Procesamiento del Lenguaje Natural, Revista nº 67, septiembre de 2021, pp. 173-181 recibido 30-06-2021 revisado 07-07-2021 aceptado 11-07-2021

VaxxStance@IberLEF 2021: Overview of the Task on Going Beyond Text in Cross-Lingual Stance Detection

VaxxStance@IberLEF 2021: Descripción de la tarea de detección de actitudes basada en el uso de información más allá del texto

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 a 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 the VaxxStance task at IberLEF 2021. The task 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 setting, providing data for Basque and Spanish languages. The objective is to explore crosslingual approaches which also complement textual information with contextual features obtained from the social network. The results demonstrate that contextual information is crucial to obtain competitive results, especially across languages.
Keywords: Stance Detection, Multilingualism, Computational Social Science, Information Extraction.

Resumen: En este artículo se describe la tarea VaxxStance celebrada en el marco de IberLEF 2021. La tarea propone detectar la actitud de un conjunto de tweets relativos a las vacunas, a un tema muy actual y polémico en estos tiempos de pandemia. La tarea se ha propuesto en un marco multilingüe, euskera y español. Además del texto de cada tweet, se ha proporcionado además información relacionada con la red social de los usuarios autores de los tweets. Los resultados de los participantes han corroborado que el uso de información de la red social permite mejorar el rendimiento en esta tarea, particularmente en un entorno crosslingüe.
Palabras clave: Detección de Actitudes, Multilingüismo, Ciencias Sociales Computacionales, Extracción de Información.

1 Introduction

Stance detection is one of the tasks within the universe of Fake News detection and as such is related to other tasks such as Hyperpartisanism (Kiesel et al., 2019), Hate Speech Detection (Basile et al., 2019), Fact-checking and Claim Verification (Thorne et al., 2018), among others. The most popular formulations are perhaps those proposed in 2016 by the SemEval-2016 Task 6: Detecting Stance in Tweets (Mohammad et al., 2016) and by the Fake News Challenge (Stage 1)[1]. In the first, stance is defined as establishing whether a given tweet expresses a FAVOR, AGAINST or NEUTRAL (NONE) attitude with respect to a given, pre-defined topic. In the second formulation, provided by the Fake News Challenge, stance has to be inferred between a claim and a text commenting on the claim. In this case the stance category can be one of Agrees, Disagrees, Discusses and Unrelated.

Other subsequent contributions have focused mostly on the static variant of stance detection (with respect to a pre-defined topic) rather than on the dynamic one (classifying stance with respect to previous message), although there are some exceptions, notably the RumourEval tasks (Derczynski et al., 2017; Gorrell et al., 2019).

Furthermore, as it is usually the case in the Natural Language Processing (NLP) field, most works have experimented on En-

[1]http://www.fakenewschallenge.org/

ISSN 1135-5948. DOI 10.26342/2021-67-15

glish only, with some exceptions. 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, Agerri, and Rigau (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 propose the VaxxStance shared task at IberLEF 2021 (Montes et al., 2021), with the aim of detecting stance in social media on vaccines in general. The task provides data in two languages, Basque and Spanish, and its objective is to promote crosslingual research on stance detection using both the text and the 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), the VaxxStance task consists of determining whether a given tweet expresses 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 language, Basque and Spanish. The dataset is publicly available in the task website[2].

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 a required to have a balanced distribution in the 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).

2.1.1 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

[2] `https://vaxxstance.github.io/`

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.

2.1.2 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"*[3] 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

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

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, 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 of this task is 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 the 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.

		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.

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*).

In summary, 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 *Task Definition*

In this task we aimed to promote research on multilingual and crosslingual approaches to stance detection in Twitter. Ideally, 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.

3.1 Tracks

As the task contains tweets in two different languages, we proposed the following participation tracks for each language (Basque and Spanish):

- Close Track: 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 is 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 is to explore how to develop systems that do not have access to text in the target language, especially using Twitter-related information.

Participants could submit their systems to any of the tracks, but it was compulsory to participate in both languages for the chosen track.

3.2 Evaluation

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:

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

The official evaluation script is distributed together with the dataset in the task website[4].

3.3 Baselines

We provide two 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

[4] `https://vaxxstance.github.io/`

Basque users from which to obtain the features.

4 Participants and Results

Twenty groups registered for the task and downloaded the datasets. However, only three groups finally submitted runs. Table 6 shows the information of the participant groups and the reference to their reports.

Team	Report
MultiAztertest	(Gonzalez-Dios and Bengoetxea, 2021)
SQYQP	(Calleja and Méndez, 2021)
WordUp	(Lai et al., 2021)

Table 6: Participants.

In total, the participants submitted 28 runs, 14 per language. While all the three teams participated in both Textual and Contextual settings of the Close Track, only one team, WordUp, participated in the Zero-shot and Open Tracks. Thus, any comparisons between the participant systems will be performed on the Close Track.

4.1 Close Track

Table 7 shows the results for the Close Track, which received 20 submission runs. We report results for each language and evaluation setting (Textual and Contextual). For all the four rankings, the best results are always obtained by the WordUp team.

As it was the case with our baseline results, the participant systems score systematically higher for Spanish. The best results for Spanish are over an 80 F1 score. These results seem to confirm that the Spanish set is easier than the Basque one.

For each language, the results improve by using contextual information, except for the MultiAztertest Basque submissions. Still, results confirm the effectiveness of employing both textual and social information.

Regarding the results of baselines, the textual baseline obtains competitive results in both languages, being only outperformed by the WordUp team. In contrast, the contextual baseline's results are improved by all the teams in Basque (except one run from WordUp which obtains a very low score) and at least one run per group, except SQYQP, in Spanish. These results suggest that, despite their simplicity and use of linear classifiers, the approaches followed by both baselines represent an adequate starting point.

Regarding the techniques followed by the participants, MultiAzterTest tested two different approaches for the Textual setting: a system which used pre-trained transformers-based language models (run 1) and another one based on training a classifier with a set of linguistic and stylistic features such as word frequencies, semantic overlap, etc. (run 2). The first approach performed much better than the second in the textual track. For the contextual track, they employ only the information relative to the user. More specifically, for each user they select the most common stance label, assuming that users tweets correspond coherently to one stance type. While this idea worked well in Spanish, it was detrimental in (run 1) in Basque.

The SQYQP team addressed the textual setting by training a LSTM initialized with multilingual BERT embeddings using the Flair toolkit (Akbik, Blythe, and Vollgraf, 2018). For the contextual setting, they added network information and measured the distance among users following the approach done by Espinosa et al. (2020). While their results for Basque are below the textual baseline, they improve results by adding contextual information.

The WordUp! team employed a large number of common features that have been proved useful for stance detection. Features were extracted mostly from stylistic and dependency analyses. Additionally, the also crawled tweets specifically for this task and topic for both languages. The tweets were then used to train FastText word embeddings and used to obtain several features. Moreover, they created a dictionary of lemmas referring to stance in English, and translated it to Spanish and Basque. In the contextual setting they tried several network-based measures to be added as features to the logistic regression classifier. They report a large number of experiments which resulted in the best performing team across all evaluation tracks.

In general, the results obtained by the participants show that, even when using very simple features, contextual information substantially improved the results obtainecd by using text only.

		Against	Favor	F1 Macro
Basque Textual	WordUp_01	57.69	56.99	57.34
	WordUp_02	55.03	54.27	54.65
	*BASELINE	51.80	57.01	54.41
	MultiAztertest_01	48.23	52.25	50.24
	SQYQP_01	38.81	46.31	42.56
	MultiAztertest_02	34.38	34.18	34.28
Spanish Textual	WordUp_02	78.36	83.47	80.92
	WordUp_01	75.54	82.58	79.06
	*BASELINE	71.38	81.68	76.53
	MultiAztertest_01	66.67	81.53	74.10
	SQYQP_01	57.14	77.61	67.38
	MultiAztertest_02	56.47	71.60	64.04
Basque Contextual	WordUp_02	82.95	72.46	77.71
	SQYQP_01	65.17	52.94	59.06
	MultiAztertest_02	25.40	48.03	36.72
	MultiAztertest_01	16.36	56.06	36.21
	*BASELINE	5.23	48.53	26.88
	WordUp_01	0.00	0.08	0.04
Spanish Contextual	WordUp_02	91.17	87.09	89.13
	WordUp_01	88.97	86.56	87.77
	MultiAztertest_01	78.77	79.84	79.31
	*BASELINE	73.14	73.73	73.43
	SQYQP_01	66.27	80.06	73.17
	MultiAztertest_02	63.93	77.17	70.55

Table 7: Close Track official results.

4.2 Open Track

The only participant in this track was the WordUp! team, which performed data augmentation. They generated FastText word embeddings (Bojanowski et al., 2017) from a set of tweets specifically obtained for this particular task and languages. They also augmented the contextual information by extracting the social network of each user. As it can be seen in the results reported in Table 8, their results are quite similar to those obtained in the Close Track - Contextual setting. This might be due to the fact that they also used the ad-hoc generated FastText embeddings also in the Close Track.

		Against	Favor	F1 Macro
Basque	WordUp_02	82.29	72.12	77.21
	WordUp_01	64.47	68.12	66.30
Spanish	WordUp_02	90.87	88.07	89.47
	WordUp_01	90.39	88.01	89.20

Table 8: Open Track official results.

4.3 Zero-shot Track

Table 9 shows the results obtained by the only participant in this track, in which the participants could not use the text (tweets) of the target language for training. The most surprising aspect of the results is perhaps the fact that, for Basque, the zero-shot results outperform the results of the Textual evaluation setting. This seems to indicate that contextual information is far more important than the texts themselves in order to perform stance detection.

		Against	Favor	F1 Macro
Basque	WordUp_01	64.47	68.12	66.30
	WordUp_02	55.70	39.74	47.72
Spanish	WordUp_01	88.03	46.13	67.08
	WordUp_02	18.63	62.77	40.70

Table 9: Zero-Shot Track official results.

5 *Concluding Remarks*

In this paper we provide an overview of the VaxxStance@IberLEF 2021 shared evaluation task, 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 techniques employed by the different participants showed that contextual information has a great impact across languages, even for small community of users such as Basque. In this sense, textual results are in general improved by adding social network features.

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. Finally, given that just one team participated in the Open and Zero-shot Tracks, one of the main objectives of the task, to promote research on crosslingual approaches to stance detection, has not completely been achieved. Therefore further work is required on this particular line of research.

Acknowledgments

This work has been partially supported by the European Social Fund through the Youth Employment Initiative (YEI 2019) and the Spanish Ministry of Science, Innovation and Universities (DeepReading RTI2018-096846-B-C21, MCIU/AEI/FEDER, UE), and by the DeepText project (KK-2020/00088), funded by the Basque Government. Rodrigo Agerri is also funded by the RYC-2017-23647 fellowship.

Finally, we are grateful to Mirko Lai and Alessandra Cignarella for sharing with us their experience organizing the SardiStance 2020 shared task.

References

Akbik, A., D. Blythe, and R. Vollgraf. 2018. Contextual string embeddings for sequence. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA.

Baly, R., M. Mohtarami, J. Glass, L. Màrquez, A. Moschitti, and P. Nakov. 2018. Integrating Stance Detection and Fact Checking in a Unified Corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 21–27, New Orleans, Louisiana, June. Association for Computational Linguistics.

Basile, V., C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, and M. Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov. 2017. Enriching word vectors with subword information. *TACL*, 5:135–146.

Calleja, J. and A. Méndez. 2021. Sqyqp@vaxxstance: Stance detection for the antivaxxers movement. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, CEUR Workshop Proceedings.

Cignarella, A. T., M. Lai, C. Bosco, V. Patti, and P. Rosso. 2020. SardiStance@EVALITA2020: Overview of the Task on Stance Detection in Italian Tweets. In V. Basile, D. Croce, M. Di Maro, and L. C. Passaro, editors, *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*. CEUR-WS.org.

Derczynski, L., K. Bontcheva, M. Liakata, R. Procter, G. Wong Sak Hoi, and A. Zubiaga. 2017. SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76, Vancouver, Canada, August. Association for Computational Linguistics.

Espinosa, M. S., R. Agerri, Á. Rodrigo, and R. Centeno. 2020. Deepreading @ sardistance 2020: Combining textual, social and emotional features. In V. Basile, D. Croce, M. D. Maro, and L. C. Passaro, editors, *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020*, volume 2765 of *CEUR Workshop Proceedings*. CEUR-WS.org.

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

Gonzalez-Dios, I. and K. Bengoetxea. 2021. Multiaztertest@vaxxstance-iberlef 2021: Identifying stances with language models and linguistic features. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, CEUR Workshop Proceedings.

Gorrell, G., E. Kochkina, M. Liakata, A. Aker, A. Zubiaga, K. Bontcheva, and L. Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Hercig, T., P. Krejzl, B. Hourová, J. Steinberger, and L. Lenc. 2017. Detecting stance in czech news commentaries. In *Proceedings of the 17th ITAT: Slovenskočeský NLP workshop (SloNLP 2017)*, volume 1885 of *CEUR Workshop Proceedings*, pages 176–180, Bratislava, Slovakia. Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, CreateSpace Independent Publishing Platform.

Kiesel, J., M. Mestre, R. Shukla, E. Vincent, P. Adineh, D. Corney, B. Stein, and M. Potthast. 2019. SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Lai, M., A. Cignarella, D. Hernandez Farias, C. Bosco, V. Patti, and P. Rosso. 2020. Multilingual Stance Detection in Social Media Political Debates. *Computer Speech & Language*, 02.

Lai, M., A. T. Cignarella, L. Finos, and A. Sciandra. 2021. Wordup! at vaxxstance 2021: Combining contextual information with textual and dependency-based syntactic features for stance detection. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, CEUR Workshop Proceedings.

Mohammad, S., S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *SemEval-2016)*, pages 31–41.

Montes, M., P. Rosso, J. Gonzalo, E. Aragón, R. Agerri, M. Ángel Álvarez Carmona, E. Álvarez Mellado, J. C. de Albornoz, L. Chiruzzo, L. Freitas, H. G. Adorno, Y. Gutiérrez, S. M. J. Zafra, S. Lima, F. M. P. de Arco, and M. T. (eds.). 2021. Proceedings of the iberian languages evaluation forum (iberlef 2021). CEUR Workshop Proceedings.

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

Thorne, J., A. Vlachos, C. Christodoulopoulos, and A. Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana, June. Association for Computational Linguistics.

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

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