentiment using XLM-T, an XLM-R-based sentiment model trained on Twitter data (Barbieri, Anke, and Camacho-Collados 2021). We use that model as it covers all the datasets' languages, albeit from a different domain. It produces three labels – {*positive, negative, neutral*}, which we rename to {*favor, against, discuss*}, to match the label inventory common for stance tasks. To obtain a target–context pair, we assign a target for each context – either the 'Title' of the article, or, if there is a subheading, the concatenation of the two. To cover as much as possible of the stance label variety, we also include *unrelated* in the inventory, which we define as *'a piece of text unrelated to the target'*: for this, we randomly match targets and contexts from the existing tuples. The latter class also serves as a regulariser, preventing overfitting to the sentiment analysis task, as it includes examples with positive or negative contexts that are not classified as such. The resulting distribution is unrelated (60%), discuss (23%), against (10%), favor (7%). This aims to match the class imbalance common for stance tasks (Pomerleau and Rao 2017; Baly et al. 2018; Lozhnikov, Derczynski, and Mazzara 2020).

Finally, we augment 50% of the examples by replacing the target (title) with the first sentence from the abstract of the Wikipedia page. These new examples are added to the original dataset, keeping both the original and the augmented ones. Our aim is to also produce long examples such as user posts, descriptions of an events, etc., which are common targets for stance. The resulting dataset contains around 300K examples, which we split into 80% for training and 10% for development and testing each, ensuring that sentences from one article are only included in one of the data splits.

4 Experiments

Models We evaluate three groups of models: (*i*) without any pre-training, i.e., baselines (see next); (*ii*) pre-trained

[4] We did not include articles in Hindi, as the *hindi* corpus contains texts in Latin, whereas the Wiki articles are in Devanagari.

on multiple English stance datasets (*‘enstance‘*), using automatically labelled instances produced using a sentiment model (*‘*Wiki‘*), see Section 2.3; and (*iii*) multi-dataset learning (MDL), i.e., we include N examples from each dataset into the training data. We train and evaluate on a single dataset, except in the case of MDL, where we train and evaluate on everything. We choose the best model based on the macro-averaged F1 on all datasets. All models use XLM-R_{Base} as their base.

Baselines In addition to our proposed models (Section 2), we compare to a number of simple baselines:

Majority class baseline calculated from the distributions of the labels in each test set.

Random baseline Each test instance is assigned a target label at random with equal probability.

Logistic Regression A logistic regression trained using TF.IDF word unigrams. The input is the concatenation of separately produced vectors for the target and the context.

XLM-R A conventionally fine-tuned XLM-R_{Base} model predicting and back-propagating the errors though the special $<s>$ token.

4.1 Quantitative Analysis

We first analyse the high-level few-shot performance of the proposed models using the averaged per-dataset F1 macro. Then, we zoom in on the dataset level and analyse the models in the two most extreme training scenarios: few-shot with 32 examples, and full-resource training.

Few-shot analysis Table 2 shows results for different types of pre-training on top of the pattern-based model (Section 2). The top of the table lists baselines, followed by ablations of training techniques. More fine-grained, performance per dataset is shown in Table 3 in Appendix C. We can see that the *‘Pattern‘* model outperforms random baselines in all shots, except zero. Moreover, there is a steady increase in performance when adding more examples. The performance saturates at around *256* examples, with the difference between it and *all* being 1.3 points F1, whereas in subsequent pairs from previous columns the margin is 3.5 to 5 points.

The middle part of Table 2 ablates the stance pre-training on top of the pattern-based model. We first analyse the models trained using the artificial dataset from Wikipedia articles, automatically labelled with a multilingual sentiment model (Section 2.3). We study the effects of the language of the pre-training data by including two setups – *enWiki* that contains only English data, and *mWiki* with equally distributed data among all languages seen in the datasets. Both variants give a sizeable improvement over the baselines in all few-shot settings, especially in low-resource ones. The increase in F1 when using *32* examples is more than 6 points on average; and these positive effects are retained when training on *all* examples. The *mWiki* model outperforms the *Pattern* baseline by 4 points and the *enWiki* – 1.4 points respectively. The multilingual pre-trained model constantly scores higher than the English pre-trained one. Moreover, we see a tendency for the gap between the two to increase with the number of examples reaching 2.6 points in *all*.

Model	Shots 0	32	64	128	256	all
Majority	25.30					
Random	30.26					
Pattern	18.25	39.17	43.79	47.16	52.15	53.43
Pattern + Pre-training						
enWiki	28.99	45.09	47.96	50.19	53.85	54.82
mWiki	28.56	45.88	48.59	51.42	<u>54.38</u>	57.40
enstance	**35.16**	**50.38**	**52.69**	**54.75**	**57.87**	61.31
Multi-dataset learning						
MDL Pattern	-	40.76	43.25	48.06	50.36	<u>61.81</u>
MDL mWiki	-	<u>47.16</u>	<u>49.82</u>	<u>51.98</u>	<u>54.33</u>	**62.25**

Table 2: Few-shot macro-average F1. The random and the majority class baselines use no training, and are constant. *en/mWiki* is a pre-trained on our sentiment-based stance task using English or Multilingual data. *enstance* is pre-trained on all English stance datasets. Multi-dataset learning (*MDL*): we train on K examples from each dataset.

For pre-training on English stance data (*enstance*), even with 32 examples, we see a large increase in performance of 11 points absolute over the *Pattern* baseline. This model is also competitive, within 3 points absolute on average, to the baseline trained on the whole dataset. Furthermore, the *enstance* model outperforms the pre-training with automatically labelled stance instances (*en/mWiki*). Nevertheless, the *en/mWiki* models stay within 3-5 points F1 in the all shots. The gap in performance is expected, as the *enstance* model is exposed to multiple stance definitions during its extensive pre-training, in contrast to the single one in the *Wiki* and its noisy labels. Finally, only *enstance* surpasses the random baselines even in the zero-shot setting, scoring 35.16 F1, demonstrating the difficulty of this task. We offer additional analyses of the zero-shot performance in Appendix C.3.

The bottom part of Table 2 shows results for multi-dataset learning (*MDL*). Here, the models are trained on N examples from each dataset, instead of N from a single dataset. The first row presents the *MDL Pattern* model (without any pre-training). Here, we can see that in few-shot setting training on multiple datasets does not bring a significant performance gains compared to using examples from a single dataset. Nonetheless, when all the data is used for training, F1 notably increases, outperforming the English stance model. Furthermore, combining MDL with the multilingual sentiment-based stance pre-training (*MDL mWiki*) yields an even larger increase – almost 9 points F1 higher than *Pattern*, 5 points better than *mWiki*, and 1 point – *enstance*. We attribute the weaker performance on few-shot and the strong performance on full-resource learning of the *MDL*-based models to the diversity of the stance definitions and domains of the datasets, i.e., *MDL* fails to generalise and overfits the training data samples in the few-shot setting, however when more data is included, it serves as a regularizer, thus the model's score improves. The phenomena is also seen in other studies on English stance (Schiller, Daxenberger, and Gurevych 2021; Hardalov et al. 2021). To some extent, the same regularisation effect comes from the pre-

Model	ans	arafc	con-ita	czech	dast	e-fra	hindi	iber-ca	iber-es	nlpcc	r-ita	rusta.	sardi.	xsta-de	xsta-fr	$\mathbf{F1_{avg}}$
Majority	26.0	20.4	27.5	22.1	21.9	28.9	23.6	25.4	24.6	19.3	24.5	19.8	26.7	33.6	35.2	25.3
Random	24.9	20.3	26.7	33.4	17.5	25.0	32.4	28.9	31.0	32.3	31.4	20.9	28.8	50.1	50.2	30.3
Logistic Reg.	31.0	32.7	31.0	29.2	21.9	33.8	33.7	45.8	39.3	29.4	60.9	24.5	32.2	62.8	64.9	38.2
XLM-$\mathrm{R_{Base}}$	83.2	35.7	42.3	54.7	26.2	33.0	29.3	65.9	54.2	58.2	87.6	19.8	49.9	73.2	72.7	52.4
Full-resource training																
Pattern	84.1	39.6	34.1	48.1	34.8	34.3	43.0	67.0	56.5	51.4	79.5	32.1	49.7	73.1	74.1	53.4
enWiki	86.9	38.5	42.8	50.8	25.5	48.9	45.5	65.3	57.0	51.4	88.3	22.4	50.7	73.7	74.7	54.8
mWiki	83.0	40.5	63.0	55.1	32.1	49.8	45.4	68.6	57.5	54.7	93.5	32.8	**52.5**	64.8	67.7	57.4
enstance	**89.0**	**46.5**	59.6	53.1	**41.5**	**54.5**	46.9	66.3	58.8	**58.7**	93.0	50.0	52.0	**74.8**	74.9	61.3
MDL	84.7	44.8	**71.7**	54.1	38.2	48.9	47.5	**70.5**	62.1	57.3	94.1	**53.0**	50.3	73.9	**76.1**	61.8
MDL mWiki	82.9	42.7	**71.8**	**56.8**	40.8	49.5	**48.9**	**70.5**	**64.0**	58.3	**96.5**	51.5	49.9	73.9	75.9	**62.3**
Few-shot (32) training																
Pattern	38.1	26.5	31.6	43.4	25.5	40.1	35.4	39.6	35.8	37.2	54.2	44.1	37.1	47.4	51.6	39.2
enWiki	39.6	33.8	46.8	44.1	27.7	47.8	$\underline{39.8}$	46.7	39.4	45.2	75.9	31.8	41.6	58.2	58.1	45.1
mWiki	45.4	32.5	46.9	46.1	26.5	50.5	39.2	42.3	40.0	$\underline{47.3}$	80.9	31.9	43.4	57.5	57.7	45.9
enstance	$\underline{68.3}$	$\underline{39.4}$	48.7	$\underline{47.3}$	27.0	$\underline{54.9}$	38.0	44.3	$\underline{40.7}$	46.8	82.1	$\underline{49.3}$	$\underline{45.2}$	$\underline{59.1}$	$\underline{64.6}$	$\underline{50.4}$
MDL	43.7	28.4	39.8	37.8	$\underline{28.3}$	38.7	37.7	37.5	38.6	38.9	68.0	40.9	33.8	47.2	52.1	40.8
MDL mWiki	47.3	31.8	$\underline{58.5}$	44.1	27.5	47.5	$\underline{39.8}$	$\underline{48.0}$	39.1	46.3	$\underline{82.8}$	35.1	44.2	57.3	58.0	47.2

Table 3: Per-dataset results with *pre-training*. In multi-dataset learning (MDL), the model is trained on N examples per dataset.

training on the artificial stance task, then the model needs to adjust its weights to the new definition, without having to learn the generic stance task from scratch.

Per-dataset analysis Next, we analyse our experiments on the dataset level. In Table 3 we present a fine-grained evaluation for each dataset covering the two most extreme data regimes that we run our models in: (*i*) full-resource training and (*ii*) few-shot training with 32 examples. We want to emphasise that we do not include state-of-the-art (SOTA) results in Table 3 as the setup in most previous work differs from ours, e.g., the data splits do not match (see Appendix B.1), or the use different metrics, etc. For more details about the SOTA refer to the Appendix C.1. For completeness, we include two standard strong baseline models, i.e., Logistic Regression and a conventionally fine-tuned XLM-R. Both baselines are trained on every dataset separately using all of the data available in its training set.

From our results it is clear that even with all data available from training, a model that does not do any pre-training or knowledge transfer such as the *Logistic Regression* struggles with the cross-lingual stance detection tasks. Even though the model surpasses the random baselines, it falls over 14 points F1 short compared to both the XLM-$\mathrm{R_{Base}}$ and the *Pattern* model. In turn, the *Pattern* model is 1 point better than the XLM-$\mathrm{R_{Base}}$ outperforming the random baselines on all datasets. Interestingly, the XLM-$\mathrm{R_{Base}}$ model fails to beat the random baselines on *hindi* and *rustance*. We attribute this to the code-mixed nature of the former, and the small number of training examples (359) in the latter.

To further understand the results of the models bootstrapped with pre-training or multi-dataset learning, we analyse their per-dataset performance next. From Table 3 we can see that the *MDL* variants achieve the highest results on 8 out of the 15 datasets, *enstance* rank best on 6 and a single win is for *mWiki* on *sardistance*.

Examining the results achieved by the sentiment-based stance pre-training (*en/mWiki*) we see between 7 and 29 points absolute increase in terms of F1 over the *Pattern* baseline for several datasets – *czech, conref-ita, e-fra* and *r-ita*. A contradiction example are the two datasets *dast* and *rustance*, where we have a notable drop in F1 compared both to the *Pattern* and the *enstance* models. On one hand this can be attributed on the skewed label distribution, especially in the *support(ing)*, *deny* and *querying* classes, on the other hand that also suggests the stance definition in these two datasets is different than the one adopted by us in the *en/mWiki* pre-training. In turn, *enstance* demonstrates a robust performance on all datasets, as it has been pre-trained on a variety of stance detection tasks.

A common characteristic uniting the datasets, where the *MDL* models achieve the highest F1, is the presence of at least one other dataset with similar topic and language: (*i*) *conref-ita* and *r-ita* are both Italian datasets about a referendum, (*ii*) *ibereval* contains tweets about the "Independence of Catalonia" in Catalan and Spanish, and (*iii*) *xstance* contains comments by candidates for elections in Switzerland. This suggests that multi-dataset learning is most beneficial when we have similar datasets.

Finally, we analyse the few-shot training with 32 examples. Here, the highest scoring model on 9 out 15 is the *enstance* one. This suggests that other models struggle to learn the stance definition from the cross-lingual datasets by learning from just 32 examples. This phenomena is particularly noticeable in datasets having a skewed label distribution with one or more of the classes being a small proportion of the dataset such as the two Arabic datasets (*ans* – other (2%), *arabicfc* – disagree (3%)). Nevertheless, *en/mWiki* models show steady sizeable improvements of 6 points F1 on average on all dataset. On the other hand, as in the full-resource setting, training on multiple datasets (*MDL mWiki*) boosts the performance of *conref-ita* and *r-ita* with 27 points F1 compared to the *Pattern* baseline. However, we must note that this holds true only when we pre-training on a stance task, as the *MDL* model has lower F1. That again is an argument in favour of our hypothesises: (*i*) few-shot training on multiple stance datasets fails to generalise, and (*ii*) combining datasets that cover the same topic and are in the language have the largest impact on the model's score.

5 Discussion

Our fine-tuning with few instances improves over random and non-neural baselines such as Logistic Regression trained on all-shots, even by more than 20 points F1 on average when training on just 32 instances. However, such models, especially when trained on very few examples, suffer from large variance and instability. In particular, for cross-lingual stance, the pattern-based model's standard deviation (σ) varies from 1.1 (*conref-ita, nlpcc*) to 8.9 (*ibereval-ca*), with 3.5 on average when trained on 32 examples. Pre-training improves stability by reducing the variance, e.g., *en/mWiki* have a σ 2.7 with minimum under 1, which is more than 5% relative change even when comparing to the highest F1 average achieved with 32 examples. The lowest σ is when the model is trained on all-shots, and especially in the *MDL* models with 1.7, and 1.4 for the *mWiki* variant.

This variability can be attributed to the known instabilities of large pre-trained language models (Mosbach, Andriushchenko, and Klakow 2021), but this does not explain it all. Choosing a right set of data points is another extremely important factor in few-shot learning that calls for better selection of training data for pre-training and fine-tuning (Axelrod, He, and Gao 2011; Ruder and Plank 2017).

Another important factor is the inconsistency of the tasks in the training data. This is visible from our *MDL* experiments, i.e., the tasks use a variety of definitions and labels. Even with more examples in the training set in comparison to single-task training (15xN examples), the models tend to overfit and struggle to generalise to the testing data. In turn, when sufficient resources are available, *MDL* yields sizeable improvements even without additional pre-training.

Having access to noisy sentiment-based stance data in the same languages helps, but transferring knowledge from a resource-rich language (e.g., English) on the same task (or set of task definitions) is even more beneficial, in contrast to the data's (see Section 4.1) and label's language (see Appendix C.5). Moreover, when using noisy labels from an external model, there is always a risk of introducing additional bias due the training data and discrepancies in the task definition (Waseem et al. 2021; Bender et al. 2021). We observed this for both the *dast* and the *rustance* datasets.

6 Related Work

Stance Detection Recent studies on stance detection have shown that mixing cross-domain English data improves accuracy and robustness (Schiller, Daxenberger, and Gurevych 2021; Hardalov et al. 2021). They also indicated important challenges of cross-domain setups such as differences in stance definitions, annotation guidelines, and label inventories. Our *cross-lingual* setup adds two more challenges: *(i)* data scarcity in the target language, which requires learning from few examples, and *(ii)* need for better multilingual models with the ability for cross-lingual knowledge transfer.

Cross-Lingual Stance Detection There have been many efforts to develop multilingual stance systems (Zotova et al. 2020; Taulé et al. 2017; Vamvas and Sennrich 2020; Agerri et al. 2021). However, most of them consider 2–3 languages, often from the same language family, thus only providing limited evidence for the potential of cross-lingual stance models to generalise across languages. A notable exception is Lai et al. (2020), who work with 5 languages, but restrict their study to a single family of non-English languages and their domain to political topics only. Our work, on the other hand, spans 6 language families and multiple domains from news (Khouja 2020) to finance (Vamvas and Sennrich 2020).

Stance and Sentiment Sentiment Analysis has a long history of association with stance (Somasundaran, Ruppenhofer, and Wiebe 2007; Somasundaran and Wiebe 2010). Sentiment is often annotated in parallel to stance (Mohammad et al. 2016; Hercig, Krejzl, and Král 2018) and has been used extensively as a feature (Ebrahimi, Dou, and Lowd 2016; Sobhani, Mohammad, and Kiritchenko 2016; Sun et al. 2018) or as an auxiliary task (Li and Caragea 2019; Sun et al. 2019) for improving stance detection. Missing from these studies, however, is leveraging sentiment annotations to generate noisy stance examples, which we explore here: for English and in a multilingual setting.

Pattern-based Training Recently, prompt or pattern-based training has emerged as an effective way of exploiting pre-trained language models for different tasks in few-shot settings (Petroni et al. 2019; Schick and Schütze 2021a). Brown et al. (2020) introduced a large language model (i.e., GPT-3), which showed strong performance on several tasks through demonstrations of the task. Schick and Schütze (2021a,b) proposed Pattern-Exploiting Training (PET), a novel approach using comparatively smaller masked language models through Cloze-style probing with task informed patterns. Tam et al. (2021) build on PET, with an additional loss that allows them to circumvent the reliance on unsupervised data and ensembling. There have been studies on aspects of prompt-based methods such as performance in the absence of prompts (Logan IV et al. 2021), quantifying scale efficiency (Le Scao and Rush 2021), learned continuous prompts (Li and Liang 2021; Lester, Al-Rfou, and Constant 2021; Qin and Eisner 2021) or gradient-based generated discrete prompts (Shin et al. 2020). Liu et al. (2021) offer a survey of prompt-based techniques. We focus on the few-shot advantages offered by PET methods and evaluating them in a cross-lingual setting.

7 Conclusion and Future Work

We have presented a holistic study of cross-lingual stance detection. We investigated PET with different (pre-)training procedures and extended it by proposing a novel label encoder that mitigates the need for translating labels into a single verbalisation. In addition to that, we introduced a novel methodology to produce artificial stance examples using a set of sentiment annotations. This yields sizeable improvements on 15 datasets over strong baselines, i.e., more than 6% F1 absolute in a low-shot and 4% F1 in a full-resource scenario. Finally, we study the impact of multi-dataset learning and pre-training with English stance data, which further boost the performance by 5% F1 absolute.

In future work, we plan to experiment with more sentiment-based models and stance task formulations, as well as different prompt engineering techniques.

References

Agerri, R.; Centeno, R.; Espnosa, M.; Fernandez de Landa, J.; and Rodrigo, A. 2021. VaxxStance: Going Beyond Text in Cross-lingual Stance Detection. In *IberLEF*. Online.

Axelrod, A.; He, X.; and Gao, J. 2011. Domain Adaptation via Pseudo In-Domain Data Selection. In *EMNLP*, 355–362. Edinburgh, Scotland, UK.

Baly, R.; Mohtarami, M.; Glass, J.; Màrquez, L.; Moschitti, A.; and Nakov, P. 2018. Integrating Stance Detection and Fact Checking in a Unified Corpus. In *NAACL-HLT*, 21–27. New Orleans, Louisiana, USA.

Barbieri, F.; Anke, L. E.; and Camacho-Collados, J. 2021. XLM-T: A Multilingual Language Model Toolkit for Twitter. *arXiv:2104.12250*.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *FAccT*, 610–623.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *arXiv:2005.14165*.

Cignarella, A. T.; Lai, M.; Bosco, C.; Patti, V.; Paolo, R.; et al. 2020. SardiStance@ EVALITA2020: Overview of the Task on Stance Detection in Italian Tweets. In *EVALITA*, 1–10.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *ACL*, 8440–8451. Online.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, 4171–4186. Minneapolis, Minnesota, USA.

Ebrahimi, J.; Dou, D.; and Lowd, D. 2016. A Joint Sentiment-Target-Stance Model for Stance Classification in Tweets. In *COLING*, 2656–2665. Osaka, Japan.

Gao, T.; Fisch, A.; and Chen, D. 2021. Making Pre-trained Language Models Better Few-shot Learners. In *ACL-IJCNLP*, 3816–3830. Online.

Hardalov, M.; Arora, A.; Nakov, P.; and Augenstein, I. 2021. Cross-Domain Label-Adaptive Stance Detection. In *EMNLP*, 9011–9028. Online and Punta Cana, Dominican Republic.

Hercig, T.; Krejzl, P.; Hourová, B.; Steinberger, J.; and Lenc, L. 2017. Detecting Stance in Czech News Commentaries. In *ITAT*, 176–180.

Hercig, T.; Krejzl, P.; and Král, P. 2018. Stance and sentiment in Czech. *Computación y Sistemas*, 22(3): 787–794.

Joshi, P.; Santy, S.; Budhiraja, A.; Bali, K.; and Choudhury, M. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *ACL*, 6282–6293. Online.

Joulin, A.; Grave, E.; Bojanowski, P.; and Mikolov, T. 2017. Bag of Tricks for Efficient Text Classification. In *EACL*, 427–431. Valencia, Spain.

Khouja, J. 2020. Stance Prediction and Claim Verification: An Arabic Perspective. In *FEVER*, 8–17. Online.

Lai, M.; Cignarella, A. T.; Hernández Farías, D. I.; Bosco, C.; Patti, V.; and Rosso, P. 2020. Multilingual stance detection in social media political debates. *CS&L*, 63: 101075.

Lai, M.; Patti, V.; Ruffo, G.; and Rosso, P. 2018. Stance Evolution and Twitter Interactions in an Italian Political Debate. In *NLDB*, 15–27.

Le Scao, T.; and Rush, A. 2021. How many data points is a prompt worth? In *NAACL-HLT*, 2627–2636. Online.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *EMNLP*, 3045–3059. Online and Punta Cana, Dominican Republic.

Li, X. L.; and Liang, P. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *ACL-IJCNLP*, 4582–4597. Online.

Li, Y.; and Caragea, C. 2019. Multi-Task Stance Detection with Sentiment and Stance Lexicons. In *EMNLP-IJCNLP*, 6299–6305. Hong Kong, China.

Lillie, A. E.; Middelboe, E. R.; and Derczynski, L. 2019. Joint Rumour Stance and Veracity Prediction. In *NoDaLiDa*, 208–221. Turku, Finland.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2021. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. arXiv:2107.13586.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692*.

Logan IV, R. L.; Balažević, I.; Wallace, E.; Petroni, F.; Singh, S.; and Riedel, S. 2021. Cutting down on prompts and parameters: Simple few-shot learning with language models. *arXiv:2106.13353*.

Lozhnikov, N.; Derczynski, L.; and Mazzara, M. 2020. Stance Prediction for Russian: Data and Analysis. In *SEDA*, 176–186.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. In *ICLR*.

Miller, G. A. 1998. *WordNet: An electronic lexical database*. MIT press.

Mohammad, S.; Kiritchenko, S.; Sobhani, P.; Zhu, X.; and Cherry, C. 2016. SemEval-2016 Task 6: Detecting Stance in Tweets. In *SemEval*, 31–41. San Diego, California.

Mosbach, M.; Andriushchenko, M.; and Klakow, D. 2021. On the Stability of Fine-tuning {BERT}: Misconceptions, Explanations, and Strong Baselines. In *ICLR*.

Pennington, J.; Socher, R.; and Manning, C. 2014. GloVe: Global Vectors for Word Representation. In *EMNLP*, 1532–1543. Doha, Qatar.

Petroni, F.; Rocktäschel, T.; Riedel, S.; Lewis, P.; Bakhtin, A.; Wu, Y.; and Miller, A. 2019. Language Models as Knowledge Bases? In *EMNLP-IJCNLP*, 2463–2473. Hong Kong, China.

Pomerleau, D.; and Rao, D. 2017. Fake News Challenge Stage 1 (FNC-I): Stance Detection.

Poth, C.; Pfeiffer, J.; Rücklé, A.; and Gurevych, I. 2021. What to Pre-Train on? Efficient Intermediate Task Selection. *arXiv:2104.08247*.

Qi, P.; Zhang, Y.; Zhang, Y.; Bolton, J.; and Manning, C. D. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *ACL: System Demonstrations*, 101–108. Online.

Qin, G.; and Eisner, J. 2021. Learning How to Ask: Querying LMs with Mixtures of Soft Prompts. In *NAACL-HLT*, 5203–5212. Online.

Ruder, S.; and Plank, B. 2017. Learning to select data for transfer learning with Bayesian Optimization. In *EMNLP*, 372–382. Copenhagen, Denmark.

Schick, T.; Schmid, H.; and Schütze, H. 2020. Automatically Identifying Words That Can Serve as Labels for Few-Shot Text Classification. In *COLING*, 5569–5578. Online.

Schick, T.; and Schütze, H. 2021a. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In *EACL*, 255–269. Online.

Schick, T.; and Schütze, H. 2021b. It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. In *NAACL-HLT*, 2339–2352. Online.

Schiller, B.; Daxenberger, J.; and Gurevych, I. 2021. Stance detection benchmark: How robust is your stance detection? *KI-Künstliche Intelligenz*, 1–13.

Shin, T.; Razeghi, Y.; Logan IV, R. L.; Wallace, E.; and Singh, S. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *EMNLP*, 4222–4235. Online.

Sobhani, P.; Mohammad, S.; and Kiritchenko, S. 2016. Detecting Stance in Tweets And Analyzing its Interaction with Sentiment. In **SEM-SemEval*, 159–169. Berlin, Germany.

Somasundaran, S.; Ruppenhofer, J.; and Wiebe, J. 2007. Detecting Arguing and Sentiment in Meetings. In *SIGDIAL*, 26–34. Antwerp, Belgium.

Somasundaran, S.; and Wiebe, J. 2010. Recognizing Stances in Ideological On-Line Debates. In *CAAGET*, 116–124. Los Angeles, CA, USA.

Sun, Q.; Wang, Z.; Li, S.; Zhu, Q.; and Zhou, G. 2019. Stance detection via sentiment information and neural network model. *Frontiers of Comp. Science*, 13(1): 127–138.

Sun, Q.; Wang, Z.; Zhu, Q.; and Zhou, G. 2018. Stance Detection with Hierarchical Attention Network. In *COLING*, 2399–2409. Santa Fe, New Mexico, USA.

Swami, S.; Khandelwal, A.; Singh, V.; Akhtar, S.; and Shrivastava, M. 2018. An English-Hindi Code-Mixed Corpus: Stance Annotation and Baseline System. *arxiv:1805.11868*.

Tam, D.; Menon, R. R.; Bansal, M.; Srivastava, S.; and Raffel, C. 2021. Improving and simplifying pattern exploiting training. *arXiv:2103.11955*.

Taulé, M.; Martí, M. A.; Rangel, F. M.; Rosso, P.; Bosco, C.; Patti, V.; et al. 2017. Overview of the task on stance and gender detection in tweets on Catalan independence at IberEval 2017. In *IberEval*, volume 1881, 157–177.

Vamvas, J.; and Sennrich, R. 2020. X-Stance: A Multilingual Multi-Target Dataset for Stance Detection. In *SwissText-KONVENS*. Zurich, Switzerland.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *BlackboxNLP*, 353–355. Brussels, Belgium.

Waseem, Z.; Lulz, S.; Bingel, J.; and Augenstein, I. 2021. Disembodied machine learning: On the illusion of objectivity in NLP. *arXiv preprint arXiv:2101.11974*.

Xu, R.; Zhou, Y.; Wu, D.; Gui, L.; Du, J.; and Xue, Y. 2016. Overview of NLPCC Shared Task 4: Stance Detection in Chinese Microblogs. In *NLPCC*, 907–916.

Zotova, E.; Agerri, R.; Nuñez, M.; and Rigau, G. 2020. Multilingual Stance Detection in Tweets: The Catalonia Independence Corpus. In *LREC*, 1368–1375. Marseille, France.