Expert Systems With Applications 170 (2021) 114547

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Semi-automatic generation of multilingual datasets for stance detection in Twitter

Elena Zotova [a], Rodrigo Agerri [b,*], German Rigau [b]

[a] SNLT group at Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Spain
[b] HiTZ Center - Ixa, University of the Basque Country UPV/EHU, Spain

ARTICLE INFO

Keywords:
Stance detection
Multilingualism
Text categorization
Fake news
Deep learning

ABSTRACT

Popular social media networks provide the perfect environment to study the opinions and attitudes expressed by users. While interactions in social media such as Twitter occur in many natural languages, research on stance detection (the position or attitude expressed with respect to a specific topic) within the Natural Language Processing field has largely been done for English. Although some efforts have recently been made to develop annotated data in other languages, there is a telling lack of resources to facilitate multilingual and crosslingual research on stance detection. This is partially due to the fact that manually annotating a corpus of social media texts is a difficult, slow and costly process. Furthermore, as stance is a highly domain- and topic-specific phenomenon, the need for annotated data is specially demanding. As a result, most of the manually labeled resources are hindered by their relatively small size and skewed class distribution. This paper presents a method to obtain multilingual datasets for stance detection in Twitter. Instead of manually annotating on a per tweet basis, we leverage user-based information to semi-automatically label large amounts of tweets. Empirical monolingual and cross-lingual experimentation and qualitative analysis show that our method helps to overcome the aforementioned difficulties to build large, balanced and multilingual labeled corpora. We believe that our method can be easily adapted to easily generate labeled social media data for other Natural Language Processing tasks and domains.

1. Introduction

The phenomenon of *fake news* is becoming notoriously common, particular within social media. Fake news have been defined as “a made-up story with an intention to deceive”[1], often for a secondary gain, and it is considered to be one of the most serious challenges facing the news industry and the political sphere.

Determining the veracity of a given article or social media message, in absence of further context or background knowledge, is often a very difficult task, even for expert fact-checkers. Thus, the organizers of the Fake News Challenge considered that *fake news* detection should be broken down into intermediate tasks[2], so that the output of each of them would provide an *indicator* to be taken into account in the overall fake news detection task. The first stage of the Fake News Challenge was stance detection. In their view, a stance detection system would allow fact-checkers to automatically know which documents, messages or users agree or disagree with a given document thereby helping to identify contentious content.

Stance detection has been addressed from at least two general perspectives, depending on whether the topic is open-ended or static. There are at least two well known shared tasks which formulate open stance detection tasks. The aforementioned Fake News Challenge, where the task is to classify whether a given document *agrees*, *disagrees*, *discusses* or is *unrelated* with respect to a previously published headline or news document. Closely related to this, the RumourEval 2017 shared task (Derczynski et al., 2017) referred to four different categories to express the stance of a tweet with respect to a triggering message (often a rumour), namely, *support*, *deny*, *query* and *comment*. Closer to our work, the second perspective defines stance with respect to a pre-defined or given topic or target, as it is often called. Thus, given a tweet and a target entity or topic, Natural Language Processing (NLP) systems should try to classify whether the stance expressed is in *favour*, *against*, or whether is

* Corresponding author.
E-mail addresses: ezotova@vicomtech.org (E. Zotova), rodrigo.agerri@ehu.eus (R. Agerri), german.rigau@ehu.eus (G. Rigau).

[1] https://www.nytimes.com/2016/12/06/us/fake-news-partisan-republican-democrat.html
[2] http://www.fakenewschallenge.org/

https://doi.org/10.1016/j.eswa.2020.114547
Received 29 July 2020; Received in revised form 4 December 2020; Accepted 24 December 2020
Available online 1 January 2021

unrelated or *none* with respect to the given target. Let us consider the following two examples:

> **Tweet:** *I still remember the days when I prayed God for strength.. then suddenly God gave me difficulties to make me strong. Thank you God! #SemST*
> **Target:** Atheism.
> **Stance:** AGAINST.
> **Tweet:** *@PH4NT4M @MarcusChoOo @CheyenneWYN women. The term is women. Misogynist! #SemST*
> **Target:** Feminist Movement.
> **Stance:** FAVOR

These two examples illustrate the nature of the task. Messages are very short, contain non-standard spelling grammar, emojis, hashtags and figurative language such as irony and sarcasm. This particular task, as defined by the Stance Detection in Twitter at SemeEval 2016 (Mohammad, Kiritchenko, Sobhani, Zhu, & Cherry, 2016), consists of classifying single tweets, without conversational structure, into one of three classes: FAVOR, AGAINST and NONE.

As it is often the case in the NLP field, research on stanced detection has mostly be done for English, although recently there have been efforts to develop annotated corpora for stance detection in languages other than English. For instance, Mohtarami, Glass, and Nakov, 2019 experiments on the Arabic corpus provided by Baly et al., 2018. A dataset in Czech was developed from comments of news and used to experiment with SVM, Maximum Entropy and Convolutional Neural Networks (CNNs) (Hercig, Krejzl, Hourová, Steinberger, & Lenc, 2017). Furthermore, Vychegzhanin and Kotelnikov, 2019 presented a method of assembling classifiers for stance detection in Russian. Finally, Evrard, Uro, Hervé, and Mazoyer, 2020 created a French Twitter corpus for stance detection.

Still, there is a clear need for resources to investigate crosslingual approaches to stance detection. To the best of our knowledge, the Catalan and Spanish corpus provided by the IberEval 2017 and 2018 shared tasks (Taulé et al., 2017; Taulé, Rangel, Martí, & Rosso, 2018) is the first work with the aim of facilitating multilingual research on stance detection. This was later complemented by Lai et al., 2020, who provided another two datasets in French and Italian. However, the datasets are about different topics on data obtained from different dates which makes it very difficult to perform crosslingual research. Moreover, the majority of the previous resources, both monolingual or multilingual, are hindered by their small size and skewed class distribution (Taulé et al., 2018; Lai et al., 2020). This is partially due to the fact that manual annotation on a tweet basis is a difficult, slow and costly task. To make things worse, stance is a highly domain- and topic-specific phenomenon, which means that each topic requires its own annotated dataset to develop state-of-the-art classifiers. This creates an endless demand of labeled data.

This paper tackles these issues by proposing a method to semi-automatically obtain multilingual annotated data for stance detection based on a categorization of Twitter users. The result of applying such method is the *Catalonia Independence Corpus* (CIC).

For our new corpus we collect and semi-automatically annotate tweets (coetaneous, and on the same topic) in two different languages: Catalan and Spanish. The availability of multilingual annotated data collected on the same dates and on the same topics facilitates comparison of models in crosslingual experimentation. Otherwise, it would be difficult to know if differences in performance are due to the learning models or to differences in topics and temporality of the obtained data.

Our present work substantially improves and extends a first version of the CIC dataset and the preliminary set of experiments presented in Zotova, Agerri, Nuñez, and Rigau, 2020. More specifically, in this work we make the following contributions.

First, we devise an alternative method to build a new version of the CIC dataset, CIC-Random, where the messages across the train, development and test splits are of different users.

Second, we provide a large set of experiments in four different datasets (SemEval 2016 and IberEval 2018, CIC and CIC-Random) for three languages (Catalan, English, Spanish) showing that systems behave consistently across datasets and languages; this in turn suggests that our methodology to build annotated datasets for multilingual and cross-lingual stance detection in Twitter helps to alleviate the difficulties faced by previous manual-based efforts (Mohammad et al., 2016; Taulé et al., 2018).

Third, we use the newly created CIC dataset to perform cross-lingual experimentation, the first of its kind for stance detection, with large pre-trained multilingual language models such as multilingual BERT and XLM-RoBERTa (Devlin, Chang, Lee, & Toutanova, 2019; Conneau et al., 2019a), comparing zero-shot approaches with translation-based strategies. These experiments also help to provide some insights about the multilingual behaviour of the transformers.

Fourth, we perform extensive error analysis to better understand the pros and cons of our method with respect to manual annotation. It seems that while the semi-automatic nature of our method introduces some noise in the annotations, user-specific information provides the extra context required to better label the individual tweets.

Fifth, we use the CIC-Random version to show that systems obtaining better results on the original CIC data, as opposed to results in previous benchmarks (Mohammad et al., 2016; Taulé et al., 2018), were not due to the systems overfitting on specific users' writing style. This result, together with the manual inspection and annotation of a sample of the CIC corpus, suggests that for stance detection in Twitter, its quality is as good as the one we obtain manually.

Finally, in order to facilitate reproducibility of results we publicly distribute the CIC and CIC-Random corpora and the variously pre-processed versions of every corpus used in this paper[3].

The rest of the paper is organized as follows: Section 2 describes previous approaches to stance detection. In Section 3 we describe our methodology to build datasets for stance detection, method that has been first employed to generate our new CIC corpus. The experimental setup is specified in Section 4. Section 5 reports on our monolingual and cross-lingual experiments, while Section 6 provides an error analysis of the obtained results. We finish with some concluding remarks in Section 7.

2. Related Work

The growing interest on stance detection is demonstrated by at least three recent surveys addressing the topic. The first one revised opinion mining methods in general with a special focus on stance towards products (Wang, Zhou, Jiang, Si, & Yang, 2019). Another recent survey study detailed research work that modeled stance detection as a text entailment task (Küçük & Can, 2020). This particular survey provided a broad coverage of stance detection methods, including works from various research domains such as Natural Language Processing (NLP), Computational Social Science, and Web science. Furthermore, it also surveyed the modeling of stance using text, network-based, and behavioural features. More recently, AlDayel and Magdy, 2020 covered new research directions on stance detection in social media.

Automatic stance detection in social media is divided into two main approaches. First, those that rely on *traditional* machine learning models combining hand-engineered features (Mohammad, Sobhani, & Kiritchenko, 2017) or static word representations (Bøhler, Asla, Marsi, & Sætre, 2016). Second, those that are based on the application of deep learning and neural networks Augenstein, Rocktäschel, Vlachos, and Bontcheva, 2016; Zarrella et al., 2016; Wei, Zhang, Liu, Chen, and Wang, 2016; Igarashi, Komatsu, Kobayashi, Okazaki, and Inui, 2016.

[3] https://github.com/ZotovaElena/Multilingual-Stance-Detection

Although other well-known English datasets for stance detection exist[4] (Derczynski et al., 2017), closer to our particular interest is the dataset from the SemEval 2016 shared task for Stance Detection in Twitter (Mohammad et al., 2016). On the supervised setting of SemEval 2016, (Mohammad et al., 2017) obtained the best results using a SVM classifier based on word n-grams and character n-grams features, outperforming other deep learning approaches (Zarrella et al., 2016; Wei et al., 2016). Later, AlDayel and Magdy, 2019 explored users interactions in Twitter and compared various features, including on-topic content, network interactions, user's preferences, and online network connections.

More recently, some other deep learning approaches improved over the SemEval 2016 official state-of-the-art results. For instance, Du, Xu, He, and Gui, 2017 proposed a neural network-based model to incorporate target-specific information by means of an attention mechanism. In the work of Benton and Dredze, 2018 recurrent neural networks are combined with pre-trained "user embeddings", which enrich the training data with additional information in user-based level. Sun, Wang, Zhu, and Zhou, 2018 presented an hierarchical attention network to weigh the importance of various linguistic information, and learn the mutual attention between the document and the linguistic information. One more approach (Wei, Mao, & Zeng, 2018) with end-to-end deep neural model leverages attention mechanism to detect stance through target and tweet interactions. The work of Siddiqua, Chy, and Aono, 2019 proposes a neural ensemble model that combines two densely connected BiLSTMs, nested LSTMs where each module is coupled with an attention mechanism.

Current best results on the SemEval 2016 dataset are reported by Ghosh, Singhania, Singh, Rudra, and Ghosh, 2019. They offer a systematic comparison of seven stance detection methods and fine-tuned a masked language pre-trained model (BERT Large) (Devlin et al., 2019) to report current state-of-the-art results on this particular benchmark.

Neural network approaches have also been successful for the SemEval 2016 Task B (weakly-supervised setting). For instance, apart from the previously mentioned systems, Augenstein et al., 2016 proposed a bidirectional Long-Short Term Memory (LSTM) encoding model. First, the target is encoded by a LSTM network and then a second LSTM is used to encode the tweet using the encoding of the target as its initial state.

Semi-supervised methods include building stance detection models with small training sets. Misra, Ecker, Handleman, Hahn, and Walker, 2016 proposes a data augmentation method for small annotated datasets based on stance-bearing hashtags. Fraisier, Cabanac, Pitarch, Besançon, and Boughanem, 2018 proposed an ensemble of systems to model the stance detection task at user level complemented by content–, interaction– and geographic–based proximity of social network profiles.

Multilingual approaches based on the TW-10 corpus were developed for the "MultiModal Stance Detection in tweets on Catalan #1Oct Referendum" task at IberEval 2018. The best Spanish system (Segura-Bedmar, 2018) consisted of a linear classifier with TF-IDF vectorization, obtaining a final 28.02 F1 macro score in the Spanish test data. The best result for Catalan (Cuquerella & Rodríguez, 2018) consisted of combining the Spanish and Catalan training sets to create a larger and more balanced corpus. They experimented with stemming of various lengths (three, four and five characters) and removing character suffixes from the word, which helped to generalize over Catalan and Spanish. Their final F1 macro was 30.68.

Lai et al., 2020 propose a multilingual stance detection system for English, Spanish, Catalan, French and Italian. The English, Catalan and Spanish data are based on the SemEval 2016 and IberEval 2017 (Lai, Cignarella, & Farías, 2017) corpus respectively, whereas the French and Italian datasets were originally presented for that paper. They explore different types of features—stylistic, structural, contextual, and affective—and their contribution in the learning process of models such as BiLSTM, CNN and SVM. Their scores for English, Catalan and Spanish were substantially improved by Zotova et al., 2020, a preliminary version of the approach we present in this paper.

Summarizing, the few multilingual approaches presented so far are hindered by the small size and skewed class distribution of the existing datasets. In the next section we will examine three of these resources, SemEval 2016 and TW-10 (Catalan and Spanish) and propose our semi-automatic method to efficiently obtain good quality annotated data for multilingual stance detection in Twitter.

3. New Dataset: The Catalonia Independence Corpus

In this section we provide a detailed description of the method developed to generate the Catalonia Independence Corpus (CIC). In order to study multilingual and crosslingual approaches for stance detection in Twitter, it is desirable to obtain annotated datasets on a common topic for more than one language and obtained on the same dates (coetaneous). Previous works include datasets in several languages, notably, IberEval 2018 (Taulé et al., 2018 & Lai et al., 2020). However, they do not provide an adequate setting for multilingual and crosslingual studies to stance detection.

With respect to IberEval 2018, they provide annotated data in Catalan and Spanish, but in the Catalan part the classes distribution is extremely skewed (Taulé et al., 2018). Regarding Lai et al., 2020, they developed two new datasets for French and Italian, but they are not coetaneous nor about the same topic. These issues make crosslingual experimentation very difficult as differences between performance across languages may be due to other issues rather than model performance, namely, one topic being more difficult than the other.

Finally, most previous datasets are quite small because they exclusively rely on manual annotation on a per tweet basis. This is very costly but also not very efficient because in many cases annotating a tweet without its background context is nigh on impossible.

Our new dataset for stance detection in Twitter aims to address these shortcomings: (i) the collected data is coetaneous across languages, (ii) on the same topic, (iii) multilingual (Spanish and Catalan), (iv) their class distribution is balanced, and (v), their annotation method more efficient and, as a consequence, their size is much larger than previous datasets for stance detection.

The first three issues are addressed in the data collection phase, but, crucially, issues (iv) and (v) are direct consequences of our methodology to obtain stance annotations. Our method for efficient annotation is based on the following *four building steps*:

- **User-based annotation**: namely, taking into account the full timeline. We assume that it is easier to annotate a full timeline rather than the text of a single tweet without context.
- **User relations**: Based on previous research on political homophily (Barberá, 2015; Himelboim, McCreery, & Smith, 2013; Zubiaga, Wang, Liakata, & Procter, 2019), we use the relations between users (retweets) to obtain more users or accounts from which to obtain the tweets for our dataset.
- **Hashtags and Keywords**: A selection of hashtags and keywords are applied to obtain tweets that are relevant to our topic of interest.
- **Topic Modelling**: We refine the extraction of on-topic tweets performed on step 2 by applying LDA. The idea is to cross user-level and topic information to provide annotated tweets for the final version of our dataset.

The result is a final balanced dataset containing 10 K annotated tweets for Spanish and for Catalan, respectively. This new dataset is the largest annotated corpus of its kind, displaying all the required features to perform crosslingual experimentation. In the following we describe each of the steps involved in the development of the CIC corpus.

It should be noted that our annotation process, described in Sections

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

Table 4
Example of a user (@manuelvalls) manually classified as AGAINST of independence of Catalonia.

Tweet by @manuelvalls	Translation
Pese a la complicidad de Chavists del Ayuntamiento de Barcelona como los populistas de izquierdas y los independentistas desde? @CiudadanosCs? y? @CiutadansBCN? llevamos años luchando por la libertad y el respeto a los derechos humanos. #VenezuelaEnLaCalle #Venezuela https://t.co/fjw1hxXyYD\end	Despite the complicity of Chavistas from the Barcelona City Council such as the left-wing populists and the independentists from? @CiudadanosCs? and? @CiutadansBCN? we have been fighting for freedom and respect for human rights for years. #VenezuelaEnLaCalle #Venezuela https://t.co/fjw1hxXyYD
El relato del golpe de Estado en Cataluña contado por los protagonistas constitucionalistas. Vargas Llosa: "Era algo gigantesco. Creo que la manifestación más grande que he visto". Por @e_bece https://t.co/Aymvhgj6cw	The story of the coup d'état in Catalonia told by the constitutionalist protagonists. Vargas Llosa:"It was something gigantic. I think it was the biggest demonstration I've ever seen". By @e_bece https://t.co/Aymvhgj6cw
La noche que el Rey defendió a España: "El 3 de octubre de 2017 el Rey cortó en seco una ensoñación separatista que llevaba años marcando goles fuera de juego al Estado ante unos árbitros que se limitaban a llevarse las manos a la cabeza". Por @AlmudenaMF https://t.co/XDDM2D96XH	The night the King defended Spain: "On 3 October 2017, the King cut short a separatist dream that had been scoring offside goals for the State for years in front of referees who were simply putting their hands on their heads." By @AlmudenaMF https://t.co/XDDM2D96XH

gathered during 12 days on February and March 2019 in Barcelona and Terrassa[5]. This collection was originally obtained for industrial research in stance detection and political ideology (left–right) prediction. The crawling was performed with full access to the Twitter API. For its use in academic research, we first separated them by language[6] obtaining 680,000 tweets in Catalan and 2 million tweets in Spanish. We discarded duplicated messages and those shorter than three words.

For the data collection, we first compiled a list of Twitter accounts from media, political parties and political activists that clearly and explicitly express their stance with respect to the independence of Catalonia. This list was manually compiled based on the high visibility of users related to the political scene in Catalonia. In total we obtain around 150 accounts of personalities, political parties and digital media.

Secondly, we extracted the most active and retweeted tweets in the crawled corpus selecting a list of 1200 accounts in which there was enough content related to the topic of interest, namely, the independence of Catalonia. In this step non-political accounts were discarded.

With respect to the data collection stage, it should be noted that in order to have some data available to learn stance about a given topic we need tweets that actually talk about that topic. That is true for any topic, not just for the one in our dataset. Thus, if we were to collect data about other topics such as *Feminism* or *Donald Trump* or *Climate Change* or *Brexit*, we would still need to identify keywords, accounts and/or hashtags that talk about those topics. Therefore, this is not an issue specific to our approach.

3.2. User-based Annotation

Annotation of the 1200 accounts obtained in the previous step was carried out using the same three labels and guidelines as in previously existing stance datasets Mohammad et al., 2016; Taulé et al., 2018. Thus, FAVOR and AGAINST refer to a positive or negative stance towards the independence of Catalonia, respectively. Finally, NONE will express neither a negative nor a positive stance, or simply that it is not

Fig. 1. Two examples of accounts which can be easily annotated as FAVOR and AGAINST, respectively. The Catalonian republican flag appears on the FAVOR-user's profile whereas the Spanish official flag is displayed in the profile of the user AGAINST the independence.

3.2, 3.3, 3.4 and 3.5, is language- and topic-independent, namely, the four steps listed above can be performed regardless of the topic, target language and even task. However, these issues do affect the collection of the data, as the objective of data collection must be obtaining on-topic tweets that can then be leveraged for classification.

3.1. Data Collection

In order to create the CIC dataset, we used a collection of tweets

possible to reach a clear decision.

Unlike previous approaches, and in order to increase the consistency of the annotations and to speed up the annotation process, we do not manually annotate each tweet in a one-by-one fashion. Instead, the annotation process was mainly based on classifying stance at user level,

[5] An industrial city 25 km away from Barcelona
[6] https://code.google.com/archive/p/language-detection/

Table 5
Distribution of the categorized users.

Label	Count
AGAINST	3,091
FAVOR	22,247
NONE	176

namely, we categorized the tweets' authors manually by checking their Twitter accounts.

The assumption was that for a human annotator it is easier to annotate a full timeline rather than the text of a single tweet without additional context. Thus, in our annotation process the decision about stance was also made taking into consideration other aspects from the users' accounts, such as the use of special emojis and symbols that may state clearly the stance towards the target (e.g.. displaying a yellow ribbon is a pro-independence symbol, whereas a Spanish flag would convey that the user is against the independence, etc.), or by the Bio section. In this step each user or Twitter account is assigned a stance label: FAVOR, AGAINST or NONE.

Table 4 presents an example of a user categorised as being AGAINST the independence of Catalonia. These tweets were written by Manuel Valls, a Catalan politician from a unionist party. He clearly expresses his opinion about the topic with typical vocabulary such es *coup d'état*, *separatist* or *constitutionality*.

Furthermore, Fig. 1 presents two *simple* examples of accounts which can be easily categorized as FAVOR (left) and AGAINST (right), just by looking at their profile. Thus, the FAVOR-user displays a flag related to the Catalonian Republic and the independence movement. Moreover, the user against independence displays a profile related to Tabarnia, an unionist symbol[7].

3.3. User Relations

In addition to the user-level annotations, we extracted the relations between users based on their retweets (Otte & Rousseau, 2002). A study of the behaviour of Twitter users (Boyd, Golder, & Lotan, 2010) mentions, among others, that retweets are motivated by "publicly agreeing with someone", "to give visibility", or "to validate other's thoughts".

Furthermore, previous research on political homophily (Barberá, 2015; Himelboim et al., 2013) shows that homophily is also reflected in social media. In other words, supporters of one particular party or ideology are more likely to interact with users of the same ideology or party.

This idea has been studied in several NLP works relevant to this work (Lai et al., 2020; Zubiaga et al., 2019). In particular, Zubiaga et al., 2019 empirically demonstrates that there is a correlation between the national identity of users (in relation to independentist movements) and the relations between users in the social network. Their case study also includes Catalonia, together with Scotland and the Basque Country. For our particular interests, this means that by using the retweets, we can increase the size of our initial data pool (1200 users) with more users that are anti- and pro-independence.

Following the principle of political homophily, we assumed that in general users retweet mostly users expressing the same political stance as the original message. While this method may introduce some noise, it allowed us to quickly obtain a large amount of annotated data. Thus, from the 1,200 accounts that were labelled manually in the previous step, we were able to automatically obtain 25,510 categorized users. We do not distinguish between Catalan and Spanish users because most of them are fully bilingual. Table 5 reports the distribution of the categorized users. The final set contains 131,022 unique tweets in Catalan and 202,645 unique tweets in Spanish.

[7] https://en.wikipedia.org/wiki/Tabarnia

Table 6
Distribution of tweets obtained by hashtags and keywords related to the "Independence of Catalonia".

Label	Catalan	Spanish
AGAINST	1,476	8,267
FAVOR	23,030	11,843
NONE	986	497

Thus, while the use of user relations (retweets) is technically motivated by the need to increase the number of accounts from which to extract the tweets to be included in the final dataset, this is theoretically and empirically motivated by previous research on political homophily (Barberá, 2015; Himelboim et al., 2013), and especially on independence movements (Zubiaga et al., 2019).

3.4. Hastags and Keywords

As we mentioned earlier, we annotated every tweet in the corpus by assigning the stance directly to the account users. However, this does not mean that we can use every tweet from the users, given that many messages may not be related to our specific target, namely, the independence of Catalonia. This issue is addressed by extracting relevant tweets using hashtags and keywords and by applying LDA topic modelling. The latter is described in the next section.

In this step we extracted all the hashtags from the corpus and selected manually those that were related to the independence of Catalonia, such as *#CataluñaesEspaña, #CatalanRepublic, #Tabarnia, #GolpeDeEstado, #independéncia, #judicifarsa, #CatalanReferendum* etc., totalling 450 hashtags. We also manually added keywords for both languages, 25 in total. We labeled each tweet as being on topic if it contained one of the relevant hashtags or keywords. Table 6 displays the distribution of tweets after applying this filtering step using hashtags and keywords.

3.5. Topic Detection

The distribution of tweets per language obtained in the previous step, as shown by Table 6, evidences that the vast majority of the tweets are labelled as FAVOR. As our aim was to obtain a balanced dataset, we needed to add more tweets to the under-represented classes. We use the MALLET (McCallum, 2002) implementation of Latent Dirichlet allocation (LDA) (Blei, Ng, & Jordan, 2003) to perform basic target detection in the corpus of categorized users described in Table 5, Section 3.3. The objective was to obtain more on-topic tweets for those classes that are under-populated (AGAINST and NONE in Catalan and NONE in Spanish). We manually revised the obtained topics and selected only those tweets which were clustered within the *independence* topic.

In a final step, we selected approximately 10,000 tweets per language (excluding those shorter than four words) keeping the proportion of users from the initial pool of crawled tweets. We split them keeping 60% for training, and 20% each for development and test. The average length of a tweet in the Catalan Independence Corpus is very similar to the average in the TW-1O dataset. Considering that our corpus does not include extra context, this means that our average length is longer than those obtained at SemEval 2016 or TW-1O. Futhermore, our corpus is also much larger (Mohammad et al., 2016; Taulé et al., 2018) and presents a more balanced distribution of classes, as shown by Table 7. The result of this final step is the *Catalonia Independence Corpus* (CIC).

3.6. User Bias

A common feature of corpora based on social media is that a small amount of users usually generate a large proportion of the posts and viceversa. We checked this issue in the original CIC corpus and found out that in the Spanish subset there are 407 users whereas for Catalan the

Table 7
Distribution of the CIC corpus separated into training, validation and test sets. A—AGAINST, F—FAVOR, N—NONE.

	Train			Validation			Test			Total
Dataset	**A**	**F**	**N**	**A**	**F**	**N**	**A**	**F**	**N**	
CIC–CA	2,680	2,545	752	1,201	506	372	937	850	205	10,048
CIC–CA-Random	2,416	2,335	1,277	820	763	427	804	752	454	10,048
CIC-ES	2,560	2,276	1,100	875	702	503	953	843	265	10,077
CIC-ES-Random	2,515	2,420	1,111	856	782	377	829	807	380	10,077

authors were 1,100. Furthermore, we also realized that the same users were present across the three partitions, train, development and test. As tweets from the same author might contain specific information about the user, such as individual writing style, vocabulary or other communicative behaviour, we wanted to double check that systems were not learning those specific features and thus overfitting to the characteristics of some specific users.

For that reason, we created a new version of the original CIC dataset, namely, CIC-Random, reorganizing the CIC corpus in such a way that users appearing in the train set were not included in the development and test sets. More specifically, we randomly sampled the list of users and created three new splits for training, development and test, according to three criteria: (i) tweets from the same user cannot occur across the three splits; (ii) the proportion of tweets of each user in each split has to be same as in the original CIC corpus and, (iii) the balance between the three stance labels must be kept. The result of this process is the *CIC-Random* dataset.

As randomness was the key in this process, the distribution is not exactly the same as in the original CIC corpus, as shown by Table 7.

4. Experimental setup

As it is explained in the previous chapter, the development of the Catalonia Independence Corpus (CIC) was motivated by the lack of large, balanced, multilingual and coetanous corpora for stance detection. Indeed, previous experiments have shown the difficulty of drawing meaningful conclusions with the TW-1O corpus, mainly due to the skewed class distribution in the Catalan set (Taulé et al., 2018; Zotova et al., 2020).

In this section we describe the setup for the experiments performed using the dataset described in the previous chapter plus two previously existing benchmarks, SemEval 2016 and TW-10. Our motivation is to show that, in addition to being faster and cheaper to generate Twitter-based annotated datasets for stance detection, our semi-automatic, user-based method produces annotated data which is as reliable for experimentation as manually-based annotated datasets. We will investigate this claim by studying the behaviour of popular text classification baselines and methods across datasets, both monolingual (SemEval 2016) and multilingual (TW-1O and CIC). We would expect the systems to exhibit similar behaviour across datasets.

Moreover, the development of a multilingual dataset such as CIC allows us to experiment with the application of large multilingual language models such as mBERT or XLM-RoBERTa (Devlin et al., 2019; Conneau et al., 2019a) in cross-lingual settings. Thus, for a scenario in which there is not training data available for the target language, we can investigate whether it would be better: (i) using the multilingual language model for the target language directly in a zero-shot fashion or, (ii) automatically translating the data from the target language to a language for which we do have training data available, namely, applying a *"translate and fine-tune"* method.

In the rest of this section we describe the benchmark dataset, the data-preprocessing methods and the systems used for experimentation. We used four different system types: (i) TF-IDF vectorization with a SVM classifier (TF-IDF + SVM); (ii) SVM trained by averaging fastText word embeddings (Grave, Bojanowski, Gupta, Joulin, & Mikolov, 2018) for the representation of tweets (FTEmb + SVM); (iii) the fastText text classification system (Joulin, Grave, Bojanowski, & Mikolov, 2017) with fastText word embeddings (FTEmb + fastText), (iv) pre-trained language models based on the transformer architecture, including monolingual (XLNET (Yang et al., 2019) and RoBERTa (Liu et al., 2019)) and multilingual language models (mBERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau et al., 2019a)).

Table 1
Number of examples per target in the SemEval 2016 English dataset.

Target	Train	Test
Atheism	513	220
Climate Change is a Real Concern	395	169
Feminist Movement	664	285
Hillary Clinton	639	295
Legalization of Abortion	603	280
Total	**2,814**	**1,249**

Table 2
Examples of tweets from SemEval 2016 dataset.

Tweet	Target	Stance
@PH4NT4M @MarcusChoOo @CheyenneWYN women. The term is women. Misogynist! #SemST	Feminist Movement	FAVOR
American conservatism has everything to do with religion with all the good stuff taking out of it. #SemST	Atheism	AGAINST

4.1. Benchmark Datasets

In this section we describe other two well-known datasets that, together with our newly created CIC Corpus, will be used for the experimentation. This means that our experiments will be evaluated on seven different datasets and three languages, namely, Catalan English, and Spanish. We first described the SemEval 2016 dataset on Stance Detection, the first of its kind, and then the multilingual TW-10 corpus.

4.1.1. SemEval 2016

The dataset presented at the Stance Detection task organized at SemEval 2016[8] (Mohammad et al., 2016), consists of English tweets labeled for both stance (AGAINST, FAVOR and NONE). In the supervised track, more than 4,000 tweets are annotated with respect to five targets: "Atheism", "Climate Change is a Real Concern", "Feminist Movement", "Hillary Clinton", and "Legalization of Abortion". For each target, the annotated tweets were ordered by their timestamps. The first 70 percent of the tweets form the training set and the last 30 percent are reserved for the test set. Table 1 presents the final distribution of training and testing examples. (See Table 2).

To prepare the dataset the organizers collected 2 million tweets containing FAVOR, AGAINST and ambiguous (NONE) hashtags for the selected targets. These hashtags were removed after manual annotation,

[8] http://alt.qcri.org/semeval2016/task6/data/

Table 3
Distribution of classes in the TW-1O trainset.

Label	Catalan	Spanish
AGAINST	120	1,785
FAVOR	4,085	1,680
NEUTRAL	479	972
Total	4,684	4,437

Table 8
Examples of the four types of text pre-preprocessing.

Pre-processing type	Result
Original	@pilarc_pilarc Ten, manipuladora te cayó el ME ESTAS HABLANDO EN POLACO?? que le suelta el fachamierda primero #ZASCA https://t.co/XQ08KuVgtI
Type A	manipulador cayo hablar polaco suelto fachamierda #zasca
Type B	ten manipuladora se te cayó el me estas hablando en polaco que le suelta el fachamierda primero #zasca
Type C	pilarcpilarc Ten manipuladora se te cayó el ME ESTAS HABLANDO EN POLACO que le suelta el fachamierda primero ZASCA
Type D	pilarc_pilarc Ten, manipuladora se te cayó el ME ESTAS HABLANDO EN POLACO?? que le suelta el fachamierda primero ZASCA

which was performed via crowdsourcing by eight different annotators. In addition to stance, annotations are provided to express whether the target is explicitly mentioned in the tweet. Table 1 shows some examples taken from the SemEval 2016 dataset.

The best result was reported by the organizers of the task with a system based on character and word n-grams to train a linear SVM model, achieving 68.98% in F1 average score (Mohammad et al., 2016). More recently, Ghosh et al., 2019 fine-tuned the pre-trained BERT Large model (Devlin et al., 2019) obtaining the highest performance so far on this dataset, with a F1 average score of 75.1.

4.1.2. TW-1O Referendum

The "MultiModal Stance Detection in tweets on Catalan #1Oct Referendum" shared task at IberEval 2018 (MultiStanceCat) proposed to detect stance (FAVOR, AGAINST, NEUTRAL) on political discourse with respect to the Referendum on the Independence of Catalonia held on the first of October, 2017. The dataset is multilingual (Catalan and Spanish) and includes images to facilitate multimodal experimentation (Taulé et al., 2018).

The dataset was collected using the hashtags #1oct, #1O, #1oct2017 and #1octl6 to search for messages in Twitter, widely used in the dates previous to the Referendum. A total of 87,449 tweets in Catalan and 132,699 tweets in Spanish were collected between between September 20–30, 2017. After various pre-processing steps, the final dataset consists of 11,398 tweets: 5,853 written in Catalan (the TW-1O-CA) and 5,545 in Spanish (the TW-1O-ES). The dataset was annotated manually by three experts. Also, the previous and next messages are included as additional context to the original tweet. Contatenating the three tweets result in an approximate average length of 38 tokens per document.

Table 3 illustrates the classes distribution in the TW-10 data. While for Spanish the distribution of classes is quite balanced, in Catalan the FAVOR class occurs 35 times more than AGAINST, and 8 times more than NEUTRAL. Obviously, this hugely skewed dataset would make it difficult to build and compare models for Catalan and across languages. In this sense, a simple most frequent class baseline for Catalan would be correct in around 95% of the cases.

4.2. Data Pre-processing

Each of the system types mentioned above benefit from different pre-processing strategies. We believe this is particularly important as the type of pre-processing has a huge influence in the final performance. We followed four different pre-processing strategies, illustrated by Table 8.

Type A: Punctuation, URLs, retweets (RTs), Twitter usernames, hashtags, digits, stopwords, words shorter than three characters, diacritics and emojis are removed. Furthermore, multiple character repetition is replaced with a single chararacter. The text is lemmatized and lowercased. Lemmatization is performed via dictionary look-up[9]. Note that this lemmatization method does not deal with ambiguity.

Type B: We remove punctuation, URLs, RTs, digits, Twitter usernames, hashtags and emojis. Repeated characters are simplified into a single character. Text is lemmatized and lowercased.

Type C: Punctuation, Twitter usernames, hashtags and URLs are removed.

Type D: Minimal pre-processing: Twitter usernames, hashtags and URLs are removed.

4.3. TF-IDF + SVM

TF-IDF (Term Frequency times Inverse Document Frequency) (Jones, 1972) is a weighting scheme broadly used in many tasks. Its goal is to reduce the impact of words that occur too frequently in a given corpus. TF-IDF is the product of two metrics, the term frequency and the inverse document frequency. We calculate the TF-IDF scores for all pre-processed unigrams in the training corpus. The number of features equals the size of the vocabulary of the dataset and represents the dimensionality of the document vector. The TF-IDF vectorizing is applied over the text pre-processed following Type A strategy.

We also used Information Gain (Cover & Thomas, 2006) for feature selection and Grid search for hyperparameter optimization. The Information Gain scores show how common a specific feature is in a target class. For example, those words that occur mainly in tweets labelled as FAVOR will be highly ranked. All the weights are normalized and the features ranked from one to zero. We then select those features that are larger than zero. Grid search is used to tune two SVM (RBF kernel) hyperparameters, namely, C and gamma.

4.4. FastText Embeddings + SVM

Word embeddings encode words as continuous real-valued representations in a low dimensional space. Word embedding models are pre-trained over large corpora and are able to capture semantic and syntactic similarities based on co-ocurrences.

FastText distributes static word embeddings for our languages of interest, namely, Catalan, English and Spanish (Grave et al., 2018). Initial experimentation showed that the Common Crawl[10] models performed better for the stance detection task. The Common Crawl models are trained using a Continuous Bag-of-Words (CBOW) architecture with position-weights and 300 dimensions on a vocabulary of 2 M words. In order to produce vectors for out-of-vocabulary words, fastText word embeddings are trained with character n-grams of length 5, and a window of size 5 and 10 negatives (Grave et al., 2018). We represent the tweet as the average of its word vectors (Kenter, Borisov, & de Rijke, 2016). In order to facilitate the look-up into the pre-trained word embedding model, we use the Type B pre-processing strategy. As for the previous system, C and gamma hyperparameters (SVM RBF kernel) are configured via grid search.

[9] https://github.com/michmech/lemmatization-lists
[10] http://commoncrawl.org/

4.5. FastText System

Apart from the pre-trained word embedding models, fastText also refers to a text classification system (Joulin et al., 2017). The fastText system consists of a linear model with rank constraint. First a weight matrix A is build via a look-up table over the words. Then the word representations are averaged to construct the tweet representation, which is then fed into a linear classifier. This is similar to the previous approach, but in the fastText system the textual representation of the tweet is a hidden variable which can be reused. The CBOW model proposed by Mikolov, Chen, Corrado, and Dean, 2013 is similar to this architecture, with the difference that the middle word is replaced by the stance label. Finally, fastText uses a softmax function to calculate the probability distribution over the predefined classes.

4.6. Transformers

BERT (Devlin et al., 2019) is a pre-trained mask language model based on the transformer architecture (Vaswani et al., 2017) which has obtained very good results on many NLP tasks. The first multilingual version of such models was multilingual BERT (mBERT), a single language model pre-trained from corpora in more than 100 languages. Another well-known model is XLM-RoBERTa (Conneau et al., 2019b), which provides a language model for 100 languages trained on 2.5 TB of Common Crawl text. Both mBERT and XLM-RoBERTa allow to perform cross-lingual knowledge transfer (Heinzerling & Strube, 2019; Pires, Schlinger, & Garrette, 2019; Karthikeyan, Wang, Mayhew, & Roth, 2020), namely, these systems can be applied to generate predictions on datasets for languages different to the ones used to fine-tune them. Thus, in this paper we will use both mBERT and XLM-RoBERTa for fine-tuning in Catalan and Spanish but also in cross-lingual experiments using the CIC corpus. Additionally, we will use two other transformers which offer pre-trained English language models: RoBERTa and XLNet.

RoBERTa (Liu et al., 2019) is an improved, optimized version of BERT. The model was trained using BERT architecture with larger batches and much more data (around 160 GB of English texts). Furthermore, the next sentence prediction objective used for pre-training the language model is discarded.

XLNet (Yang et al., 2019) is an auto-regressive method based on permutation language modelling (Uria, Côté, Gregor, Murray, & Larochelle, 2016) without using any masking symbols. XLNet integrates two-stream self-attention and a Transformer-XL architecture (Dai et al., 2019) to the pre-training process with the objective of learning long-range dependencies.

Every experiment with the transformers uses the datasets pre-processed following the Type D strategy, which minimally removes hashtags, Twitter usernames and URLs from the tweets.

Furthermore, we use the base version of each transformer so that we can fine-tune the pre-trained model in a basic GPU of 12 GB RAM. For each dataset, we tune the hyperparameters (batch size, minimum sequence length, learning rate and number of epochs) on the development data if available, otherwise, the training set is used both for training and development.

4.7. Evaluation

The models are evaluated with the metric and script provided by the organizers of SemEval 2016 (Mohammad et al., 2016) which reports F1 macro-average score of two classes: FAVOR and AGAINST, although the NONE class is also represented in the test data:

$$F1_{avg} = \frac{F1_{favor} + F1_{against}}{2}$$

Table 9
Overall results for the SemEval 2016 English dataset.

System	$F1_{against}$	$F1_{favor}$	$F1_{avg}$
TF-IDF + SVM	73.24	53.52	63.38
FTEmb + SVM	72.06	53.56	62.81
FTEmb + fastText	72.94	61.58	67.26
mBERT (msl 256, batch 32, lr 5e-5, 5e)	73.90	62.29	68.10
XLM-R (msl 128, batch 16, lr 2e-5, 10e)	75.30	65.03	70.17
XLNet (msl 128, batch 16, lr 2e-5, 10e)	76.23	66.57	71.40
RoBERTa (msl 256, batch 16, lr 2e-5, 10e)	76.87	67.43	**72.15**
Ghosh et al., 2019 (msl 128, batch 16, lr 2e-5, 50e)	-	-	**75.10**

Table 10
SVM hyperparameters for the SemEval 2016 English dataset.

	TF-IDF + SVM		FTEmb + SVM	
Target	C	Gamma	C	Gamma
Atheism	700	0.001	100	0.1
Climate Change	700	0.001	10	0.75
Feminist Movement	700	0.001	10	1.0
Hillary Clinton	1000	0.0001	10	0.75
Legalization of Abortion	700	0.001	10	1.0

5. Experimental Results

In this section we report the results obtained by applying the experimental setup from Section 4 to the datasets described in Section 3.

More specifically, in this section we will show, via experimentation, that our proposed method to develop the CIC corpus generates multilingual datasets for stance detection in Twitter which are at least as reliable for experimentation as manually annotated ones. In this sense, our method would help to facilitate the development of new multilingual and cross-lingual approaches for stance detection.

In the following, we first report the results obtained for each of the datasets in a standard in-domain setting. Finally, we perform cross-lingual experiments on the CIC corpus using multilingual language models (mBERT and XLM-RoBERTa).

5.1. SemEval 2016

In Table 9 we can see the results of the experiments performed on the SemEval 2016 dataset. The first three rows refer to the systems based on linear classification, namely, SVM using TF-IDF 4.3, SVM with averaged fastText embeddings 4.4, and the fastText system itself 4.5. The next four rows provide the results obtained using the monolingual (XLNet and RoBERTa) and multilingual (mBERT and XLM-RoBERTa) Transformer-based pre-trained language models.

For the SVM-based systems, we performed grid search on the training data to find optimal values for the hyperparameters C and gamma. These can be seen in Table 10.

The fastText system is used off-the-shelf, training domain-specific word embeddings on each target's training set with one exception: after some experimentation on the "Feminist Movement" train set, we increased the number of epochs to 60.

The train set from the "Feminist Movement" target was also used to obtain the hyperparameters for mBERT, XLNet, RoBERTa and XLM-RoBERTa. Thus, for mBERT we used a maximum sequence length of 256, 32 batch size, 5e-5 learning rate and 5 epochs.

The results show that, as it has been the case for other text classification tasks (Devlin et al., 2019; Yang et al., 2019) the Transformers-based pre-trained language models outperform any other approach. In fact, we obtain state-of-the art results, improving over any other approach presented in Section 2, using the *base* version of RoBERTa (Liu et al., 2019). Ghosh et al., 2019 obtains best overall results by applying BERT Large uncased with the hyperparameters specified in the last row

Table 11
Results on the TW-1O Catalan testset.

System	$F1_{against}$	$F1_{favor}$	$F1_{avg}$
TF-IDF + SVM (C = 100, Gamma = 0.01)	22.86	94.68	58.77
FTEmb + SVM (C = 10, Gamma = 1)	0.00	93.88	46.94
FTEmb + fastText	12.90	94.60	53.78
mBERT (msl 128, batch 32, lr 2e-5, 10e)	28.57	94.33	**61.45**
mBERT ca + es (msl 128, batch 32, lr 2e-5, 10e)	05.71	94.03	49.87
XLM-R (msl 128, batch 32, lr 2e-5, 10e)	21.62	94.86	58.24
Baseline			
Cuquerella and Rodríguez, 2018	-	-	30.68

Table 12
Results on the TW-10 Spanish testset.

System	$1_{against}$	$F1_{favor}$	$F1_{avg}$
TF-IDF + SVM (C = 500, Gamma = 0.001)	68.50	64.53	66.52
FTEmb + SVM (C = 300, Gamma = 0.75)	63.65	58.85	61.25
FTEmb + fastText	69.58	65.37	**67.48**
mBERT (msl 256, batch 32, lr 5e-5, 5e)	66.80	65.11	65.96
mBERT ca + es (msl 256, batch 32, lr 5e-5, 5e)	66.67	62.16	64.42
XLM-R (msl 256, batch 32, lr 2e-5, 10e)	65.54	59.26	62.40
Baseline			
Segura-Bedmar, 2018	-	-	28.02

of Table 9. This suggests that these large pre-trained language models require less training data than classic machine learning approaches. Finally, they also show that for fine-tuning on small datasets it is convenient to increase the number of epochs.

5.2. *IberEval 2018*

Tables 11 and 12 report our results for Catalan and Spanish respectively. The hyperparameters were chosen following the procedure explained for the SemEval 2016 dataset. The only difference in the setup refers to the fastText system. While for the SemEval data the word embeddings were directly trained on the stance training data, for this particular dataset results were better if we used the fastText pre-trained word embedding models from Common Crawl Grave et al., 2018. Furthermore, we trained fastText for 20 epochs.

With respect to the results for the Catalan language, it seems to us that its skewed class distribution is the most important issue, given that every system struggles to correctly predict the *against* class. The best results are obtained by mBERT, with a very low 28.57 $F1_{against}$ score.

For Spanish the results are more balanced. Interestingly, for this language the fastText linear classifier combined with fastText embeddings (FTEmb + fastText) obtains better results than mBERT or XLM-RoBERTa.

We also tried by to augment the training data available to fine-tune the multilingual pre-trained models by concatenating the training sets of both languages (e.g.. see “ca + es” results). However, this strategy was not beneficial.

In any case, our results substantially improve over previous state-of-the-art in both languages. However, it should be noted that they are comparatively lower than those obtained with the English SemEval 2016 data. In spite of this, the systems ranking across the English and Spanish datasets is quite similar, the Catalan results being the exception.

At this point, one question is whether the results obtained by mBERT and XLM-RoBERTa are lower for Catalan because that language is not as well represented as English or Spanish in the multilingual language models, or whether it is just due to the skewed distribution of classes. In the next sections we will look into these and other research questions using our CIC corpus.

Table 13
Results on the Catalan testset of the Catalonia Independence Corpus (CIC–CA).

	CIC			CIC-Random		
System	$F1_{against}$	$F1_{favor}$	$F1_{avg}$	$F1_{against}$	$F1_{favor}$	$F1_{avg}$
TF-IDF + SVM	68.89	72.91	70.90	70.20	72.49	71.35
FTEmb + SVM	59.43	64.46	61.95	67.91	65.77	66.84
FTEmb + fastText	70.73	72.21	71.47	69.62	70.63	70.13
mBERT	70.64	77.52	74.08	69.98	75.60	72.79
mBERT ca + es	53.13	77.98	65.56	60.90	77.52	69.21
XLM-R	70.69	78.67	**74.68**	71.63	78.10	**74.87**

Table 14
Results on the Spanish testset of the Catalonia Independence Corpus (CIC-ES).

	CIC			CIC-Random		
System	$F1_{against}$	$F1_{favor}$	$F1_{avg}$	$F1_{against}$	$F1_{favor}$	$F1_{avg}$
TF-IDF + SVM	70.67	71.50	71.09	69.56	74.12	**71.84**
FTEmb + SVM	64.24	62.51	63.38	64.39	66.99	65.69
FTEmb + fastText	73.20	71.13	72.43	69.76	71.22	70.49
mBERT	75.17	74.27	**74.72**	70.20	71.35	70.78
mBERT ca + es	69.54	71.83	70.69	68.02	72.86	70.44
XLM-R	74.68	72.45	73.57	70.01	70.75	70.38

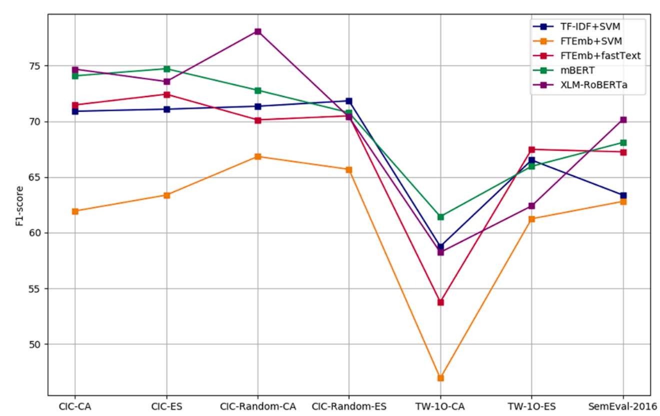

Fig. 2. Systems' behaviour across the datasets.

5.3. *CIC Corpus*

The results obtained with both versions of the CIC corpus can be seen in Tables 13 and 14. Hyperparameters were chosen on the development set: (i) for TF-IDF + SVM C and Gamma values were 500 and 0.001, respectively; (ii) FTEmb + SVM, C = 100 and Gamma = 1; (iii) for fastText we used the setting described for TW-1O; (iii) mBERT and XLM-RoBERTa were fine-tuned over 10 epochs using the following settings: maximum sequence length 128, batch 32, learning rate 2e-5.

The results from Tables 13 and 14 arise several issues. First, it is clear that there is a consistency in the behaviour of the systems across both languages. In fact, it turns out that for the CIC *random* version of the dataset results are higher in Catalan. Second, systems display similar behaviour across both versions of the CIC dataset. This is also true with respect to the English Semeval 2016 results: the best linear classifier is usually fastText whereas the Transformer models obtain the best overall results. Therefore, this means that there is not any flaw in our method for the creation of Twitter-based datasets that may cause the systems to overfit on the tweets from a specific author's writing style, and so on. Third, augmenting the training data (“ca + es”) does not help. Finally, both the Catalan and Spanish results are higher across languages and systems than those obtained using the TW-1O and SemEval 2016 datasets. These general patterns are visualized in Fig. 2. Thus, while for CIC

Table 15
Comparing zero-shot or translating across languages on CIC corpus.

	Zero-shot			Translate		
	$F1_{against}$	$F1_{favor}$	$F1_{avg}$	$F1_{against}$	$F1_{favor}$	$F1_{avg}$
CIC Corpus						
mBERT ca-es	46.15	56.97	**51.56**	46.51	54.75	50.63
mBERT es-ca	22.07	65.47	43.77	42.92	60.73	**51.83**
XLM-R ca-es	48.91	56.65	**52.78**	44.80	55.69	50.25
XLM-R es-ca	33.61	62.82	48.22	47.21	59.75	**53.48**
CIC-Random Corpus						
mBERT ca-es	46.96	58.18	**52.57**	45.33	56.29	50.81
mBERT es-ca	33.20	60.88	47.04	50.49	55.20	**52.85**
XLM-R ca-es	49.43	54.64	**52.04**	43.35	57.71	50.53
XLM-R es-ca	27.87	53.10	40.49	50.75	56.58	**53.67**

and SemEval the systems display similar behaviour, with the TW-1O they are much more unstable.

In terms of specific results, it was unexpected that TF-IDF + SVM would outperform every other approach (in CIC-Ramdom-ES). Furthermore, it is also surprising that mBERT would obtain better results than XLM-RoBERTa for Spanish; this was also the case with the TW-1O Spanish data. In addition, in TW-1O the *favor* class seems easier to learn than *against*.

We believe that the experimental results here presented allows us to conclude that our semi-automatic method for the annotation of tweets presented in this paper is faster and produces annotations of comparable quality with the human annotated data at tweet level.

5.4. Zero-Shot vs Translation

For our final set of experiments, we apply mBERT and XLM-RoBERTa in a zero-shot scenario. The idea is the following: assuming that we do not have annotated data for stance detection in a given language, which would be the optimal strategy? To answer this question, we consider two alternatives: (i) fine-tune the models in a given language and predict in the other (zero-shot approach) or, (ii) fine-tune in the language for which we do have annotated data and predict in the machine translated version of the other language's test (translation approach).

Table 15 reports the result of this experiment. The *zero-shot* and *translate* multicolumns contain the scores obtained when fine-tuning on a language and predicting in the other. For example, the first zero-shot row means that mBERT was fine-tuned on the Catalan CIC training set and evaluated on the CIC-ES test, obtaining 51.56 $F1_{avg}$ score. Its corresponding *translate* result means that mBERT was fine-tuned in Catalan and that the prediction was performed on the translated (into Catalan) CIC-ES test set, obtaining 50.63 $F1_{avg}$ score. For the automatic translation of the target language test sets, we used the MarianMT system (Junczys-Dowmunt et al., 2018) via the Hugginface Transformers API (Wolf et al., 2019) which offers MarianMT models trained on the OpusMT corpus (Tiedemann, 2012).

By looking at Table 15 we can see that the *translation* approach works better than applying *zero-shot* whenever the source language is Spanish, namely, by fine-tuning on Spanish and predicting on the translated Catalan test set. We believe that this responds to two claims already mentioned in the literature.

First, while these deep learning multilingual models performed very well for tasks involving high-resourced languages such as English, their performance drops when applied to low-resource languages (Agerri et al., 2020). As languages share the quota of substrings, which partially depends on corpus size, larger languages such as Spanish may be better represented than lower resourced languages such as Catalan.

Second, it has also been claimed that these multilingual models seem to behave better for structurally similar languages (Karthikeyan et al., 2020).

Our results reinforce these two separate claims, since the *zero-shot* approach works quite well when fine-tuning on Catalan and making the predictions in Spanish (similar languages), and translating is preferable if the source language is a large and well represented language (such as Spanish). By plotting the results in Fig. 3, it is easier to confirm this and appreciate two further issues. On the one hand, *translation* results are more stable because for "ca-es" we also fine-tune on the low-resource language. On the other, the *zero-shot* approach suffers when predicting onto low-resourced languages. Following this line of argumentation, we also experimented by translating from Catalan and Spanish into English, but results were substantially worse, possibly because performance from Catalan to Spanish and viceversa benefits from their similar grammatical structure.

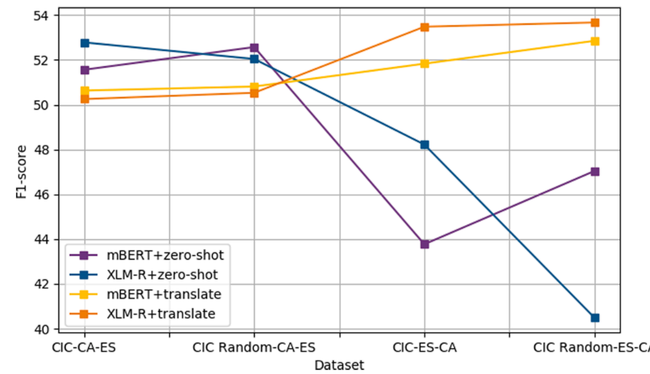

Fig. 3. Comparison of zero-shot and translation approaches.

Table 16
Measuring loss after translation.

	CIC			CIC-Random		
	$F1_{against}$	$F1_{favor}$	$F1_{avg}$	$F1_{against}$	$F1_{favor}$	$F1_{avg}$
mBERT **ca-es**	69.12	74.84	**71.98**	68.21	70.72	**69.47**
mBERT es-ca	72.60	71.00	71.80	67.17	69.73	68.45
XLM-R **ca-es**	67.14	75.12	**71.13**	67.01	73.40	**70.21**
XLM-R es-ca	70.21	69.97	70.09	67.51	68.89	68.20

As a final experiment, we wanted to know how much does machine translation affect the results reported in Table 15 and Fig. 3. In order to quantitatively assess this, we fine-tune in a given language and dataset and evaluate on its translated test set (in a zero-shot manner). For example, the first row in Table 16 means that we fine-tuned mBERT with the CIC–CA training set and evaluated it on the translated test set. The setting is the same as in Section 5.3, but making the predictions on the translated test.

As it was the case for the *zero-shot* and *translation* comparison, translating into Spanish produces better results. We attribute this to the the fact that, while being quite similar, Spanish is a better represented language in the multilingual pre-trained models. In any case, it is remarkable how good the performance of mBERT and XLM-RoBERTa is across languages, losing just around 3 points in $F1_{avg}$ score.

6. Error Analysis

After showing in the previous section, by means of quantitative results, that our method generates good quality annotated data for multilingual experimentation on stance detection in Twitter, in this section we offer a qualitative analysis to better understand the features of the Catalonia Independence Corpus (CIC).

In order to do so, we manually revise and annotate a sample from the CIC training set in order to compare the semi-automatically obtained annotations with those given by humans. Furthermore, we inspect the output of the best classifiers, analyzing the tweets that were incorrectly labeled, and calculate an upperbound score to compare it with our best

Table 17
Example of a tweet categorized differently with our method and with human annotation.

Tweet 1	*Arrimadas irá a Waterloo este domingo para recordar a Puigdemont que la república no existe*
Our method	NONE
Human	AGAINST/FAVOR
Language	Spanish
Translation	*Arrimadas will go to Waterloo this Sunday to remind Puigdemont that the Republic does not exist*
Tweet 2	*@unprecisionman @jordisalvia Quan l'advocat preguntava sobre certes contradiccions d'un incident concret q havia explicat el Millo, el jutge ha dit q això no era rellevant per la causa*
Our method	AGAINST
Human	FAVOR/NONE
Language	Catalan
Translation	*When the lawyer asked him about certain contradictions with respect to a specific incident which Millo had explained, the judge said that it was not relevant.*

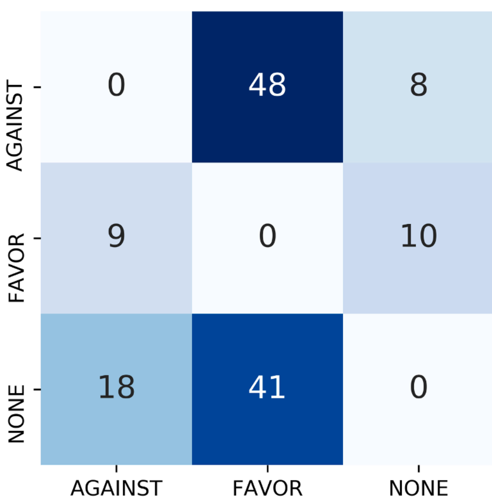

Fig. 4. Confusion matrix for CIC–CA made with majority voting.

trained models.

6.1. Annotation Errors

The manual error analysis of the semi-automatic annotation of the CIC corpus was carried out as follows. We took a random sample of 100 tweets per language from the training sets. The obtained sample was then manually revised by two human annotators. After independently annotating each of the samples, the annotators tried to agreed upon a common label for the contentious examples. Overall, we found out that the error rate in the Spanish sample was around 20%, whereas for the Catalan sample was slightly higher, around 35%. It should be noted that the annotators found it very difficult to agree on their correct annotation. This was due to several reasons.

First, the meaning of the tweets is usually under-specified, namely, there is not enough context or background information available in order to take an informed decision. Second, many tweets use figurative language such as sarcasm, irony or require extra commonsense and/or domain-specific knowledge. Finally, other tweets referred to the topic in a indirect manner without clearly establishing a given stance with respect to the topic.

Table 17 offers two examples of contentious tweets for which it is quite difficult to decide whether our semi-automatic method provided a correct annotation or not. Tweet 1 is classified as NONE in the CIC dataset, but it is possible to assign both AGAINST (assuming that the writer supports Arrimadas's action) and FAVOR (if the message was to

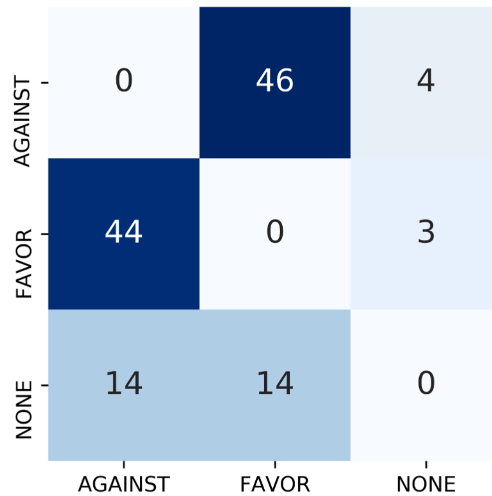

Fig. 5. Confusion matrix for CIC-ES made with majority voting.

be a case of irony). In Tweet 2, although it seems to be of type NONE, a case can be made for it to be both in FAVOR (assuming that the Spanish judge has an anti-independence bias) and AGAINST (in this case the tweet would be agreeing with the judge's decision).

This manual annotation exercise showed that labelling stance in tweets is a difficult task for humans, partly because it depends greatly on the annotator's background knowledge and intuition. In addition, annotating tweets one by one, as opposed to our user-based annotation, very often suffers from a lack of context.

6.2. Prediction Errors

For this analysis we chose the five best classifiers per language from Tables 13 and 14 and identified those tweets that were incorrectly labeled by at least three of those five classifiers. In total we obtained 134 tweets in Catalan and 125 in Spanish.

The type of errors are shown in Figs. 4 and 5. It can be seen that the most frequent error for both languages is AGAINST being predicted as FAVOR. Furthermore, the second most common source of errors is for Catalan NONE being predicted as FAVOR whereas for Spanish is FAVOR mistakenly predicted as AGAINST.

After manual inspection of the misclassified tweets these were the most common sources of error:

- **Annotation error**: Sometimes when users retweet or quote a tweet expressing the opposite political stance without further comment. This causes our semi-automatic method to generate wrong labels.
- **Underspecification**: The tweet is just too short, or the target is not explicitly mentioned or referred to.
- **Missing conversation structure**: Replies to an unknown trigger tweet are often difficult to label.

Tweet 3 provides an example of a quotation of a message expressing the opposite stance. In this particular case the message is using content or words related to an AGAINST stance to express FAVOR.

Tweet 3: *"Joder con los indepes que no se venden como hacía Pujol". Voy a informarme de cuál es el grado de cumplimiento de las promesas económicas a los catalanes en general y de los PGE en particular. Por qué no reclamas que te hostien y luego te prometan 4 de tus perras? https://t.co/842QMqkj2W*
Label: FAVOR.
Automatic classification: AGAINST (all systems).
Language: Spanish.

Table 18
Upperbound obtained from the predictions of the best 5 systems.

	CIC			CIC-Random		
System	$F1_{against}$	$F1_{favor}$	$F1_{avg}$	$F1_{against}$	$F1_{favor}$	$F1_{avg}$
Upperbound CA	94.44	93.68	94.06	91.70	92.53	92.12
XLM-R	70.69	78.67	74.68	71.63	78.10	74.87
Upperbound ES	93.63	93.71	93.67	93.23	93.67	93.45
mBERT	75.17	74.27	74.72	70.20	71.35	70.78

Translation: *"Bloody independentists who don't sell themselves out like Pujol did." I'm going to find out what is the degree of fulfilment of the economic promises made to the Catalans in general and with respect to the Budget in particular. Why don't you ask to be beaten and then let them promise you four cents? https://t.co/842QMqkj2W*

6.3. Upperbound Score

In order to understand how much room for improvement there is in the CIC dataset, we calculated an upperbound score consisting of assigning a given label if at least one of the five best systems predicted it correctly.

Table 18 shows that the gap between the best results and the upperbound scores is quite large, which means that we still have a large margin for developing better stance detection systems in this particular dataset.

7. Conclusion

In this paper we have shown that our methodology to build annotated datasets for multilingual and cross-lingual stance detection in Twitter helps to alleviate a number of problems present in previous manual-based efforts. Our method to build the *Catalonia Independence Corpus* (CIC) is faster and requires less manual effort while obtaining larger and more balanced datasets. Furthermore, we have empirically demonstrate that the behaviour of the systems evaluated on the CIC data is consistent with respect to previous benchmarks (e.g.. SemEval 2016 and TW-1O). In fact, overall results are higher on the CIC corpus. We have discarded that those results were due to any overfitting to users' idiosyncrasies in their writing style by creating the CIC-Random version of the corpus. Moreover, a qualitative analysis and manual annotation of a corpus sample showed that the obtained semi-automatic annotations are of comparable, if not better, quality than the manual ones. We attribute this to the user-based nature of our method, which helps to overcome the under-specification or lack of context of the individual tweets.

The availability of the CIC corpus has also allowed us to explore cross-lingual approaches, comparing zero-shot with translation based strategies. These experiments have also provided insights about the behaviour of large pre-trained multilingual language models.

We believe that our method would be helpful to obtain good quality annotated data quicker and more efficiently for many types of Social Media Analysis and Natural Language Processing tasks which use Twitter data. In this sense, in future work we would like to explore the application of our method to various types of tasks and domains. We see two possible research avenues: (i) on-demand generation of domain-specific annotated data and, (ii) applying our method to open stance detection, namely, to generate labeled data to classify the stance of a message not to a specific target, but to a previous message, news headline or tweet.

CRediT authorship contribution statement

Elena Zotova: Conceptualization, Methodology, Software, Data curation, Writing - original draft, Validation. **Rodrigo Agerri:** Conceptualization, Methodology, Software, Data curation, Writing - original draft, Supervision, Writing - review & editing, Software, Validation. **German Rigau:** Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work has been partially funded by the Spanish Ministry of Science, Innovation and Universities (DeepReading RTI2018-096846-B-C21, MCIU/AEI/FEDER, UE), *Ayudas Fundación BBVA a Equipos de Investigación Científica 2018* (BigKnowledge), and DeepText (KK-2020/00088), funded by the Basque Government. Rodrigo Agerri is also funded by the RYC-2017–23647 fellowship and acknowledges the donation of a Titan V GPU by the NVIDIA Corporation.

References

Agerri, R., San Vicente, I., Campos, J. A., Barrena, A., Saralegi, X., Soroa, A., & Agirre, E. (2020). Give your Text Representation Models some Love: the Case for Basque. In *Proceedings of The 12th Language Resources and Evaluation Conference* (pp. 4781–4788). Marseille, France: European Language Resources Association.

AlDayel, A., & Magdy, W. (2019). Your Stance is Exposed! Analysing Possible Factors for Stance Detection on Social Media. *Proceedings of the ACM on Human-Computer Interaction, 3*. https://doi.org/10.1145/3359307

AlDayel, A., & Magdy, W. (2020). Stance Detection on Social Media: State of the Art and Trends. arXiv preprint arXiv:2006.03644.

Augenstein, I., Rocktäschel, T., Vlachos, A., & Bontcheva, K. (2016). Stance detection with bidirectional conditional encoding, in. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing Austin, Texas* (pp. 876–885).

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

Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political Analysis, 23*, 76–91.

Benton, A. & Dredze, M. (2018). Using Author Embeddings to Improve Tweet Stance Classification. In: Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text, Association for Computational Linguistics, Brussels, Belgium. pp. 184–194.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Bøhler, H., Asla, P., Marsi, E., & Sætre, R. (2016). IDI@NTNU at SemEval-2016 Task 6: Detecting Stance in Tweets Using Shallow Features and GloVe Vectors for Word Representation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* (pp. 445–450). San Diego, California: Association for Computational Linguistics. https://doi.org/10.18653/v1/S16-1072.

Boyd, D., Golder, S., & Lotan, G. (2010). Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter, in: Proceedings of the 2010 43rd Hawaii International Conference on System Sciences, IEEE Computer Society. p. 1–10.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2019a). Unsupervised cross-lingual representation learning at scale. ArXiv abs/1911.02116.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2019b). Unsupervised Cross-lingual Representation Learning at Scale. arXiv preprint arXiv:1911.02116.

Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. New York, NY, USA: Wiley-Interscience.

Cuquerella, C.A. & Rodríguez, C.C. (2018). Crica team: Multimodal stance detection in tweets on catalan 1oct referendum (multistancecat), in: IberEval@SEPLN, pp. 167–172.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J.G., Le, Q.V., & Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context. In: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pp. 2978–2988.

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

Devlin, J., Chang, M., Lee, K., & Toutanova, K., (2019). BERT: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, Volume 1 (Long and Short Papers), pp. 4171–4186.

Du, J., Xu, R., He, Y., & Gui, L. (2017). Stance classification with target-specific neural attention networks, in: Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI 2017, pp. 3988–3994.

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

Fraisier, O., Cabanac, G., Pitarch, Y., Besançon, R., & Boughanem, M., (2018). Stance Classification through Proximity-Based Community Detection, in: Proceedings of the 29th on Hypertext and Social Media, Association for Computing Machinery, New York, NY, USA. p. 220–228.

Ghosh, S., Singhania, P., Singh, S., Rudra, K., & Ghosh, S. (2019). Stance detection in web and social media: a comparative study. In *International Conference of the Cross-Language Evaluation Forum for European Languages* (pp. 75–87). Springer.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7–12, 2018, pp. 3483–3487.

Heinzerling, B., & Strube, M. (2019). Sequence Tagging with Contextual and Non-Contextual Subword Representations: A Multilingual Evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistic* (pp. 273–291). Florence, Italy: Association for Computational Linguistics.

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

Himelboim, I., McCreery, S., & Smith, M. (2013). Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on twitter. *Journal of Computer-Mediated Communication, 18*, 154–174.

Igarashi, Y., Komatsu, H., Kobayashi, S., Okazaki, N., & Inui, K. (2016). Tohoku at SemEval-2016 Task 6: Feature-based Model versus Convolutional Neural Network for Stance Detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* (pp. 401–407). San Diego, California: Association for Computational Linguistics.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28*, 11–21.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2* (pp. 427–431). Valencia, Spain: Short Papers.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A.F., Bogoychev, N., Martins, A.F.T., & Birch, A. (2018). Marian: Fast neural machine translation in C++, in: Proceedings of ACL 2018, System Demonstrations, Association for Computational Linguistics, Melbourne, Australia. pp. 116–121.

Karthikeyan, K., Wang, Z., Mayhew, S., & Roth, D. (2020). Cross-lingual ability of multilingual BERT: An empirical study. In *International Conference on Learning Representations (ICLR)* (pp. 1–8).

Kenter, T., Borisov, A., & de Rijke, M. (2016). Siamese CBOW: optimizing word embeddings for sentence representations. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7–12, 2016, Berlin, Germany, Volume 1: Long Papers, pp. 941–951.

Küçük, D., & Can, F. (2020). Stance Detection: a Survey. *ACM Computing Surveys (CSUR), 53*, 1–37.

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

Lai, M., Cignarella, A.T., & Farías, D.I.H, (2017). itacos at ibereval2017: Detecting stance in catalan and spanish tweets, in: IberEval@SEPLN, pp. 1–8.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. ArXiv abs/1907.11692.

McCallum, A.K. (2002). MALLET: A Machine Learning for Language Toolkit. Http://mallet.cs.umass.edu.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In: 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2–4, 2013, Workshop Track Proceedings, pp. 1–12.

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. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* (pp. 420–427). San Diego, California: Association for Computational Linguistics.

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., & Cherry, C. (2016). Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), San Diego, California* (pp. 31–41).

Mohammad, S. M., Sobhani, P., & Kiritchenko, S. (2017). Stance and Sentiment in Tweets. *ACM Transactions on Internet Technology, 17*, 26:1–26:23.

Mohtarami, M., Glass, J., & Nakov, P. (2019). Contrastive Language Adaptation for Cross-Lingual Stance Detection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 4441–4451). Hong Kong, China: Association for Computational Linguistics.

Otte, E., & Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. *Journal of Information Science, 28*, 441–453.

Pires, T., Schlinger, E., & Garrette, D. (2019). How Multilingual is Multilingual BERT?. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4996–5001).

Segura-Bedmar, I. (2018). Labda's early steps toward multimodal stance detection. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018) September 18th, 2018, Sevilla, Spain* (pp. 180–186).

Siddiqua, U.A., Chy, A.N., & Aono, M. (2019). Tweet Stance Detection Using an Attention based Neural Ensemble Model. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota. pp. 1868–1873.

Sun, Q., Wang, Z., Zhu, Q., & Zhou, G. (2018). Stance detection with hierarchical attention network. In *Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, New Mexico, USA* (pp. 2399–2409).

Taulé, M., Martí, M.A., Rangel, F., Rosso, P., Bosco, C., & Patti, V. (2017). Overview of the task on stance and gender detection in tweets on catalan independence ibereval 2017. In: Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017), Murcia, Spain. pp. 158–177.

Taulé, M., Rangel, F., Martí, M.A., & Rosso, P. (2018). Overview of the task on multimodal stance detection in tweets on catalan 1oct referendum. In: IberEval 2018. CEUR Workshop Proceedings. CEUR-WS.org, Sevilla, Spain. pp. 149–166.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 2214–2218). Istanbul, Turkey: European Language Resources Association (ELRA).

Uria, B., Côté, M. A., Gregor, K., Murray, I., & Larochelle, H. (2016). Neural Autoregressive Distribution Estimation. *Journal of Machine Learning Research, 17*, 7184–7220.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4–9 December 2017, Long Beach, CA, USA* (pp. 5998–6008).

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

Wang, R., Zhou, D., Jiang, M., Si, J., & Yang, Y. (2019). A survey on opinion mining: from stance to product aspect. *IEEE Access, 7*, 41101–41124.

Wei, P., Mao, W., & Zeng, D. (2018). A Target-Guided Neural Memory Model for Stance Detection in Twitter. In *2018 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8).

Wei, W., Zhang, X., Liu, X., Chen, W., & Wang, T. (2016). Pkudblab at SemEval-2016 task 6: A specific convolutional neural network system for effective stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)* (pp. 384–388). San Diego, California: Association for Computational Linguistics.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & Brew, J. (2019). HuggingFace's Transformers: State-of-the-art Natural Language Processing. ArXiv abs/1910.03771.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *NeurIPS*, 1–18.

Zarrella, G. & Marsh, A. (2016). MITRE at SemEval-2016 Task 6: Transfer learning for stance detection. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), Association for Computational Linguistics, San Diego, California. pp. 458–463.

Zotova, E., Agerri, R., Nuñez, M., & Rigau, G. (2020). Multilingual Stance Detection in Tweets: The Catalonia Independence Corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference* (pp. 1368–1375). Marseille, France: European Language Resources Association.

Zubiaga, A., Wang, B., Liakata, M., & Procter, R. (2019). Political homophily in independence movements: Analyzing and classifying social media users by national identity. *IEEE Intelligent Systems, 34*, 34–42.