Information Processing and Management 60 (2023) 103294

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Spanish Corpora of tweets about COVID-19 vaccination for automatic stance detection

Rubén Yáñez Martínez [a], Guillermo Blanco [a,b,c], Anália Lourenço [a,b,c,d,*]

[a] Universidade de Vigo, Department of Computer Science, ESEI-Escuela Superior de Ingeniería Informática, Edificio Politécnico, Campus Universitario As Lagoas s/n 32004 Ourense, Spain
[b] CINBIO, The Biomedical Research Centre, Universidade de Vigo, Campus Univesitario Lagoas-Marcosende, 36310 Vigo, Spain
[c] SING, Next Generation Computer Systems Group, Galicia Sur Health Research Institute (IIS Galicia Sur), SERGAS-UVIGO, Vigo, Spain
[d] CEB, Centre of Biological Engineering, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal

ARTICLE INFO

Keywords:
Social media
Stance detection
Corpus annotation
Semi-supervised learning
Transformer embeddings
Density-based clustering

ABSTRACT

The paper presents new annotated corpora for performing stance detection on Spanish Twitter data, most notably Health-related tweets. The objectives of this research are threefold: (1) to develop a manually annotated benchmark corpus for emotion recognition taking into account different variants of Spanish in social posts; (2) to evaluate the efficiency of semi-supervised models for extending such corpus with unlabelled posts; and (3) to describe such short text corpora via specialised topic modelling.

A corpus of 2,801 tweets about COVID-19 vaccination was annotated by three native speakers to be in favour (904), against (674) or neither (1,223) with a 0.725 Fleiss' kappa score. Results show that the self-training method with SVM base estimator can alleviate annotation work while ensuring high model performance. The self-training model outperformed the other approaches and produced a corpus of 11,204 tweets with a macro averaged f1 score of 0.94. The combination of sentence-level deep learning embeddings and density-based clustering was applied to explore the contents of both corpora. Topic quality was measured in terms of the trustworthiness and the validation index.

1. Introduction

1.1. Background

Today, social media platforms are key in the dissemination and consumption of Health information (Aiello, Renson, & Zivich, 2020; Suarez-Lledo & Alvarez-Galvez, 2021). Social media communities revolving around health and medical topics are growing in number and activity (Chen & Wang, 2021). Most notably, Health crises, such as the COVID-19 pandemic, have stimulated highly dynamic activity on social platforms (Abd-Alrazaq, Alhuwail, Househ, Hamdi, & Shah, 2020; Chawla et al., 2021; Zhou, Xiu, Wang, & Yu, 2021). Therefore, social media platforms have become a new, appealing source of information that offers a unique perspective regarding voluntarily reported health-related information.

The analysis of interpersonal and official communications (e.g. comments on leaving with chronic conditions or posts about crisis

[*] Corresponding author.
E-mail addresses: ruyanez@alumnos.uvigo.es (R.Y. Martínez), guillermo.blanco.gonzalez@uvigo.es (G. Blanco), analia@uvigo.es (A. Lourenço).

https://doi.org/10.1016/j.ipm.2023.103294
Received 21 April 2022; Received in revised form 18 November 2022; Accepted 20 January 2023
Available online 30 January 2023

updates and measures) are key to understanding the needs and concerns of the public, detecting misinformation spreading, and strengthening communication between the public and the authorities. One main aspect of these analyses is the automatic determination of the stance of posts towards specific targets (ALDayel & Magdy, 2021). For example, conversations about rumours, fake news, controversial topics and crises are expected to exhibit contrasting stances (Kumari, Ashok, Ghosal, & Ekbal, 2021, 2022; Roy, Bhanu, Saxena, Dandapat, & Chandra, 2022). Likewise, stance detection is often an intermediate task of citizen sensing (Chawla et al., 2021) and Public Health applications (Kunneman, Lambooij, Wong, Bosch, & Mollema, 2020).

However, research on social media application differs from earlier stance detection works due to the multiplicity of contexts present in social media (ALDayel & Magdy, 2021; Alkhalifa & Zubiaga, 2022). Social posts are short and informal texts and contain plenty of idiosyncratic abbreviations, as well as special tokens, such as hashtags, user mentions, and emojis. Currently, few benchmarking corpora include social media posts, which can degrade the performance of classifiers to be applied to the social domain (Kunneman et al., 2020). Moreover, research to carry out stance detection and build language resources for other languages than English is still limited.

Due to its worldwide prevalence and the strategic role played by social media in crisis management, the COVID-19 pandemic has encouraged research on stance detection in social media for English as well as other affluent languages. Of particular interest to the present work, several infodemiology studies analysed Spanish Twitter conversations about COVID-19 vaccination and described the content that was tweeted and the mood of the users (Herrera-Peco et al., 2021; Santoveña-Casal, Gil-Quintana, & Ramos, 2021). The development and evaluation of automated methods of stance detection would be of practical utility to scale up and extend such studies as well as to address open challenges such as the detection and curbing of misinformation, the improvement of health awareness and well-being, and the minimisation of social media inequalities.

1.2. Motivation and contributions of this paper

This work aims to produce new corpora for automated stance detection in Spanish Twitter and, more specifically, corpora useful for Health applications. So, the first objective is to develop a manually annotated benchmark corpus. The second objective is to evaluate the benefits of using a semi-supervised learning approach for the task of corpus augmentation, i.e. extending the manually annotated corpus with unlabelled tweets. The last objective is to develop an unsupervised topic model for social posts that can mine coherent topics for these short texts and produce a characterisation of the corpus, e.g. regarding the existing stances.

Therefore, the contributions of this work are as follows:

- two new, public semantically annotated corpora to evaluate stance detection classifiers over tweets in Spanish.
- comparative performance study of semi-supervised learning models for short texts and Spanish language.
- a semi-supervised learning model for stance annotation in Spanish social posts about COVID-19 vaccination.
- a topic model combining a pre-trained sentence embedding model (SBERT) and density-based clustering (HDBSCAN) to depict the most meaningful topics of the social corpora.

To the best of the authors' knowledge, the shared task VaxxStance at IberLEF 2021 was the first to address stance detection in Spanish Health-related social posts (Agerri, Centeno, Espinosa, Fernandez De Landa, & Rodrigo, 2021) and no similar benchmarking corpus exists to date. So, the corpora presented in this work are considered a new, valuable resource to keep promoting this line of research in support of practical health applications for Spanish speakers. Most notably, the new corpora include for the first time conversations in both Peninsula/European Spanish and the varieties of Spanish spoken in Latin America, which is of practical interest given the large population of native speakers in America. The conducted semi-supervised learning experiments bring new insights into the performance of trained transformer encoder models in Spanish varieties and the performance of semi-supervised learning algorithms regarding stance detection on short and colloquial texts. Besides being specialised in short texts, the implemented topic modelling approach grants the ability to characterise corpus contents throughout time, i.e. compare the contents of different versions of the corpus, and thus, be able to understand how the developed classifiers "endure" changes in the contexts and contents of the compiled short texts.

The remaining sections of the paper are as follows. Section 2 reviews existing corpora for social media stance detection, the application of semi-supervised learning methods for the semantic annotation of social posts, and the topic modelling of short texts. The methodologies utilised to prepare the tweet corpora on COVID-19 vaccination stance and to perform the analysis of the conversational topics of the tweets are explained in Section 3. Section 4 describes the manually annotated corpus, the augmented corpus and the topic models of both corpora. Section 5 debates the main findings and contributions of the work as well as current limitations. To conclude, Section 6 summarises the present work and proposes future work directions.

2. Related work

The next subsections underline the current research approaches to stance detection in social media posts, the application of semi-supervised learning to social post annotation and the topic modelling of short texts.

2.1. Corpora for stance annotation in social posts

The detection of stance is a well-recognised natural language processing problem with many practical application areas (Küçük &

Fazli, 2020). Nevertheless, existing methods revealed themselves ineffective to deal with the short length, informality, dynamics and idiosyncrasies of social media posts (Al-Ghadir, Azmi, & Hussain, 2021).

The SemEval international workshops (https://en.wikipedia.org/wiki/SemEval) were the first to propose a community challenge on automatic stance detection in social posts (Mohammad, Kiritchenko, Sobhani, Zhu, & Cherry, 2016; Nakov et al., 2016). These challenges released three corpora, i.e. a Twitter corpus targeting six selected topics (i.e. politics, feminism, climate change, atheism and legalization of abortion), a Twitter multi-topic rumour corpus, and a corpus of Reddit posts. More recently, two broader corpora were proposed: the Will-They-Won't-They (WT-WT) is a topic agnostic, large size corpus (Conforti et al., 2020) and the tWT–WT corpus, which describes how the target entities can be used to detect tweet stance (Kaushal, Saha, & Ganguly, 2021).

Finding stance annotated post corpora for other languages than English is difficult. A corpus that covers Twitter trending subjects in the French language was presented at LREC2020 (Evrard, Uro, Hervé, & Mazoyer, 2020). The SardiStance shared task at EVALITA2020 released the first-ever task for stance detection in Italian tweets (Giorgioni, Politi, Salman, Croce, & Basili, 2020). And, two IberEval shared tasks released Spanish tweets corpora, i.e. the TW-10 Referendum Dataset, which covers politics and aims to provide multilingual stance-annotated data in Catalan and Spanish (Zotova, Agerri, Nuñez, & Rigau, 2020), and the VaxxStance corpus, which aims to explore cross-lingual approaches (Agerri et al., 2021).

The present work contributes with new resources for the Spanish language. Similar to VaxxStance, the new corpora include tweet corpora of practical relevance to Health applications. As added value, the new corpora present a broader collection of social conversations, accounting for both Peninsula/European Spanish and the varieties of Spanish spoken in Latin America. Moreover, the present work studied the performance of semi-supervised learning approaches and topic models in combination with language-specific encodings to secure the representativeness of the new resources (a larger volume of posts from a wider range of users) and the ability to extend such resources according to the natural, ever evolving nature of Health discussions.

2.2. Semi-supervised learning for social posts

The manual annotation of a corpus by domain experts is the best option to guarantee the high quality of the resource, but it is too laborious and time-consuming. In particular, the manual annotation of social media corpora is quite challenging, since the number of posts for controversial or timely relevant topics can scale up to huge numbers very quickly.

Semi-supervised methodologies have been proposed as a valid approach for extending manually annotated corpora with unlabelled data, reducing human effort in annotation while enabling the automatic labelling of a large corpus. Specifically, several works have proposed semi-supervised approaches to tweet stance classification (Darwish, Stefanov, Aupetit, & Nakov, 2019; Dutta, Caur, Chakrabarti, & Chakraborty, 2022; Giasemidis, Kaplis, Agrafiotis, & Nurse, 2020). Moreover, the WNUT-2020 workshop proposed a specific task, i.e. semi-supervised learning for the identification of informative COVID-19 English tweets. Here, one of the participating teams reported noticeable performance improvements in their classifiers by using a pseudo-labelling scheme (Sancheti, Chawla, & Verma, 2020). Other works describe the extension of an Arabic sentiment corpus using the self-learning technique (Al-Laith, Shahbaz, Alaskar, & Rehmat, 2021) and the application of a self-trained approach to the preparation of hate and offensive speech corpora from social media (Alsafari & Sadaoui, 2021). Another research exploited stance-bearing hashtags to extend the corpus provided in the SemEval-2016 task (Misra, Ecker, Handleman, Hahn, & Walker, 2016). On a similar note, a study evaluated the performance of the label propagation and label spreading algorithms over a large dataset of tweets about 72 different rumours (Giasemidis et al., 2020). Of particular interest to the present research, semi-automatic annotation was also applied by the organisers of the IberEval competition to construct the Catalonia Independence Corpus for stance detection in Spanish and Catalan tweets (Zotova et al., 2020).

2.3. Sentence-based topic modelling

Topic modelling methods aim to discover latent semantic structure in large collections of documents. Latent Dirichlet Allocation (LDA) and Probabilistic Latent Semantic Analysis are the most widely used methods, but they are not ideal to model short texts. There is insufficient information on word co-occurrence and statistics among texts cannot fully capture words that relate semantically but rarely co-occur.

Recent works proposed new models that enhance the contextual information of the short texts. The Dirichlet Multinomial Mixture Model-based model (DMM) proposes a modified LDA model that ensures that each text is sampled from a single latent topic. Moreover, the EM-based algorithm and a collapsed Gibbs Sampling algorithm were proposed for DMM to address the sparsity and high-dimensionality of the short texts. One limitation of these models is that they do not account for the lack of enough information on word co-occurrence. Topic models based on graph neural networks have been suggested as an alternative, preferably in combination with pre-trained word embeddings to focus on specific salient topics (Murakami & Chakraborty, 2021; Zhao et al., 2021). Another possible algorithm is the Hierarchical Density-based Spatial Clustering of Application with Noise (HDBSCAN), which can handle data with variable density (Campello, Moulavi, & Sander, 2013). Specifically, the combination of FastText and BERT-alike embeddings for text representation, the Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) for dimension reduction and HDBSCAN for clustering has shown promising results for short text clustering (Meng, Qureshi, & Khandker, 2021; Salmi, Mérelle, Gilissen, van der Mei, & Bhulai, 2022).

In general, pre-trained contextualised representations are achieving important performance gains across various tasks of natural language processing, including semantic similarity detection (Pan et al., 2022). In contrast to the typical individual term embeddings (i.e. words, n-grams or phrases), these embeddings leverage semantic information implicitly contained in the texts to determine the relative semantic similarity of the corresponding words. Currently, there are four main embeddings for sentences or short texts.

Doc2Vec (Le & Mikolov, 2014) is an extension of the popular Word2Vec embedding that offers two representations, i.e. the Distributed Memory version of Paragraph Vector and the Distributed Bag of Words version of Paragraph Vector. The Sentence-Bert (SBERT) calculates embeddings for sentences and paragraphs based on a pre-trained BERT network and Siamese network structures (Reimers & Gurevych, 2019). The pre-trained word embeddings of InferSent are GloVe vectors and enable GRUs, LSTMs, and BiLSTMs encoder architectures (Conneau, Kiela, Schwenk, Barrault, & Bordes, 2017). Finally, the Universal Sentence Encoder combines two encoder models, i.e. transformer and deep averaging networks (Cer et al., 2018). So far, only the SBERT sentence embeddings have been trained in the Spanish language.

3. Materials and methods

This section describes the methodology implemented to produce and analyse the semantically annotated corpora. Fig. 1 outlines the main steps of the workflow.

3.1. *Tweets retrieval*

The Tweepy Python library was applied to retrieve the tweets about COVID-19 vaccination and the corresponding user metadata (Roesslein, 2020). The search pursued public tweets, written in the Spanish language, that contained hashtags and keywords usually associated with COVID-19 and its vaccines. Specifically, the search queries included the common names of the disease and the names of the vaccines with most media exposure: #coronavirus, #coronavirusespaña, #coronavirusmadrid, #COVID19, "coronavirus", "cov-id-19″, "Pfizer", and "vacuna coronavirus", "vacuna", "inyec-", "incocul-", "Pfizer", "Moderna", "biontech", "AstraZeneca", "vacunaoxford", "oxford", "Johnson & Johnson", "Novavax", and "Sputnik". Tweets were compiled from March 1st 2020 to January 4th 2022.

As described in Fig. 2, this retrieval process returned 2833,596 tweets that were then examined. The pycld2 library was applied to verify the language of the collected tweets and filter out those that were not in Spanish. Likewise, retweets and duplicate tweets were removed, so that the dataset only contained unique tweets posted by their original authors. The final collection accounted for unique tweets written in different Spanish variants, namely Peninsula/European Spanish and the varieties of Spanish spoken in Latin America.

A total of 828,741 tweets met the inclusion criteria. From these, a stratified random sample (based on the original time distribution) of 16,664 tweets was created to proceed with corpus creation. Supplementary Material 1 describes the set of 16,664 tweets.

3.2. *Manual annotation guidelines and inter-annotator agreement*

Considering the guidelines proposed for the "SemEval-2016 Task 6: Detecting Stance in Tweets", the present task of stance annotation was formulated as a three-class classification task (Mohammad et al., 2016). That is, from reading the tweet, its stance was labelled:

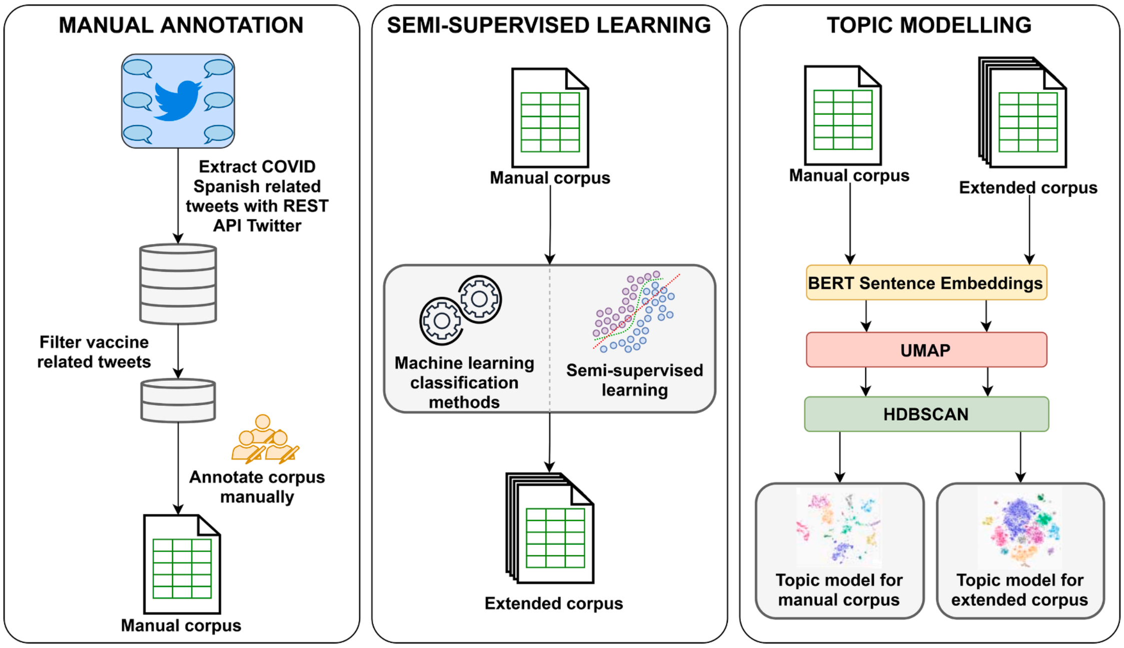

Fig. 1. Schematic of the preparation of the tweets corpora on COVID vaccine stance and further topic modelling.

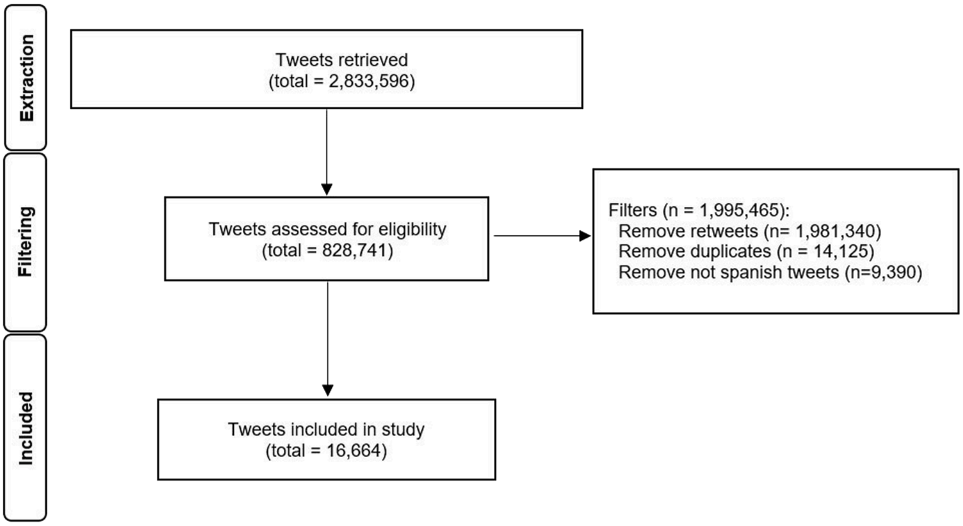

Fig. 2. Search and process of tweets about COVID-19 vaccination.

IN FAVOUR, when the tweeter supports COVID vaccination.

AGAINST, when the tweeter is against COVID vaccination.

NEITHER, when the tweeter shows a neutral stance towards COVID vaccination.

Supplementary Material 1 also provides examples for each stance label. Many tweets reflect personal testimonials encouraging people to get the vaccine (e.g., “Ayer entrevisté a una persona que ya se ha puesto la vacuna/ Yesterday I interviewed a person who has already had the vaccine” or “ya mi mami está vacunada/My mom is already vaccinated”) or expressions of fear and distrust towards vaccination (e.g., “Claro, si te da un trombo y te mueres, coronavirus ya no te va a dar/Of course, if you get a thrombus and you die, you will not get coronavirus”). Some users try to persuade other people not to vaccinate. (e.g., “VACUNARSE es más peligroso que contagiarse … ESTA MENTIRA TIENE QUE ACABAR/GETTING VACCINATED is more dangerous than getting infected… THIS LIE HAS TO END”). Comments on other people’s opinions, without expressing the tweeter’s stance, were labelled as neither (e.g., “Médico francés, tras un análisis estadístico, considera…/ French doctor, after a statistical analysis, considers…”). Reports of the progression of the pandemics were also labelled as neither (e.g., “India supera las 80.000 muertes/ India exceeds 80,000 deaths”). To ensure the greatest variety of tweets, the RT annotation was excluded to avoid similar text

The annotation process involved three rounds of annotation that led to a final corpus of 2801 manually annotated tweets, presented in Section 4.1. In each round of annotation, each tweet was independently annotated by three native Spanish speakers, as "Against", "In favour", "Neither" or "To discard". In the first round of annotation, the annotators labelled 500 tweets, this round served to create the first rules of annotation guide. The second and the third rounds entailed the annotation of 1000 and 1500 additional tweets, respectively. The inter-annotator agreement (IAA) was calculated using the Fleiss’ kappa score (Fleiss, 1971; Landis & Koch, 1977). Disagreements were discussed to obtain consensual labels as well as to improve the annotation guidelines (in Supplementary Material 2). This annotation process aimed to ensure that multi-faceted stance, rhetorical, and irony were considered adequately. In each round, a new collection of tweets was annotated and thus, ensure that the initial annotation guidelines were valid throughout time and topic changes. The second and third rounds of annotation aimed to ensure consistent interpretation of the colloquial expressions and slang used in conversations taking into account the different language variants. Moreover, the evaluation of hashtags and emoticons as indicators of the post’s stance was also treated in annotator consensus. Tweets were discarded whenever deemed off-topic (e.g., the tweet was about influenza vaccination and did not mention COVID-19) by the majority of the annotators, or a majority consensus could not be reached.

3.3. Semi-supervised learning for corpus annotation

Tweet pre-processing started with the elimination of line breaks and sequences of blank spaces and lowercase conversion using a simple Python script. The Python library clean-text supported further removal of URLs, e-mails, punctuation marks, phone numbers, and non-ASCII characters, and the NLTK library was used to remove Spanish stopwords. Finally, the Spacy-Stanza library was applied to perform token lemmatization.

Several experiments were conducted with the base estimators and semi-supervised methodologies discussed below. The Python Scikit-Learn and Pytorch libraries supported all these experiments. The language-specific BETO transformer model (Cañete et al., 2020) in combination with LNN (Linear Neural Network) and Bi-LSTM (Long Short-Term Memory) architectures was compared with a traditional machine learning approach that trains the algorithms in features retrieved by the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm. Support Vector Machines (SVM) (Cortes, 1995), Naive Bayes (Rennie, Shih, Teevan, & Karger, 2003), and Random Forest (Breiman, 2001) were chosen as base estimators. A grid search was performed to find the optimal hyperparameters for the classification models (Supplementary Material 1).

Three semi-supervised learning methods were evaluated using different ratios of unlabelled data as follows:

- **Self-training**: the classification model is built with different base estimators and different ratios of unlabelled data (2.8k, 5.6k, and 8.4k).
- **Label propagation and spreading**: the topic similarity was favoured over sentiment polarity or other similarity measures (da Silva, Coletta, Hruschka, & Hruschka, 2016). Different ratios of unlabelled data (2.8k, 5.6k, and 8.4k) were explored.
- **Delta-training**: variation of self-training that uses two sets of classifiers, i.e. one set of classifiers is randomly initialised whilst the other classifiers apply pre-trained word embeddings, to reduce error accumulation. Different ratios of unlabelled data (2.8k, 5.6k, and 8.4k) were explored (Jo & Cinarel, 2019).

The analysis of performance is based on 10-fold cross-validation over 90% of the tweet set and the other (randomly selected) 10% are used as a test set. Section 4.2 presents the results of these experiments. Comparison is based on the obtained macro averaged f1 scores. Accuracy, precision, and recall values complement this analysis.

3.4. *Topic modelling for tweet clustering*

As an initial step, the tweets were processed into the SBERT sentence embeddings considering the pre-trained BERT models below:

- **distiluse-base-multilingual-cased-v2**: a model based on DistilBertModel with a 512-dimensional dense vector space.
- **paraphrase-multilingual-mpnet-base-v2**: a model based on XLMRobertaModel with a 768-dimensional dense vector output.
- **paraphrase-multilingual-MiniLM-L12-v2**: a model that uses a 384-dimensional dense vector space to represent sentences and paragraphs.

The application of the Uniform Manifold Approximation and Projection (UMAP) algorithm ensured dimensionality reduction while preserving a significant portion of the three high-dimensional local structures in lower dimensionality (Mcinnes, Healy, & Melville, 2020). The algorithm was parameterized based on the evaluation of the trustworthiness (T) metric (Stasis, Stables, & Hockman, 2016):

$$\mathbf{T}(\mathbf{k}) = 1 - \frac{2}{\mathbf{k}(2\mathbf{n} - 3\mathbf{k} - 1)} \sum_{i=1}^{n} \sum_{\mathbf{j} \in \mathbf{U}_{\mathbf{i}}^{(\boldsymbol{k})}} (\mathbf{r}(\mathbf{i}, \mathbf{j}) - \mathbf{k})$$

Then, the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) was applied to model tweets by topic (Campello et al., 2013). The evaluation of the quality of the model was based on the Density-Based Clustering Validation index (DBCV) which computes the least dense region inside a cluster and the densest region between the clusters (Moulavi, Jaskowiak, Campello, Zimek, & Sander, 2014):

$$DBCV(C) = \sum_{i=1}^{i=l} \frac{|C_i|}{|O|} V_c(C_i)$$

Supplementary Material 1 describes the hyperparametrization conducted at each of the aforementioned steps.

The optimal topic models depict the tweets and their corresponding embeddings, namely the vector showing the predicted topic for each tweet and the array of probabilities. Each row gives the probabilities of the corresponding tweet being part of the different topics. Finally, the class-based TF-IDF (c-TF-IDF) was implemented for the visual analysis of the obtained results, namely to enable the identification of interpretable topics (Özgür, Özgür, & Güngör, 2005). These results are described in Section 4.3.

4. Results

This work contributes with new, publicly available corpora of stance annotation on short, colloquial texts written in Spanish that may be of practical use to the research community, namely those developing health applications that use social media data. Thus, the next sections report the results obtained during the processes of annotation of the two corpora and then explore their contents.

4.1. *Manually annotated corpus*

Three annotators, who are regular users of Twitter and are Spanish native speakers, participated in the labelling of this corpus. Three rounds of annotation led to a corpus of 2801 manually annotated tweets.

The annotation process was multi-round and was based on the analysis of the inter-annotator agreement. The Fleiss kappa score obtained in the first round was 0.536, in the second round was 0.619 and in the last round was 0.725. That is, there was a fair agreement in the first round and a substantial agreement in the last two rounds of annotation (Landis & Koch, 1977). The initial round of annotation brought forward ambiguous cases and disagreements that were discussed to redefine and clarify the annotation guidelines as best as possible. For example, the annotators disagreed when evaluating the stance of the tweets that contained emojis, used irony, used social media slang and jargon, or referred to public figures unfamiliar to them. The improvement of IAA in the next rounds corroborated that the amendment of the annotation guidelines was successful and the final version is clear and specific and may be used as standard annotation guidelines for stance annotation in social posts.

The final corpus contains only tweets for which at least two out of the three annotators agreed on the stance label. Specifically, whenever a consensus was not possible or the tweet was deemed off-topic, it was eliminated. The main cause for persistent disagreement was “multi-faceted” stance, often associated with irony or sarcasm, such as: ”Con un pie en el cementerio y otra cosa ya muchos tienen la vacuna puesta del Covid-19 que les importa / With one foot in the cemetery and something else, many already have the Covid-19 vaccine that matters to them” or “No eres vidente ni muy espabila por las chorradas que dices. Inmunizarte de virus en Verano que son de Invierno. Venga Lumbreras. Ponte la vacuna / You are not a psychic or very smart because of the bullshit you say. Immunize yourself from viruses in Summer that are from Winter. Come smartass. get the vaccine”.

Tables 1 and 2 describe the characteristics of the manually annotated corpus in terms of general contents and per stance label, respectively. For the most part, the source of the tweet was different, which contributes to diversity (in terms of opinions, content phrasing, etc.). On average, there is a 62% chance of a tweet containing a hashtag whereas only 26% of the tweets contain mentions (i.e. replied to or mentioned Twitter users).

The corpus is imbalanced, i.e. the label “neither” is more represented than the other two: 904 in favour, 674 against, and 1223 do not show a stance about COVID-19 vaccination. Regarding tweet contents, corpus distribution per annotation label shows a similar amount of hashtags and mentions for the three labels. The tweets “against” vaccination contain more tokens and those with an undetermined stance are usually shorter. Pre-processing affects the number of tokens ever so slightly. The dehydrated tweets of the manually annotated corpus are available in Supplementary Material 3.

4.2. Semi-supervised corpus

This section reports the construction of an extended version of the manually annotated corpus through a semi-supervised learning approach. Table 3 presents the results of the experiments, i.e. the performance scores obtained by the baseline classifiers and the semi-supervised models for the independent test set, using the manually annotated corpus (MC) and different sets of unlabelled tweets. Label propagation and label spreading do not require a kernel, whereas self-training was evaluated using different kernels. Delta training experiments accounted for two possible sets of deep learning classifiers. Supplementary Material 1 provides further details on such evaluation.

One can observe that the language-specific deep learning architectures outperformed the rest of the models for the manually annotated corpus (i.e. an F1 score of 0.92). The self-training with an SVM kernel was the best performing semi-supervised strategy and the derived SVM and the language-specific deep learning classifiers BETO+LNN and BETO+Bi-LSTM achieved the best performance (i.e. 0.94, 0.92 and 0.93, respectively).

The label propagation and spreading strategies attained poorer results but were similar to those reached by the self-training method with the language-specific deep learning kernel. The delta-training method performs better, most notably, using the Bi-LSTM model. In these experiments, the SVM classifier always outperforms the other classifiers.

Tables 4 and 5 describe the characteristics of the final version of the extended corpus in terms of general contents and contents per stance label, respectively. This corpus contains 11,204 tweets from 5573 unique users, i.e. 5746 tweets in favour of COVID-19 vaccination, 3.553 against, and 1905 do not show a stance. The corpus is imbalanced, i.e. the label “in favour” is more represented than the other two. For the most part, the source of the tweets was different, which contributes to diversity (in terms of opinions, content phrasing, etc.). On average, there is a 78% chance of a tweet containing a hashtag but only 28% of the tweets contain mentions (i.e. replied to or mentioned Twitter users).

Regarding the content of the tweets, the distribution of the corpus by annotation tag shows a different distribution of hashtags and mentions by stance. While "in favour" tweets show a higher number of hashtags, "neither" tweets have almost no mentions. The tweets “against” vaccination contain more tokens and those with an undetermined stance are usually shorter. Pre-processing affects the number of tokens ever so slightly.

The dehydrated tweets of the extended corpus are available in Supplementary Material 3.

Table 1
Descriptive statistics on the manually annotated corpus.

Feature	Value
Number of tweets	2801
Number of tweets “in favour”	904
Number of tweets “against”	674
Number of tweets “neither”	1223
Number of unique users	2383
Number of unique hashtags	1688
Number of unique mentions	1041
Number of tokens before processing	14,369
Number of tokens after processing	10,352
Average of tokens per tweet before processing	20
Average of tokens per tweet after processing	17

Table 2
Descriptive statistics on the labels of the manually annotated corpus.

Feature	In Favour	Against	Neither
Number of unique users	850	551	1060
Number of unique hashtags	657	441	837
Number of unique mentions	396	368	371
Number of tokens before processing	6185	5722	6924
Number of tokens after processing	4433	3764	5565
Average of tokens per tweet before processing	21	23	18
Average of tokens per tweet after processing	17	18	15

Table 3
Summary of performance scores for the base classifiers and those using the semi-supervised methods on the independent test set. The manually annotated corpus (MC) is augmented with different ratios of unlabelled data (2.8k, 5.6k, and 8.4k). The performance is reported in terms of the macro averaged f1 score (F1).

		F1 score				
		SVM	Naive Bayes	Random Forest	BETO+LNN	BETO+Bi-LSTM
Baseline	**MC**	0.69	0.67	0.6	0.92	0.92
Self-training+SVM	MC+2.8k	0.86	0.83	0.71	0.88	0.88
	MC+5.6k	0.9	0.87	0.76	0.9	0.9
	MC+8.4k	**0.94**	0.9	0.84	**0.92**	**0.93**
Label propagation	MC+2.8k	0.79	0.76	0.67	0.86	0.84
	MC+5.6k	0.83	0.8	0.69	0.86	0.87
	MC+8.4k	0.84	0.81	0.68	0.86	0.87
Label spreading	MC+2.8k	0.8	0.8	0.62	0.8	0.78
	MC+5.6k	0.84	0.81	0.63	0.81	0.8
	MC+8.4k	0.82	0.79	0.58	0.75	0.76
Self-training- BERT+LNN	MC+2.8k	0.81	0.78	0.72	0.83	0.83
	MC+5.6k	0.87	0.82	0.81	0.85	0.83
	MC+8.4k	0.78	0.81	0.81	0.84	0.84
Self-training-BERT+Bi-LSTM	MC+2.8k	0.82	0.79	0.75	0.83	0.85
	MC+5.6k	0.87	0.8	0.8	0.85	0.86
	MC+8.4k	0.87	0.79	0.76	0.8	0.8
Delta training-BERT+LNN	MC+2.8k	0.80	0.78	0.72	0.83	0.84
	MC+5.6k	0.84	0.78	0.72	0.79	0.82
	MC+8.4k	0.83	0.76	0.76	0.77	0.78
Delta training-BERT+Bi-LSTM	MC+2.8k	0.81	0.79	0.75	0.84	0.84
	MC+5.6k	0.87	0.81	0.79	0.83	0.81
	MC+8.4k	0.88	0.79	0.80	0.80	0.79

Table 4
Descriptive statistics on the extended corpus.

Feature	Value
Total of tweets	11,204
Number of tweets "in favour"	5746
Number of tweets "against"	3553
Number of tweets "neither"	1905
Number of unique users	5573
Number of unique hashtags	4474
Number of unique mentions	3182
Number of tokens before processing	30,424
Number of tokens after processing	25,261
Average of tokens per tweet before processing	20
Average of tokens per tweet after processing	17

4.3. Topic mining

The contents of the manual corpus and the extended corpus were further explored. S-BERT models at optimal parameters, followed by a UMAP reduction, supported tweet encoding. The HDBSCAN models at optimal parameters were rendered in terms of the biggest c-TF-IDF values.

Table 6 displays the topics generated for the manually annotated dataset and the extended dataset. For the manually annotated collection, the HDBSCAN model returned 2 clusters for "In favour" tweets as well as for the "Against" tweets, and 6 clusters for tweets labelled as "Neither". For the extended corpus, there is a similar number of clusters for "In favour" and "Against" tweets, and 4 clusters

Table 5
Descriptive statistics on the labels of the extended corpus.

Feature	In Favour	Against	Neither
Number of unique users	2591	1885	1356
Number of unique hashtags	2422	1339	1297
Number of unique mentions	1396	1533	532
Number of tokens before processing	16,004	14,779	8488
Number of tokens after processing	14,601	9619	7061
Average of tokens per tweet before processing	20	22	18
Average of tokens per tweet after processing	18	18	16

for the “Neither” tweets. In all cases, the cluster labelled “Topic -1” denotes the identified noise.

Looking into the “In favour” collections, topics draw attention to the vaccination of certain population groups. Tweets that celebrate the vaccination of elder people, using terms to subjects (“mama/mom”, “papa/dad”) and positive emotions (“alegría/happiness”, “feliz/happy”), and pregnant women, using terms to denote targets (“embarazo/pregnancy”, “gestacion/gestation”). For example, “💉#YoMeVacuno: Recuerda que esta semana continúa la vacunación contra #COVID19, tanto para dosis de refuerzo como para rezagados desde los 3 años, segundas dosis y embarazadas a partir de la semana 16 de gestación. 😷La pandemia no ha terminado. ¡Es importante seguir cuidándonos!/ 💉 #IGetVaccinated: Remember that this week vaccination against #COVID19 continues, both for booster doses and for those left behind from the age of 3, second doses and pregnant women from week 16 of gestation. 😷 The pandemic is not over. It is important to continue taking care of ourselves!”.

These topics also express the belief that vaccine can protect us from the disease and raising awareness is key to fight misinformation through actions (“proteger/protect”, “evitar/avoid”). For example, “Cuidado con lo que ven en las redes sociales sobre la vacuna 💉. ¡Mucho es FALSO! ❌ Lo cierto e importante es que al aplicársela, protegerán a su familia y a la comunidad. #JuntosSanamos #Covid19 #Coronavirus #Vacuna #Vaccine #pandemia/Be careful what you see on social media about the vaccine 💉. Much is FALSE! ❌ The truth and important thing is that by applying it, they will protect their family and the community. #TogetherWeHeal #Covid19 #Coronavirus #Vaccine #pandemic”. Vaccination goals, with direct mentions (“calendario/schedule”) and goals (“#dosisderefuerzo/#boosterdose”, “dosis/dose”), are often indicated. For instance, “Parece que muy pronto tendremos una vacuna segura y efectiva para la #COVID19, ¡y nosotros estamos listos! A final de año tendremos 520 millones de jeringuillas: ¡podríamos dar una vuelta y media al mundo con ellas!/ It seems that we will have a safe and effective vaccine for #COVID19 very soon, and we are ready! By the end of the year we will have 520 million syringes: we could go around the world one and a half times with them!”.

The “Against” groups clearly express their opposition to vaccination campaigns in both collections of tweets. Topics point to the

Table 6
Results of the topics extracted from the manually and extended corpora.

		Tweets	Users	T	DBCV	Topic	Words
Manual annotated corpus	In favour	904	850	0.91	2.66e-04	−1	salud, gracias, noticia, publico, pfizer, pais
						0	papa, mama, salir, esperar, alegría, querer, feliz
						1	proteger, vacunacovid, enfermedad, virus, evitar
	Against	674	551	0.90	1.12E-05	−1	muerte, efecto, querer, plandemia, astrazeneca, adverso
						0	muerto, plandemia, yonomevacuno, plaga, politico, despertar
						1	antivacuna, gente, romper, menos, aviso, momento, solución
	Neither	1223	1060	0.92	1.87E-01	−1	primero, dosis, pfizer, salud, mundo, millón
						0	rusia, sputnik, sputnikv, nuevo, anuncio, putin, efectividad
						1	efecto, adverso, moderna, pfizer, farmacéutica
						2	poder, llegar, dia, isabel, tendencia, reina
						3	astrazeneca, trombo, trombosis, caso, relacion, cerebral, vinculo
						4	pandemia, oms, ola, mes, duro, crisis, mundial, alertar, contagio
						5	vacunacovid, muerte, contagio, inmunizado, campaña, informacion
Extended corpus	In favour	5746	2591	0.96	1.81E-01	−1	#dosisderefuerzo, porcentaje, objetivo, actualizar, campaña, octubre, total
						0	embarazado, gestacion, astrazeneca, pfizer, sinovac, población, deg
						1	noticia, calendario, pfizer, #vacunacovid, #dosisderefuerzo, dosis
	Against	3553	1885	0.95	3.61E-01	−1	surcoreano, arrestado, fabricar, manipulación, crimen, empresa, #totalnormalidad
						0	#covidhoax, fraude, plandemic, #falsapandemia, falso, plandemer, #yonomevacuno
						1	plandemia, #SLR, pandemia, gate, gates, #censuraenlar, NOM
	Neither	1905	1356	0.92	7.36E-01	−1	internacional, ensayo, johnson, rusia, oms, reino, unido
						0	chino, wakuchin, #noticiasdejapon, xing, xin, sinophar,
						1	vacunado, muerte, #ultimahora, moderna, europeo, oms, ensayo
						2	radar, rd, new, #radarrd, covax, escasez, ue
						3	rusia, sputnik, sputnikv, putin, ruso, vladimir, presidente

existence of plots and fabricated news about the need for massive vaccination, with direct mentions ("fraude/fraud", "#falsapandemia/#fakepandemic" and "#covidhoax") and derogatory adjectives ("plandemer", "plandemic"), and show great apprehensiveness for serious side effects and even death, using terms to denote actions ("yonomevacuno/I don't vaccinate", "despertar/wake up"). For example, "#CensuraEnLaRed #COVID19 #Plandemia si el CEO de #Pfizer NO SE VACUNA...¿Por qué lo tengo que hacer yo? QUE SE VACUNEN POLITICOS Y EMPRESARIOS FARMACEUTICOS...van a ver como se acaba la "pandemia" en 2 días!!! acá la noticia: https://t.co/O0v/#InternetCensorship #COVID19 #Plandemic if the CEO of #Pfizer DOES NOT GET VACCINATED...Why do I have to do it? LET POLITICIANS AND PHARMACEUTICAL BUSINESSERS GET VACCINATED... they will see how the "pandemic" ends in 2 days!!! here the news: https://t.co/O0v", or "Y que hacemos de ser así? Nos dejamos vacunar? #yonomevacuno El Gobierno Británico anticipa una ola de muertes/lesiones por la vacuna COVID-19/ And what do we do if we are? Do we let ourselves be vaccinated? #Idontvaccinate The British Government anticipates a wave of deaths / injuries from the COVID-19 vaccine".

In the extended corpus, Topic 1 emphasizes the role of certain individuals in these alleged plots, with direct mentions ("Gates") and references to the possibility of new world order ("#SLR", "NOM"). For example, "La gente con PCR (fraude!) positiva de bicho fake con múltiples e incoherentes síntomas. Al final neumonías BACTERIANAS. Ahora con las vacunas la gente se muere de LO DE Siempreee excepto por las vacunas. PIENSAA https://t.co/xmzpQLixMP/ People with PCR (fraud!) positive for fake bug with multiple and inconsistent symptoms. In the end BACTERIAL pneumonias. Now with the vaccines people die of the Always except for the vaccines. THINK https://t.co/xmzpQLixMP" Classifier in topic 1 and in the topic 0 "es una falsa pandemia de covid 19 las vacunas de covid son un fraude #yonomevacuno #plandemia #plandemic #falsapandemia #covidHOAX / It's a fake covid 19 pandemic covid vaccines are a fraud #idontvaccinate #plandemic #plandemic #fakepandemic #covidHOAX".

In both corpora, "Neither" is the most diverse group and discusses an array of issues related to the virus, the vaccines, and social distance.

There is a topic about the use of the Russian vaccine, with direct mentions ("sputnik"," sputnikv"), mentions to actors ("Rusia", "Putin") and using terms to denote motivation ("efectividad/ effectiveness"). For example, "#Sputnik pequeño fragmento del Informe publicado en The Lancet, sobre los resultados de la Fase 3 de la vacuna Sputnik. Efectividad y margen de adversos. Aprobada. #COVID19 #coronavirus #Vacunas Y si se pudieran fabricar aquí?/ "#Sputnik small fragment of the Report published in The Lancet, on the results of Phase 3 of the Sputnik vaccine. Effectiveness and margin of adverse events. approved. #COVID19 #coronavirus #Vaccines What if they could be created here?". A similar topic discusses vaccination in China and Asia, using terms to denote subjects ("chino/chinese", "xing", "xin") and direct mentions to the vaccine ("wakuchin", "sinophar"). For instance, "[WAKUCHIN] #AstraZeneca solicita autorización para aplicar su #vacuna en el #Japón Detalles: #ワクチン #アストラゼネカ #新型コロナウイルス #noticiasnippon #noticiasdejapón #ノティシアスニッポン #coronavirus/"[WAKUCHIN] #AstraZeneca requests authorization to apply its #vaccine in #Japan Details: #ワクチン #アストラゼネカ #新型コロナウイルス #nipponnews #japanesenews #ノティポアン #coronavirusアッ".

Some topics include information about the study of possible side effects of the vaccines, with direct mentions of the vaccines (Astrazeneca, Moderna, Pfizer) as well as the side effects ("trombo/thrombus", "thrombosis/thrombosis", "cerebral/brain"). Other topics cover announcements and statistics about the pandemics, with mentions of targets ("campaña/campaign","oms/who"), derogatory adjectives ("duro/hard"), motivation for vaccination ("ola/wave","mundial/world","contagio/contagion", "muerte/death") and action ("alertar/alert"). For example, "¡LA NOTICIA DEL AÑO! ✋ ■ Vladimir Putin, presidente de Rusia, anuncia que acaba de registrar la primera VACUNA del mundo contra Covid19. Está denominada SputnikV en honor al primer satélite soviético. Forma la inmunidad de hasta 2 años. Actuarán de forma rápida en Rusia.", or "●Sigue #EnDirecto la actualidad de la pandemia La OMS concluye que la vacuna de AstraZeneca aporta más beneficios que riesgos y no cree que aumente los trombos/THE NEWS OF THE YEAR! ✋ ■ Vladimir Putin, President of Russia, announces that he has just registered the world's first VACCINE against Covid19. It is called SputnikV in honor of the first Soviet satellite. Forms immunity up to 2 years. They will act quickly in Russia.", or "●Follow #Live the latest on the pandemic The WHO concludes that the AstraZeneca vaccine provides more benefits than risks and does not believe it will increase thrombi".

Overall, most users show political neutrality, even though political motivation is one of the major arguments against COVID-19 vaccination, along with hesitancy motivated by limited testing. The possibility of death and serious side effects is equally used as an argument by those that are in favour and against vaccination. Besides, the clusters of "neither" stance are usually populated by news and statistics reports, i.e. campaign dates, pandemic statistics or the study of the potential association between vaccines and heart diseases.

5. Discussion

5.1. Principal findings

The application of machine learning methodologies holds a huge potential for aiding in social media data analysis. For example, gaining a better understanding of social media stance about COVID vaccination is key to the early detection of misinformation, vaccine reluctance or decay in vaccine confidence. Yet, the manual and continuous monitoring of social conversations is virtually impossible. Machine learning methodologies can be applied to stance detection, but current research struggles with the lack of benchmarking resources. Notably, little research exists in languages other than English, even affluent languages such as Spanish. So, the main contributions of this work are the high-quality corpora for stance detection in Spanish Twitter, the experiments on semi-supervised corpus annotation, and further corpus topic modelling. To the best of authors' knowledge, these are the first resources prepared to train stance detectors on variants of Spanish and thus, be able to analyse Spanish social media conversations worldwide.

The three-round manual annotation process led to a corpus of over 2k tweets with a substantial IAA. For the most part, the annotation process aimed to ensure that multi-faceted stance, rhetorical, and irony were considered adequately. Each round of annotation managed a different collection of tweets, to make sure that the initial annotation guidelines were valid throughout time, i.e. facing new concerns and arguments as well as different ways of defending similar and opposite stances, and different ways of postulating such arguments. For the most part, the second and third rounds of annotation aimed to ensure consistent interpretation of colloquial expressions and slang used in conversations, mainly Latin American, which were not familiar to the annotators. The evaluation of hashtags and emoticons as indicators of the post's stance was also treated in annotator consensus. There were also cases of very similar tweets, which annotators decided to eliminate for the sake of diversity.

Nevertheless, the process was not optimal for several reasons. For example, it was often the case that the same annotator labelled nearly identical tweets differently due to finding small nuances in the tweets. After discussing the reasons behind such inconsistency, it became clear that the proposed three classes are not exhaustive of all the possible cases. In particular, out of the topic tweets and those tweets whose meaning is not completely understood by the annotator would not fit any of the classes. Conversely, the interpretation of irony or social media argot led to disagreements that did not fit into the three-class scheme either. Moreover, the in favour, against and neither classes are not perfectly dissociated and could be used simultaneously, which generates undesired ambiguity. For instance, the tweet may support the general stance of a referred source and, at the same time, comment or even question part of the content. Therefore, the annotator would have to assess whether the tweet mostly agrees or adds contrasting information to the source tweet, and then decide between in favour or against, respectively. Taking into account all these limitations, whenever a majority consensus could not be reached, tweets were eliminated because the text was deemed to offer little context to decide upon the right label.

Considering the inherently dynamic nature of social media as a public source of information, and the low cost of retrieving unlabelled posts, the present study also proposes the application of a semi-supervised learning strategy to augment the manually annotated corpus and thus, generate a larger, high-quality resource for stance detection in social posts. Specifically, the evaluation of a SOTA language-specific embedding combined with deep learning architectures and the delta-training variant of self-training bring novel insights. Interestingly, the self-training with SVM consistently outperformed the other models. Results suggest that the language-specific embeddings may be useful but an embedding trained on the topic (i.e. COVID-19 vaccination) would probably yield better results. That is, how well the embeddings capture the conversation contexts is the key to their success on short text learning tasks.

Further segmentation of corpora into topics was deemed beneficial for gaining a better understanding of the contents, as an added means of quality inspection and an automated means of characterisation for release and community consumption. The application of a density-based clustering algorithm using SBERT embeddings and a suitable dimension reduction algorithm enabled the depiction of semantically meaningful topics for such short texts.

5.2. Limitations

Using the public Twitter API is a limitation since it restricts the number of requests that can be made per day and is only able to retrieve tweets from the past 14 days. Another limitation is the study of geolocation, i.e. comparing what (and how) contents are being discussed in Spanish-speaking countries, which could rely on voluntarily submitted data (usually available for a small fraction of the users) or would require the training of a specialised model. Likewise, the study could also include the co-official languages of Spain (namely, Catalan, Basque and Galician). These extensions could be useful to explore language idiosyncrasies and boost multi-lingual research, even though it would be necessary to adjust the annotation guidelines to those new objectives. Moreover, it would call for the participation of a more diversified group of native speakers.

Regarding the proposed experiments, semi-supervised learning is of benefit to diversifying the corpora (in terms of phrasing, use of idiosyncratic expressions, and presence of irony, among others), but the costs of training and fine-tuning the models should be taken into account. Likewise, one should not forget the specificities of social media. Often, a few users generate a large share of the posts and vice versa. As tweets from the same author are likely to reflect the writing style, vocabulary and other communicative behaviour of the individual, models could easily overfit the characteristics of given users. This is not the case with the presented corpora. In the manually labelled corpus, there are 2383 users and 2801 tweets while, in the extended corpus, there are 5573 authors for 11,204 tweets. That is a user in these corpora generated, on average, no more than 2 tweets. However, if one would wish to enlarge the extended corpus even further, a random selection of tweets may not be the best option.

On a similar note, the unlabelled tweets used by the semi-supervised learning strategies were drawn randomly to prevent any constrain on the underlying classifier. However, a more sophisticated selection policy could be of interest if one can ensure the main classifier properties.

6. Conclusions

This paper introduces two new, high-quality tweet corpora for stance detection. Specifically, the new corpora cover stance on Spanish tweets about COVID-19 vaccination. Stance detection is considered a pivotal task in practical analyses of social media interactions and Health is a major topic of discussion on these platforms, especially since the burst of the pandemic. While some corpora exist for training stance detectors for English, classifiers for other languages are scarce. Therefore, the new corpora can aid in the advancement of stance detection research and, by doing so, in the implementation of tools of practical utility for Public Health.

The evaluation of semi-supervised learning strategies for corpus annotation is another relevant contribution of this work. Since social media users are quite active and dynamic in their interactions, the costs of building manually annotated corpora can easily become unmanageable. Especially, during crises, when these platforms become preferred vehicles to search for and share information. The experiments conducted here show that semi-supervised learning can be successfully applied to extend social, short text corpora and thus, account for class representativeness as well as semantic drifts and contents diversity, in a timely and cost-effective way. Similarly, the combination of sentence embeddings and density-based clustering showed a good ability to model the contents of social posts. This is of practical utility to obtain a first impression of corpus contents as well as to uncover possible biases, underrepresentation of topics and other issues regarding corpus annotation.

The corpus annotations (including the tweet ids), the annotation scheme and details on the performed experiments are publicly available in the Supplementary Materials.

CRediT authorship contribution statement

Rubén Yáñez Martínez: Methodology, Investigation, Software, Writing – original draft, Writing – review & editing. **Guillermo Blanco:** Methodology, Investigation, Software, Writing – original draft, Writing – review & editing. **Anália Lourenço:** Conceptualization, Supervision, Validation, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

None to declare.

Data availability

Data will be made available on request.

Acknowledgements

This study was supported by MCIN/AEI/ 10.13039/501100011033 under the scope of the CURMIS4th project (Grant PID2020–113673RB-I00), the Consellería de Educación, Universidades e Formación Profesional (Xunta de Galicia) under the scope of the strategic funding of ED431C2018/55-GRC Competitive Reference Group, the "Centro singular de investigación de Galicia" (accreditation 2019–2022), and the Portuguese Foundation for Science and Technology (FCT) under the scope of the strategic funding of UIDB/04469/2020 unit. SING group thanks CITI (Centro de Investigación, Transferencia e Innovación) from the University of Vigo for hosting its IT infrastructure. Funding for open access charge: Universidade de Vigo/CISUG.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ipm.2023.103294.

References

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378–382. https://doi.org/10.1037/h0031619

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159. https://doi.org/10.2307/2529310

Cortes, C. (1995). *Support-Vector Networks, 20*, 273–297.

Breiman, L. (2001). *Random Forests, 45*, 5–32.

Rennie, J. D. M., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the poor assumptions of naive Bayes text classifiers. In *ICML'03: Proceedings of the Twentieth International Conference on International Conference on Machine Learning*.

Özgür, A., Özgür, L., & Güngör, T. (2005). Text categorization with class-based and Corpus-based keyword selection. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), *3733 LNCS*, 606–615. https://doi.org/10.1007/11569596_63.

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. *Lecture notes in computer science (Including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (pp. 160–172). https://doi.org/10.1007/978-3-642-37456-2_14. *7819 LNAI*(PART 2).

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In *, 4. 31st International Conference on Machine Learning, ICML 2014* (pp. 2931–2939). https://doi.org/10.48550/arxiv.1405.4053

Moulavi, D., Jaskowiak, P.A., Campello, R.J.G.B., Zimek, A., & Sander, J. (2014). *Density-based clustering validation*.

da Silva, N. F. F., Coletta, L. F. S., Hruschka, E. R., & Hruschka, E. R., Jr. (2016). Using unsupervised information to improve semi-supervised tweet sentiment classification. *Information Sciences*, 348–365. https://doi.org/10.1016/j.ins.2016.02.002. *355–356*.

Misra, A., Ecker, B., Handleman, T., Hahn, N., & Walker, M. (2016). *NLDS-UCSC at SemEval-2016 Task 6: A semi-supervised approach to detecting stance in tweets. Proceeding*, 420–427.

Mohammad, S.M., Kiritchenko, S., Sobhani, P., Zhu, X., & Cherry, C. (2016). *SemEval-2016 Task 6: Detecting Stance in Tweets*. 31–41. http://alt.qcri.org/semeval2016/task6/.

Nakov, P., Rosenthal, S., Kiritchenko, S., Mohammad, S. M., Kozareva, Z., Ritter, A., Stoyanov, V., & Zhu, X. (2016). Developing a successful SemEval task in sentiment analysis of Twitter and other social media texts. *Language Resources and Evaluation, 50*(1), 35–65. https://doi.org/10.1007/s10579-015-9328-1

Stasis, S., Stables, R., & Hockman, J. (2016). Semantically controlled adaptive equalisation in reduced dimensionality parameter space. *Applied Sciences, 6*(4), 116. https://doi.org/10.3390/app6040116

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017). *Supervised learning of universal sentence representations from natural language inference data.* https://arxiv.org/abs/1705.02364.

Darwish, K., Stefanov, P., Aupetit, M., & Nakov, P. (2019). Unsupervised user stance detection on Twitter. In *Proceedings of the 14th International AAAI Conference on Web and Social Media, ICWSM 2020* (pp. 141–152). https://arxiv.org/abs/1904.02000v3.

Jo, H., & Cinarel, C. (2019). Delta-training: Simple semi-supervised text classification using pretrained word embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 3458–3463).

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference* (pp. 3982–3992). https://doi.org/10.18653/v1/d19-1410

Abd-Alrazaq, A., Alhuwail, D., Househ, M., Hamdi, M., & Shah, Z. (2020). Top concerns of tweeters during the COVID-19 pandemic: Infoveillance study. *Journal of Medical Internet Research, 22*(4), e19016. https://doi.org/10.2196/19016

Aiello, A. E., Renson, A., & Zivich, P. N. (2020). Social media– and internet-based disease surveillance for public health. *Annual Review of Public Health, 41*(1), 101–118. https://doi.org/10.1146/annurev-publhealth-040119-094402

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., & Pérez, J. (2020). Spanish pre-trained Bert model and evaluation data. Practical ML for Developing Countries Workshop @ICLR 2020. https://github.com/josecannete/spanish-corpora.

Conforti, C., Berndt, J., Pilehvar, M.T., Giannitsarou, C., Toxvaerd, F., & Collier, N. (2020). *Will-they-won't-they: A very large dataset for stance detection on Twitter.* 1715–1724. https://doi.org/10.18653/v1/2020.acl-main.157.

Evrard, M., Uro, R., Hervé, N., & Mazoyer, B. (2020). *French Tweet Corpus for automatic stance detection.* 11–16.

Giasemidis, G., Kaplis, N., Agrafiotis, I., & Nurse, J. R. C. (2020). A semi-supervised approach to message stance classification. *IEEE Transactions on Knowledge and Data Engineering, 32*(1), 1–11. https://doi.org/10.1109/TKDE.2018.2880192

Giorgioni, S., Politi, M., Salman, S., Croce, D., & Basili, R. (2020). *UNITOR @ Sardistance2020: Combining transformer-based architectures and transfer learning for robust stance detection.* https://en.wikipedia.org/wiki/Sardines_movement.

Küçük, D., & Fazli, C. A. N. (2020). Stance detection. *ACM Computing Surveys (CSUR), 53*(1). https://doi.org/10.1145/3369026

Kunneman, F., Lambooij, M., Wong, A., Bosch, A. V. D., & Mollema, L (2020). Monitoring stance towards vaccination in twitter messages. *BMC Medical Informatics and Decision Making, 20*(1), 1–14. https://doi.org/10.1186/s12911-020-1046-y

Mcinnes, L., Healy, J., & Melville, J. (2020). *UMAP: Uniform manifold approximation and projection for dimension reduction.*

Roesslein, J. (2020). *Tweepy: Twitter for Python!* https://github.com/tweepy/tweepy.

Sancheti, A., Chawla, K., & Verma, G. (2020). *LynyrdSkynyrd at WNUT-2020 Task 2: Semi-supervised learning for identification of informative COVID-19 English Tweets.* https://arxiv.org/abs/2009.03849.

Zotova, E., Agerri, R., Nuñez, M., & Rigau, G. (2020). Multilingual stance detection in Tweets: The Catalonia Independence Corpus - ACL Anthology. In *Proceedings of the 12th Language Resources and Evaluation Conference.* https://aclanthology.org/2020.lrec-1.171/.

Agerri, R., Centeno, R., Espinosa, M., Fernandez De Landa, J., & Rodrigo, A. (2021). *VaxxStance@IberLEF 2021: Overview of the task on going beyond text in cross-lingual stance detection.* https://doi.org/10.26342/2021-67-15.

Al-Ghadir, A. I., Azmi, A. M., & Hussain, A. (2021). A novel approach to stance detection in social media tweets by fusing ranked lists and sentiments. *Information Fusion, 67*, 29–40. https://doi.org/10.1016/j.inffus.2020.10.003

Al-Laith, A., Shahbaz, M., Alaskar, H. F., & Rehmat, A. (2021). AraSenCorpus: A semi-supervised approach for sentiment annotation of a large arabic text corpus. *Applied Sciences 2021, 11*(5), 2434. https://doi.org/10.3390/APP11052434. *Vol. 11, Page 2434.*

ALDayel, A., & Magdy, W. (2021). Stance detection on social media: State of the art and trends. *Information Processing & Management, 58*(4), Article 102597. https://doi.org/10.1016/J.IPM.2021.102597

Alsafari, S., & Sadaoui, S. (2021). Semi-supervised self-training of hate and offensive speech from social media. *Applied Artificial Intelligence.* https://doi.org/10.1080/08839514.2021.1988443/FORMAT/EPUB

Chawla, Y., Radziwon, A., Scaringella, L., Carlson, E. L., Greco, M., Silveira, P. D., de Aguiar, E. P., Shen, Q., Will, M., & Kowalska-Pyzalska, A. (2021). Predictors and outcomes of individual knowledge on early-stage pandemic: Social media, information credibility, public opinion, and behaviour in a large-scale global study. *Information Processing & Management, 58*(6), Article 102720. https://doi.org/10.1016/j.ipm.2021.102720

Chen, J., & Wang, Y. (2021). Social media use for health purposes: systematic review. *Journal of Medical Internet Research, 23*(5), e17917. https://doi.org/10.2196/17917

Herrera-Peco, I., Jiménez-Gómez, B., Romero Magdalena, C. S., Deudero, J. J., García-Puente, M., Benítez De Gracia, E., & Ruiz Núñez, C. (2021). Antivaccine movement and COVID-19 Negationism: A content analysis of Spanish-written messages on Twitter. *Vaccines, 9*(6), 656. https://doi.org/10.3390/vaccines9060656

Kaushal, A., Saha, A., & Ganguly, N. (2021). tWT–WT: A Dataset to Assert the Role of Target Entities for Detecting Stance of Tweets. 3879–3889. https://doi.org/10.18653/V1/2021.NAACL-MAIN.303.

Kumari, R., Ashok, N., Ghosal, T., & Ekbal, A. (2021). Misinformation detection using multitask learning with mutual learning for novelty detection and emotion recognition. *Information Processing & Management, 58*(5), Article 102631. https://doi.org/10.1016/j.ipm.2021.102631

Meng, W., Qureshi, Z., & Khandker, R. (2021). PND66 topic landscape analysis of Reddit social media submissions in insomnia. *Value in Health, 24*, S171. https://doi.org/10.1016/j.jval.2021.04.850

Murakami, R., & Chakraborty, B. (2021). Neural topic models for short text using pretrained word embeddings and its application to real data. In *2021 IEEE 4th International Conference on Knowledge Innovation and Invention (ICKII)* (pp. 146–150). https://doi.org/10.1109/ICKII51822.2021.9574752

Santoveña-Casal, S., Gil-Quintana, J., & Ramos, L. (2021). Digital citizens' feelings in national #Covid 19 campaigns in Spain. *Heliyon, 7*(10), e08112. https://doi.org/10.1016/j.heliyon.2021.e08112

Suarez-Lledo, V., & Alvarez-Galvez, J. (2021). Prevalence of health misinformation on social media: Systematic review. *Journal of Medical Internet Research, 23*(1). https://doi.org/10.2196/17187

Zhao, X., Wang, D., Zhao, Z., Liu, W., Lu, C., & Zhuang, F. (2021). A neural topic model with word vectors and entity vectors for short texts. *Information Processing & Management, 58*(2), Article 102455. https://doi.org/10.1016/j.ipm.2020.102455

Zhou, C., Xiu, H., Wang, Y., & Yu, X. (2021). Characterizing the dissemination of misinformation on social media in health emergencies: An empirical study based on COVID-19. *Information Processing & Management, 58*(4), Article 102554. https://doi.org/10.1016/j.ipm.2021.102554

Alkhalifa, R., & Zubiaga, A. (2022). Capturing stance dynamics in social media: Open challenges and research directions. *International Journal of Digital Humanities.* https://doi.org/10.1007/s42803-022-00043-w

Dutta, S., Caur, S., Chakrabarti, S., & Chakraborty, T. (2022). Semi-supervised stance detection of tweets via distant network supervision. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining* (pp. 241–251). https://doi.org/10.1145/3488560.3498511

Kumari, R., Ashok, N., Ghosal, T., & Ekbal, A. (2022). What the fake? Probing misinformation detection standing on the shoulder of novelty and emotion. *Information Processing & Management, 59*(1), Article 102740. https://doi.org/10.1016/j.ipm.2021.102740

Pan, M., Wang, J., Huang, J. X., Huang, A. J., Chen, Q., & Chen, J. (2022). A probabilistic framework for integrating sentence-level semantics via BERT into pseudo-relevance feedback. *Information Processing & Management, 59*(1), Article 102734. https://doi.org/10.1016/j.ipm.2021.102734

Roy, S., Bhanu, M., Saxena, S., Dandapat, S., & Chandra, J. (2022). gDART: Improving rumor verification in social media with Discrete Attention Representations. *Information Processing & Management, 59*(3), Article 102927. https://doi.org/10.1016/j.ipm.2022.102927

Salmi, S., Mérelle, S., Gilissen, R., van der Mei, R., & Bhulai, S. (2022). Detecting changes in help seeker conversations on a suicide prevention helpline during the COVID− 19 pandemic: In-depth analysis using encoder representations from transformers. *BMC Public Health, 22*(1), 530. https://doi.org/10.1186/s12889-022-12926-2

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Strope, B., & Kurzweil, R (2018). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 169–174). Universal Sentence Encoder for English. https://doi.org/10.18653/v1/D18-2029.