VaxxStance: A Dataset for Cross-Lingual Stance Detection on Vaccines

Rodrigo Agerri [1], Roberto Centeno [2], María Espinosa [2], Joseba Fernandez de Landa[1], Álvaro Rodrigo [2]

[1]HiTZ Center - Ixa, University of the Basque Country UPV/EHU
[2]NLP&IR group at Universidad Nacional de Educación Distancia (UNED)
rodrigo.agerri@ehu.eus, rcenteno@lsi.uned.es, mespinosa@lsi.uned.es,
joseba.fernandezdelanda@ehu.eus, alvarory@lsi.uned.es

Abstract

This paper describes a dataset which proposes to detect stance in tweets referring to vaccines, a relevant and controversial topic in the current pandemia. The task is proposed in a multilingual and crosslingual setting, providing data for both Basque and Spanish languages. In addition to facilitating research on crosslingual techniques, the dataset also provides relational data (retweets, friends) from social media with the aim of complementing the textual input to perform stance classification. Finally, apart from describing the data collection and annotation of the dataset, we also include two basic baselines using both textual and relational information. In summary, we believe that the VaxxStance dataset provides an interesting benchmark to investigate crosslingual approaches to stance detection based on both textual and contextual features.

1 Introduction

In this paper we understand stance as establishing whether a given tweet expresses a FAVOR, AGAINST or NEUTRAL (NONE) attitude with respect to a given, pre-defined topic (Mohammad et al. 2016). As it is usually the case in the Natural Language Processing (NLP) field, the majority of work on this research topic have been focused on English, although this situation is rapidly changing. An Arabic corpus integrated the tasks of fact-checking and stance detection (Baly et al. 2018), a dataset from comments of news was developed for Czech language (Hercig et al. 2017), and there also works for French (Evrard et al. 2020) and Russian (Vychegzhanin 2019). Finally, an interesting new dataset for Italian was released in 2020 as part of the SardiStance@Evalita 2020 shared task (Cignarella et al. 2020), which included not only the texts of the tweets labeled with stance, but also social network information relative to the authors of the tweets. This social network information includes retweets, user accounts profile, friends and followers, among others.

Other interesting works have tried to address stance detection from a multilingual point of view. The IberEval 2017 and 2018 shared tasks (Taulé et al. 2018) provided a dataset in Catalan and Spanish to classify stance with respect to the Independence of Catalonia, while Lai et. al (2020) provided datasets in French and Italian. However, these multilingual efforts are hindered by the extremely skewed class distribution in the Catalan IberEval data, or by the fact that the data for each language was not collected on the same timeframe and addressed different topics. This makes it very difficult to investigate multilingual and crosslingual approaches to stance detection. While Zotova et al. (2021) propose a method to address these shortcomings by providing a semi-automatically generated multilingual stance detection corpus, they do not include social network features in their dataset.

In this context, we present the VaxxStance dataset with the aim of detecting stance in social media on vaccines in general. The data includes two languages, Basque and Spanish, and its objective is to promote crosslingual research on stance detection using both the text and the relational information provided by the Twitter social network. Thus, and unlike previous approaches, we provide, for a given topic, multilingual coetanous data of gold-standard quality in a corpus which allows to experiment using both social and textual features in multilingual and crosslingual settings.

2 Multilingual Dataset

Following the formulation of stance provided by Mohammad et al. (2016), for VaxxStance we annotate tweets as expressing an AGAINST, FAVOR or NEUTRAL (NONE) stance towards vaccines. Additionally, and inspired by the SardiStance 2020 shared task (Cignarella et al. 2020), the dataset includes two different types of data: Textual and Contextual (retweets, friends and user data), for two languages, Basque and Spanish. The dataset and baselines are publicly available[1].

2.1 Collection and Annotation

In a first attempt we tried to do the data collection and annotation for both languages in the same manner. However, as it will be explained below, due to the idiosyncrasies of Basque it was necessary to devise an alternative, more viable, method for that language, especially to obtain the required textual data. In any case, we did specify a number of criteria that both languages needed to comply with. First, the datasets were required to have a balanced distribution in the

[1]https://vaxxstance.github.io/

ratio users/tweets to avoid that a large number of tweets belonged to a very few users. Second, the tweets in the training set had to be written by different users from those contained in the test set. This is to avoid obtaining artificially high results due to the existence of user-based information in both the training and test sets. As such, the general idea is that both the textual and user-based (or contextual) knowledge would help each other in order to better classify stance. Finally, we use the annotation guidelines from the SemEval 2016 task (Mohammad et al. 2016).

Basque Basque is spoken by roughly the 30% of the population in the Basque Country, and understood by around 50%. Due to the fact that Basque is a co-oficial language, it does have presence in the regional public administration, as well as in the education system and some news media, including a public television broadcaster. Still, the presence of Basque in mass media is extremely low, especially when compared to Spanish, the 4th most spoken language in the world.

In this context, the increasing popularity of Twitter among Basque speakers is of particular importance for a low resource language, as a relatively large amount of textual content written in Basque is generated in that social network. This provides a valuable resource to study new NLP tasks such as stance detection not only for large and popular languages, but also for low resourced ones. Still, the collection process of enough tweets relevant to the VaxxStance task was rather challenging.

At first we experimented with a keyword extraction method using the following specific keywords: *"txertoa"* (vaccine) and *"txertaketa"* (vaccination), *"negazionista"* (negationist), *#pfizer*, *#moderna*, *#astrazeneca* and their respective inflections. However, it was surprising to find that the traffic of Basque tweets relative to those topics were relatively low.

We therefore decided to try an alternative, more brute-force, method. First, we collected all the available timelines of users that are identified to write mostly in Basque (around 10k users). The content of these timelines amount to around 8M tweets. Second, relevant tweets were selected following a simple keyword search using the same keywords listed for the previous attempt. Third, a first annotator manually labeled a set of around $1,400$ tweets. Finally, those same $1,400$ tweets, belonging to 210 users, were blindly annotated by a second annotator. The final composition of the textual part of the dataset can be seen in Table 1.

		Train	Test
Tweets		1,072	312
	Favor	327	85
	Neutral	524	135
	Against	219	92
Users		149	61

Table 1: Textual data in the Basque dataset.

We would like to note that the most difficult part in the process was finding enough users that explicitly expressed a stance AGAINST vaccines.

Spanish Around $2,700$ tweets written in Spanish stating an opinion about "vaccines" were collected and annotated, as shown by Table 2. In order to avoid a potential bias derived from the current COVID-19 pandemic situation, the tweets were collected from the beginnings of Twitter until current time. They were also restricted to the peninsular variant of the Spanish language in order to avoid problems derived from the use of different terms in other variants such as Colombian, Peruvian, etc. To guide this process we used the Google tool *"Google Trends"*[2] which allowed us to locate temporal spaces where events related to vaccines had occurred, identifying the type of event and the date on which it happened. Some examples are the peaks in traffic for and against the vaccination against measles, which was a consequence of some measles outbreaks that happened in Spain during 2019. By using keywords related to the event and restricting the dates obtained, we managed to introduce tweets related to events other than the COVID-19 vaccination process.

		Train	Test
Tweets		2,003	694
	Favor	937	359
	Neutral	591	195
	Against	475	140
Users		1,261	414

Table 2: Textual data in the Spanish dataset.

In addition to the tweets collected through the events identified in Google Trends, for the rest of the tweets collected we followed the following process. First, we used a set of keywords such as "vaccine", "vaccination", as well as terms related to diseases whose vaccines have generated some controversy in society and in anti-vaccine movements, e.g., "chickenpox", "autism", "MMR", etc. After a first manual analysis, we observed that the vast majority of the tweets collected did not express a stance. In order to solve this problem, we then extracted the hashtags most commonly used in these tweets and manually analysed those that were used to express a position in favour and/or against vaccines. Some examples of these hashtags are *#YoMeVacuno*, *#VaccinesWork*, *#COVID19*, *#vacuna*, *#yomevacuno*, *#VacunaCOVID19*, *#YoNoMeVacuno*, *#gripe*, *#Plandemia*, *#yosimevacuno*, etc.

By using these hashtags, we managed to increase the number of tweets to start with the manual labeling. The labelling was performed manually by two annotators, using a third annotator to resolve disagreements. For this we used the web platform created by Cignarella et al. (2020), to whom we would like to thank for their help using it.

Once the manual annotation was completed, the set of AGAINST tweets was much smaller than those expressing a FAVOR or NEUTRAL stance. To address this issue,

[2]https://trends.google.es/trends/?geo=ES

we identified several accounts of users that may potentially be identified as supporters of anti-vaccine movements and manually collected tweets from these users expressing an AGAINST stance. This step was performed taking care in complying with the general criteria of not including more than 10 tweets per user in the final corpus, as well as not overlapping users between the training and evaluation set. In this final process we managed to increase by about 200-250 tweets the AGAINST class.

2.2 Social Media Information

The main objective in developing our dataset was studying the usefulness of the context provided by social media information to classify stance in a crosslingual setting. With this objective in mind, we collected contextual information relative to the *friends* of the authors of the tweets as well as their *retweets*. The context provided by *friends* and *retweets* can be leveraged to generate relation graphs that in turn may be used to improve the classifiers.

Table 3 shows the social media data gathered with respect to the tweets in the train and test partitions for each of the languages. In addition to the retweets of the tweets included in the datasets, for Basque we also decided to collect all the retweets made by the users, namely, by extracting the retweets from the users' timelines (TL). This strategy was applied in order to alleviate the small number of retweets obtained from the tweets in the train and test partitions.

		Train	Test
Basque	Friends	119,977	53,029
	Retweets	203	0
	Retweets (TL)	130,369	61,438
Spanish	Friends	1,708,396	438,586
	Retweets	6,832	2,148

Table 3: Social Media Information by language.

Finally, apart from social media information, the dataset also includes the meta information of each annotated tweet as well as the information related to each user.

2.3 Final Dataset

Table 4 shows the composition of the VaxxStance dataset, including both textual and contextual information. Regarding the textual information, it can bee seen that the Spanish set is roughly double in size with respect to the Basque one, although the distribution of classes across the train and test set, as shown by Tables 1 and 2, is quite similar.

With respect to the contextual information we can see that for Basque there are very few users, around 10% of the number of users for Spanish. This is a reflection of the much smaller community of Twitter users that write in Basque. In this sense, the *friends* graph also reflects the same ratio, as the number of *friends* relations is around 10% of those obtained for Spanish. If we look at the *retweets*, however, we can see that for Basque we only managed to obtain very few of them. That is why we decided to also provide the retweets for each user in the train and test sets (*retweets TL*).

		Train	Test
Basque	Tweets	1,072	312
	Users	149	61
	Friends	119,977	53,029
	Retweets	203	0
	Retweets (TL)	130,369	61,438
Spanish	Tweets	2,003	694
	Users	1,261	414
	Friends	1,708,396	438,586
	Retweets	6,832	2,148

Table 4: Composition of the VaxxStance 2021 dataset.

In summary, we believe that the VaxxStance dataset provides an interesting benchmark to investigate crosslingual approaches to stance detection based on both textual and contextual features. While the Basque set is slightly smaller than some previous approaches (Taulé et al. 2018; Cignarella et al. 2020; Zotova, Agerri, and Rigau 2021) it is still larger than the data provided for any of the topics in the SemEval 2016 dataset, which is perhaps the most popular benchmark for stance detection (Mohammad et al. 2016).

3 Tasks Definition

As we aim to promote research on multilingual and crosslingual approaches to stance detection in Twitter combining both textual and contextual data, we envisage a number of tasks that may be of interest to develop systems for stance detection on our dataset. We believe that this type of research requires annotated datasets on a common topic for more than one language and obtained on the same dates (coetaneous data). However, while previous work mentioned in the Introduction includes datasets in several languages, they do not provide an adequate evaluation setting for multilingual and crosslingual studies to stance detection. The initial tasks for experimentation may be the following:

- Close task: Language-specific evaluation. Only the provided data for each of the languages is allowed. There are two evaluation settings:
  - Textual: Only the provided tweets in the target language can be used for development. No data augmentation allowed.
  - Contextual: Text plus given Twitter-related information will be used by the participants. Contextual information refers to features related with user-based Twitter information: friends, retweets, etc. described in Section 2.2.
- Open Track: Participants can use any kind of data, including additional tweets obtained by the participants. The main objective consists of exploring data augmentation and knowledge transfer techniques for cross-lingual stance detection.
- Zero-shot Track: Texts (tweets) of the target language cannot be used for training. The main objective would be to explore how to develop systems that do not have access to text in the target language, especially using Twitter-related information.

Following previous work, we evaluate the systems with the metric provided by the SemEval 2016 task on Stance Detection (Mohammad et al. 2016) which reports F1 macro-average score of two classes, FAVOR and AGAINST, although the NONE class is also represented in the test data:

$$F1_{avg} = \frac{F1_{favor} + F1_{against}}{2} \quad (1)$$

We also provide an official evaluation script distributed together with the dataset in the task website[3].

3.1 Baselines

As starting point, we train two simple baselines, one using only textual information and a second one using just social or contextual features:

- Textual: The textual baseline is based on a SVM classifier with RBF kernel function. The text of the tweets is vectorized using a TF-IDF vectorizer and then feed to the classifier. Both *C* and *Gamma* hyperparameters are tuned by means of grid search and 5 fold CV on the training data. The best configuration is used to evaluate on the test.
- Social: This classifier uses the metadata related to each user and tweet to obtain a number of features (friends count, status count, emojis in bio, etc.) which are then used to train a XGBoost classifier. Before feeding the classifier, each class data is weighted in order to create a balanced sample.

		Against	Favor	Average
Basque	Textual	51.80	57.01	54.41
	Social	5.23	48.53	26.88
Spanish	Textual	71.38	81.68	76.53
	Social	73.14	73.73	73.43

Table 5: Baseline results on Test set.

The results obtained by the baselines show that both tracks are harder for Basque. With respect to the Textual track, stance in Spanish seems to be expressed more explicitly. Regarding the social baseline, the low results were probably caused by the low number of Basque users from which to obtain the features.

4 Concluding Remarks

In this paper we provide an overview of the VaxxStance dataset, in which the objetive is to detect stance towards vaccines across two different languages: Basque and Spanish. As a novelty for stance detection in these languages, systems can use textual and contextual information to train their systems in multilingual and crosslingual settings.

The datasets for both languages were built following the same criteria and objectives. However, further analysis is required to understand why results are systematically better for Spanish than those obtained for Basque. Further work includes developing good relational representations for the Twitter-based contextual information (retweets, friends) to improve the results obtained by the textual classifiers.

[3]https://vaxxstance.github.io/

Acknowledgements

This work has been partially funded by the UPV/EHU Colab 19/19 project "Tools for the analysis of parliamentary discourses: polarization, subjectivity and affectivity in the post-truth era". Rodrigo Agerri acknowledges the support received from the RYC-2017–23647 fellowship and from the ANTIDOTE - EU CHIST-ERA project (PCI2020-120717-2) of the Agencia Estatal de Investigación through the INT-Acciones de Programación Conjunta Internacional (MINECO) 2020 call.

References

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 *In NAACL Volume 2 (Short Papers)*, 21–27.

Cignarella, A. T.; Lai, M.; Bosco, C.; Patti, V.; and Rosso, P. 2020. SardiStance@EVALITA2020: Overview of the Task on Stance Detection in Italian Tweets. In *In EVALITA 2020*.

Evrard, M.; Uro, R.; Hervé, N.; and Mazoyer, B. 2020. French Tweet Corpus for Automatic Stance Detection. In *Proceedings of The 12th Language Resources and Evaluation Conference*, 6317–6322. Marseille, France: European Language Resources Association.

Hercig, T.; Krejzl, P.; Hourová, B.; Steinberger, J.; and Lenc, L. 2017. Detecting Stance in Czech News Commentaries. In *Proceedings of the 17th ITAT: Slovenskočeský NLP workshop (SloNLP 2017)*, volume 1885, 176–180.

Lai, M.; Cignarella, A.; Hernandez Farias, D.; Bosco, C.; Patti, V.; and Rosso, P. 2020. Multilingual Stance Detection in Social Media Political Debates. *Computer Speech & Language* .

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

Taulé, M.; Pardo, F. M. R.; Martí, M. A.; and Rosso, P. 2018. Overview of the Task on Multimodal Stance Detection in Tweets on Catalan# 1Oct Referendum. In *IberEval@ SEPLN*, 149–166.

Vychegzhanin, S. V. Kotelnikov, E. V. 2019. Stance Detection Based on Ensembles of Classifiers. *Programming and Computer Software* 228–240.

Zotova, E.; Agerri, R.; and Rigau, G. 2021. Semi-automatic generation of multilingual datasets for stance detection in Twitter. *Expert Systems with Applications* 170: 114547.