MultiAzterTest@VaxxStance-IberLEF 2021: Identifying Stances with Language Models and Linguistic Features

Itziar Gonzalez-Dios[1][0000−0003−1048−5403] and Kepa Bengoetxea[1][0000−0002−0289−6897]

Ixa Group, HiTZ center / University of the Basque Country (UPV/EHU)
{kepa.bengoetxea,itziar.gonzalezd}@ehu.eus

Abstract. Detecting stances is a Natural Language Processing task that has focused mainly on analysing debates and controversial topics. In this case, the VaxxStance@IberLEF 2021 shared task has focused on the Antivaxxers movement in Basque and Spanish tweets. In this paper, we present the participation of the MultiAzterTest team and test two approaches: a language model based approach and a linguistic and stylistic feature based approach. We also introduce the "one stance per *tuiter@lari*" heuristic to integrate contextual information. The best results are obtained with language models, but the linguistic and stylistic feature based approach offers more interpretability.

Keywords: Stance detection · VaxxStance-IberLEF · Language Models · Linguistic and stylistic features

1 Introduction

Identifying stances in social media has gained a lot of interest in the Natural Language Processing (NLP) community and debates [21] and political debates have been the main topics [15]. Detecting Stance in tweets as shared task was first organised in SemEval-2016 [17], but since them similar shared tasks have been carried out e.g. about Catalan referendum [22] or about the Sardines movement in Italy [8]. These tasks usually try to detect position on controversial and trendy topics. In this case, VaxxStance@IberLEF 2021 share task [1], which is organised in IberLEF 2021 [18], focuses on the Antivaxxers movement in Basque and Spanish. The aim of the task is to state if a tweet expresses an against, favor or neutral (none) stance.

In this paper we present the participation of the MultiAzterTest team in the close track of VaxxStance@IberLEF 2021, a language-specific evaluation where only the provided data for each language is allowed to use. There are, moreover, two settings in this track: i) textual, where only the tweets can be used and ii) the

IberLEF 2021, September 2021, Málaga, Spain.

contextual, where in addition to the texts, features related to user-based Twitter information can be used. To tackle this task, we present two approaches: the first one is based on language models and the second one is based on linguistic and stylistic features plus a classical machine learning classifier. In order to include the contextual information, we apply a heuristic inspired by the "one sense per discourse" [10].

Language models have been proved to be very effective in many NLP tasks. However, they lack of interpretability. That is why, we think that the use of linguistic and stylistic features may help to understand the underlying linguistic characteristics that are used when expressing a con or pro opinion. With that in mind, our aim is to explore if these features help in the task.

This paper is structured as follows: in Section 2 we present the corpus analysis carried out with MultiAzterTest, in Section 3 we describe our approaches and the experimental set-up, in Section 4 we present the results and we conclude and outline the future work in 5.

2 Exploratory Analysis of the VaxxStance corpus with MultiAzterTest

In this section we present the exploratory analysis of the linguistic features of VaxxStance corpus [1]. In order to carry out this analysis we have used MultiAzterTest [5]. MultiAzterTest is an open source tool and web application which analyses more than 125 linguistic and stylistic features in Basque (125 features) English (163 features), and Spanish (141 features). Following, we briefly explain how MultiAzterTest works:

1. **Preprocessing:** This step carries out all the necessary analysis in raw texts in order to be processed. This includes multilingual parsing (in our case Stanza [20]), syllable splitting, and stopword removing.
2. **Linguistic and stylistic profiling:** Based on the previous text analysis, this step calculates the linguistic and stylistic features. These features are grouped in the following types: descriptive and raw features, lexical diversity, classical readability formulae, word frequencies, vocabulary knowledge, morphological information, syntax, semantic information, semantic overlap (semantic similarity), referential cohesion (overlaps) and logical cohesion (connectives). There are five types of indicators: absolute numbers, mean, standard deviation, incidence out of 1000 and ratios.
3. **Classification:** Based on the linguistic and stylistic features, a machine learning classifier is applied. This classifier varies depending on the task. In the case of readability assessment, for example, support vector machines seem to be the most adequate.

Based on the linguistic and stylistic profiling of MultiAzterTest, we present in Table 1 the mean of some descriptive linguistic features of the VaxxStance dataset.

Table 1. Linguistic features of the dataset (train)

Variable (mean)	Basque	Spanish
Word length	7.404	5.284
Lemma length	6.469	5.192
Sentence length	11.240	16.65
Depth per sentence	4.295	5.021
Propositions per sentence	2.421	2.613
Polysemy index	6.102	4.031
Incidence of different words	908.8	873.1
Incidence of different rare words	42.65	170.9
Incidence of content words	592.1	434.3
Incidence of negation	18.07	12.94
Incidence of connectives	90.52	60.93

As we can see, the words and the lemmas are longer in Basque than in Spanish, but sentences are longer and deeper in Spanish. The propositions per sentence is similar in both languages. Regarding lexico-semantic measures, there are more different words in Basque, but there are less rare words. This may indicate that although there are many different words, these words are common. The incidence of content words is also bigger in Basque due to its typology. In the Basque words, there are more words that express the negation (negative particles) than in Spanish. However, the incidence is low and this leads us to think that contrary opinions may be subtle and not so direct. The use of connective is much bigger in Basque. It would be interesting to see if Basque tweets are written in a more formal, elaborated register than the Spanish ones.

Due to the high number of features, it is possible that some of them highly correlate in this dataset. We have also analysed the correlation with the python's package Feature Selector [14] and we see that 17 features in Basque and 37 in Spanish have a correlation magnitude greater than 0.98. For example, the incidence of the adjectives and adverbs correlate at 1.0 with adjective and adverb density respectively in both languages, which may indicate that both features are representing the same information in this dataset. Curiously, we see that in Basque stem and noun overlap have a correlation of 0.9978, which may indicate that nouns are widely used in this dataset and in Spanish, the mean of rare words and the mean of distinct rare words correlate at 0.9949, which may show that rare words are used few times.

Finally, we have also analysed the most predictive features with Weka's [12] Infogain (Table 2). From a linguistic and stylistic point of view, in both languages descriptive features, morphological features and morpho-syntactic features are on the top. In Basque there is a tendency to use normalised metrics (means, incidences), while in Spanish the raw numbers play an important role. In the case of tweets, as they have a limit for characters and more or less they do not differ to much in length, this may not be so important but in the case of text of different sizes, raw numbers may lead to misleading conclusions. It seems that vocabulary related features e.g. features related to content words, or rare

Table 2. Top10 features according to InfoGain in Basque and Spanish.

Basque	**Spanish**
lemma length (mean)	number of different forms
words length (mean)	number of words without punctuation
noun phrase density (incidence)	number of words
number of proper nouns (incidence)	number of paragraphs (incidence)
number of verbs (incidence)	number of lexical words
verb density	number of 1st person incidence
verb phrase density (incidence)	sentence length without stopwords (mean)
words length without stopwords (mean)	words length (std)
ratio of proper nouns per nouns	words length without stopwords (std)
number of verbs	lemma length (std)

words, syntactic features (sentence depth...), semantic features (polysemy index) or pragmatic features (incidence of connectives) do not play an important role.

3 Approaches

In this section we present the approaches we have followed to perform the stance classification task.

3.1 Language Model Approach

The Language Model (LM) approach uses BERTeus [2] for Basque (ixa-ehu/berteus-base-cased) and BETO [7] for Spanish (dccuchile/bert-base-spanish-wwm-uncased), both downloaded from HuggingFace [23].

For the experiments, we have truncated the texts with more than 200 tokens and padded, the shorter with zeroes. We have added two tokens to mark the beginning and the end of the sequence to each input text, [CLS] and [SEP] respectively. We have applied a pre-processing step besides the standard byte-pair encoding. This pre-processing step consists on segmenting the hashtags, replacing '&amp;' with '&', removing trailing white-spaces and finally converting the text to lowercase only in Spanish. We have used the PyTorch framework to create our model. We have probed with two sequential models on top of BERT:

- A dropout layer to fight overfitting. The dropout probability was set equal to 0.1. On top of the dropout Layer, we have added a linear layer and sigmoid activation function. The input dimension of the linear layer was 768 and the output 3 (equal to the number of classes).
- A linear layer, ReLU activation function and linear layer model. The input dimension of the first linear was 768 and the output 50, and the input dimension of the second linear was 50 and the output 3 (equal to the number of classes).

For each of the outputs, we have used the cross-entropy loss function.

To train the model, we have split the training data into 80 % for train and 20 % for validation. The training batch size was made equal to 32 and the model was trained for 10 epochs using early stopping technique. The best result in the validation data was obtained after running 9 epochs in Basque and 6 epochs in Spanish, setting the tweet length to 200, and the learning rate to 5e-06 with Linear-ReLU-Linear sequential model and the Adam optimizer [13]. We have done these experiments in the Google Colaboratory framework.

Regarding the evaluation, the metric we use is $F1$ Macro, the one used in the VaxxStance shared task and proposed by Mohammad et al. [17] for the SemEval 2016 task on Stance Detection. This metric reports the $F1$ macro-average score of FAVOR and AGAINST classes (although the NONE class is also represented in the data).

Table 3. LM results in the training and validation data ($F1$ Macro)

Setting	**Basque**	**Spanish**
LM-train	**0.75**	**0.80**
LM-validation	**0.62**	**0.71**

In Table 3 we present the results in the training and validation data. As we can see, the results seem to be competitive and they are higher in Spanish than in Basque.

3.2 Approach based on Linguistic Features and Machine Learning

The second approach consists on the use of linguistic and stylistic plus a classical machine learning classifier.

Obtaining linguistic features First, in order to get the linguistic and stylistic features, we have used MultiAzterTest [5], but, in the case of Spanish we have added more features: descriptive+, advanced morpho-syntactic, named entities, social media and abusive terms. The descriptive+ features include indicators about number of words and sentence per tweet, numerical expressions and punctuation marks. The advanced morpho-syntactic features take into account the subcategories of the PoS. The entity types considered are person, location, organisation and miscellaneous. The social media features measure mentions, hashtags, stretched words and emojis. The abusive words rely on HurtLex [4], the multilingual lexicon of words to hurt and in this case we take all the categories contained in the lexicon together. This new version of the tool is called MultiAzterTest-Social (MATS). Some of the new features are inspired by Fersini et al. [9], but others are based on other readability assessment works e.g. *ErreXail* [11]. In total, we have analysed 125 features for Basque and 246 for Spanish.

Selecting the classifier The second step is to choose a classifier. As we are training the system for the VaxxStance shared task, henceforth, we will call MATS-VaxxStance the adaptation created for this task. We have tested the Sequential Minimal Optimization (SMO) [19] classifier with different feature selection according to InfoGain: 125, 75, 50, 25 and Top10 features (Table 2) with the aim of seeing if feature reduction can help, due to the fact that many features highly correlate. We have used 10 fold cross-validation.

Table 4. MATS-VaxxStance results in the training data ($F1$ Macro)

Method	**Basque**	**Spanish**
MATS-VaxxStance-SMO-All	**0.42**	**0.66**
MATS-VaxxStance-SMO-125	**0.42**	0.65
MATS-VaxxStance-SMO-75	0.41	0.61
MATS-VaxxStance-SMO-50	0.34	0.61
MATS-VaxxStance-SMO-25	0.26	0.59
MATS-VaxxStance-SMO-Top10	0.27	0.56

In Table 4 we present the results ($F1$ macro) of these experiments. Contrary to what happens in readability assessment [5], feature selection and feature reduction seem not to be competitive in this task and we have decided to use all the features in this exploratory work.

3.3 Use of Contextual information

For the contextual evaluation setting, we have decided to use only the information of the user. Inspired by the “one sense per discourse” idea by Gale et al. [10], which was successfully implemented for named entities [3], we have decided to apply the “one stance per *tuiter@lari*” (OSPT) idea. That is, we take for granted that each user (*tuiter@lari*) has (in a short period of time) the same opinion about a topic.

So, for each user, we take the most predicted label by the system and apply it to the rest of its tweets. In the case of tie, we apply the favor label. In Algorithm 1, we present the OSPT algorithm.

4 Results in test data

In this section we present the results obtained in the test data for the textual and the contextual evaluation settings as provided by the organisers.

Looking at the results of the textual setting (Table 5), we see that the LM approach gets better results in the $F1$ macro than the linguistic features plus SMO (MATS-Vaxxstance) in both Basque and Spanish. Moreover there is a big difference between them: almost 16 points in Basque and 10 in Spanish. It is also remarkable that the LM approach works much better in Spanish than in

Algorithm 1: “One stance per *tuiter@lari*” (OSPT) algorithm

```
for each tweet do
    userid=Get the owner id of the tweet ;
    countusertweets=Count the number of tweets from that user ;
    if countusertweets $\geq 2$ then
        countuserfavortweets=counts predicted user tweets as favor;
        countuseragainsttweets=counts predicted user tweets as against;
        if countuserfavortweets == 0 and countuseragainsttweets == 0 then
            pass;
        else
            if countuserfavortweets $\geq$ countuseragainsttweets then
                changeallusertweets(userid,'FAVOR');
            else
                changeallusertweets(userid,'AGAINST');
            end
        end
    end
end
```

Basque (around 24 points of difference). There are also important differences in the retrieval of *Against* and *Favor* instances, except for the case of the MATS-VaxxStance in Basque.

Table 5. Results in the test data (textual)

Lang.	Method	Against	Favor	$F1$ macro	Ranking
EU	LM	48.23	52.25	**50.24**	3
EU	MATS-VaxxStance	34.38	34.18	34.28	4
ES	LM	66.67	81.53	**74.10**	3
ES	MATS-VaxxStance	56.47	71.60	64.04	5

Concerning the contextual setting (Table 6), we obtain mixed results. In Spanish, the LM+OSPT approach obtains the highest $F1$ macro while in Basque MATS-VaxxStance+OSPT performs better, although the between approaches difference is insignificant. In Spanish, however, the differences are bigger (9 points). Regarding the retrieval of the *Against* and *Favor* instances, there are also big differences, except for the case of the Spanish LM.

As we have only applied the “one stance per *tuiter@lari*” in the contextual evaluation, we can see which is the consequences of applying this heuristic. In Basque, applying the heuristic in the LM approach worsens the results, while they are improved in the Spanish LM. Regarding the MATS-VaxxStance approach, it helps in both languages. This lead us to think that OSPT gives consistency to the MATS-VaxxStance results.

Table 6. Results in the test data (contextual)

Lang.	Method	Against	Favor	$F1$ macro	Ranking
EU	LM+OSPT	16.36	56.06	36.21	4
EU	MATS-VaxxStance+OSPT	25.40	48.03	**36.72**	3
ES	LM+OSPT	78.77	79.84	**79.31**	3
ES	MATS-VaxxStance+OSPT	63.93	77.17	70.55	5

All results considered, we see that the language models, although they do not offer any interpretability and are far from perfect in this task, are more competitive than the linguistic information. We also want to point out that a corpus and resource analysis will be necessary to know why the results obtained in Spanish are better that the ones in Basque.

5 Conclusion and Future Work

In this paper we have presented the participation of the MultiAzterTest team in the VaxxStance@IberLef 2021 shared task. We have participated in the two evaluation settings (textual and contextual) of the close track where we have presented two approaches: the first is based on well-known languages models, exactly BERTeus for Basque and BETO for Spanish; and the second approach is based on linguistic and stylistic features provided by the open source tool MultiAzterTest plus SMO as classifier. To integrate the contextual features, inspired by the "one sense per discourse" idea and we have created the "one stance per tuiter@lari" heuristic, where if a user has mainly an opinion, we apply that label to the rest of its tweets. Regarding the results, the approach based on language models obtains in general better results and the results got for Spanish are better than those for Basque.

As this is a mainly exploratory work, there is a lot of work to do. Regarding MultiAzterTest-Social more features need to be integrated in Basque. Besides, more classifiers can be tested e.g. Random Forest [6], Simple Logistics [16]... Combinations of both approaches can also be carried out and ways to integrate the remaining contextual approaches can be explored. It is also necessary to make an analysis of the resources used in order to understand the differences in the results in Basque and Spanish. Finally, from a linguistic point of view it would be very interesting to make a feature analysis of favour and against stances to see which are the strategies which are used, if they differ or not.

Acknowledgments

We acknowledge the following projects: DeepText (KK-2020/00088), DeepReading RTI2018-096846-B-C21 (MCIU/AEI/FEDER, UE), BigKnowledge for Text Mining, BBVA and IXA taldea, A motako ikertalde finkatua (IT1343-19).

References

1. Agerri, R., Centeno, R., Espinosa, M., Fernandez de Landa, J., Rodrigo, A.: Vaxxstance@iberlef 2021: Going beyond text in crosslingual stance detection. Procesamiento del Lenguaje Natural 67(0) (2021)
2. Agerri, R., San Vicente, I., Campos, J.A., Barrena, A., Saralegi, X., Soroa, A., Agirre, E.: Give your text representation models some love: the case for basque. In: Proceedings of The 12th Language Resources and Evaluation Conference. pp. 4781–4788 (2020)
3. Barrena, A., Agirre, E., Cabaleiro, B., Penas, A., Soroa, A.: "one entity per discourse" and "one entity per collocation" improve named-entity disambiguation. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. pp. 2260–2269 (2014)
4. Bassignana, E., Basile, V., Patti, V.: Hurtlex: A multilingual lexicon of words to hurt. In: 5th Italian Conference on Computational Linguistics, CLiC-it 2018. vol. 2253, pp. 1–6. CEUR-WS (2018)
5. Bengoetxea, K., Gonzalez-Dios, I.: MultiAzterTest: a Multilingual Analyzer on Multiple Levels of Language for Readability Assessment. Manuscript from author (2021)
6. Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001)
7. Cañete, J., Chaperon, G., Fuentes, R., Ho, J.H., Kang, H., Pérez, J.: Spanish Pre-Trained BERT Model and Evaluation Data. In: PML4DC at ICLR 2020 (2020)
8. Cignarella, A.T., Lai, M., Bosco, C., Patti, V., Rosso, P.: Overview of the evalita 2020 task on stance detection in italian tweets (sardistance). Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020). CEUR-WS. org (2020)
9. Fersini, E., Nozza, D., Boifava, G.: Profiling italian misogynist: An empirical study. In: Proceedings of the Workshop on Resources and Techniques for User and Author Profiling in Abusive Language. pp. 9–13 (2020)
10. Gale, W.A., Church, K., Yarowsky, D.: One sense per discourse. In: Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992 (1992)
11. Gonzalez-Dios, I., Aranzabe, M.J., Díaz de Ilarraza, A., Salaberri, H.: Simple or complex? assessing the readability of basque texts. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. pp. 334–344. DCU and ACL, Dublin, Ireland (Aug 2014), https://www.aclweb.org/anthology/C14-1033
12. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter **11**(1), 10–18 (2009)
13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Koehrsen, W.: Feature Selector: Simple Feature Selection in Python. https://github.com/WillKoehrsen/feature-selector (2019)
15. Lai, M., Cignarella, A.T., Farías, D.I.H., Bosco, C., Patti, V., Rosso, P.: Multilingual stance detection in social media political debates. Computer Speech & Language **63**, 101075 (2020)
16. Landwehr, N., Hall, M., Frank, E.: Logistic model trees **95**(1-2), 161–205 (2005)
17. Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., Cherry, C.: Semeval-2016 task 6: Detecting stance in tweets. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). pp. 31–41 (2016)

18. Montes, M., Rosso, P., Gonzalo, J., Aragón, E., Agerri, R., Ángel Álvarez Carmona, M., Álvarez Mellado, E., de Albornoz, J.C., Chiruzzo, L., Freitas, L., Adorno, H.G., Gutiérrez, Y., Zafra, S.M.J., Lima, S., de Arco, F.M.P., (eds.), M.T.: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021. CEUR Workshop Proceedings (2021)
19. Platt, J.: Fast training of support vector machines using sequential minimal optimization. In: Schoelkopf, B., Burges, C., Smola, A. (eds.) Advances in Kernel Methods - Support Vector Learning. MIT Press (1998), http://research.microsoft.com/j̃platt/smo.html
20. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D.: Stanza: A python natural language processing toolkit for many human languages. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 101–108 (2020)
21. Somasundaran, S., Wiebe, J.: Recognizing stances in online debates. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. pp. 226–234 (2009)
22. Taulé, M., Pardo, F.M.R., Martí, M.A., Rosso, P.: Overview of the task on multimodal stance detection in tweets on catalan# 1oct referendum. In: IberEval@ SEPLN. pp. 149–166 (2018)
23. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al.: Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771 (2019)