Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1368–1375
Marseille, 11–16 May 2020

Multilingual Stance Detection: The Catalonia Independence Corpus

Elena Zotova[1], **Rodrigo Agerri**[1], **Manuel Nuñez**[2], **German Rigau**[1]
[1] IXA Group, HiTZ Centre, University of the Basque Country UPV/EHU, Donostia-San Sebastian, Spain
[2] Intercom Strategys, Madrid, Spain
zotova.el@gmail.com, rodrigo.agerri@ehu.eus, manuel.nunez@lunigtuk.com, german.rigau@ehu.eus

Abstract

Stance detection aims to determine the attitude of a given text with respect to a specific topic or claim. While stance detection has been fairly well researched in the last years, most the work has been focused on English. This is mainly due to the relative lack of annotated data in other languages. The TW-10 Referendum Dataset released at IberEval 2018 is a previous effort to provide multilingual stance-annotated data in Catalan and Spanish. Unfortunately, the TW-10 Catalan subset is extremely imbalanced. This paper addresses these issues by presenting a new multilingual dataset for stance detection in Twitter for the Catalan and Spanish languages, with the aim of facilitating research on stance detection in multilingual and cross-lingual settings. The dataset is annotated with stance towards one topic, namely, the independence of Catalonia. We also provide a semi-automatic method to annotate the dataset based on a categorization of Twitter users. We experiment on the new corpus with a number of supervised approaches, including linear classifiers and deep learning methods. Comparison of our new corpus with the with the TW-1O dataset shows both the benefits and potential of a well balanced corpus for multilingual and cross-lingual research on stance detection. Finally, we establish new state-of-the-art results on the TW-10 dataset, both for Catalan and Spanish.

Keywords: Stance Detection, Text Categorization, Less-Resourced Languages

1. Introduction

The rise of social media has given rise to the "fake news" phenomenon. According to the Fake News Challenge, "Fake news, defined by the New York Times as "a made-up story with an intention to deceive"[1], often for a secondary gain, is arguably one of the most serious challenges facing the news industry today."[2]

Determining the veracity of a given document or story, namely, whether it is fake or legitimate, is a very complex task, even for expert fact-checkers. Thus, previous work breaks down the fake news detection task in different stages, the first of which is establishing what other news sources are saying about the given document or story (whether they agree, disagree, etc. with the news story), namely, determining their stance with respect to that document or news story. Following this, the first stage of the Fake News Challenge was *Stance Detection*. This decision was supported by two main ideas: (i) a stance detection system should allow a human fact checker to enter a document (headline, message, claim, etc.) and retrieve the top documents from other news sources that agree, disagree or discuss the given document and, (ii) based on the previous step, it would be possible to build a "truth-labeling" system based on the weighted credibility of the various news organizations from which the stance has been retrieved.

Automatic stance detection has been defined as the task of classifying the attitude expressed in a text towards a given target or claim. Most of the work on stance detection has been undertaken in English using the data provided by the Detecting Stance in Tweets shared task organized at SemEval 2016 (Mohammad et al., 2016), RumourEval 2017 (Derczynski et al., 2017) and the Fake News Challenge. The SemEval 2016 task was formulated as follows: given a tweet text and a target entity or topic, automatic natural language systems must determine whether the tweet expresses a stance in **favor** of the given target, **against** the given target, or whether **none** of those inferences are likely. For example, consider the following target−tweet pairs:

> **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 examples illustrate the nature of the task, in which tweets are very short, full of specific vocabulary, non-standard spelling grammar, emojis, hashtags, and high on irony and sarcasm. The task aimed to detect stance from single tweets, without taking into account the conversational structure of tweet threads or any information about authors.

Following the model of the SemEval 2016 initiative, two shared tasks were organized as part of IberEval workshop (Taulé et al., 2017; Taulé et al., 2018). They provided tweets annotated for Stance in Catalan and Spanish. The target of the 2017 edition was the "Catalan Independence" whereas the 2018 edition (TW-10 dataset) focused on the "Catalan referendum on the 1st of October". In both editions the classes distribution was hugely skewed, which makes it difficult to explore and compare stance detection methods in multilingual and cross-lingual settings.

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

In this context, we propose the new Catalan Independence Corpus (CIC) for stance detection in Catalan and Spanish. By doing so, we aim to promote research in other languages different to English. Furthermore, the corpus presents a balanced distribution between classes so that researchers can explore multilingual and cross-lingual methods.

The contributions of this paper are the following: (i) we present a new dataset in Catalan and Spanish to work on multilingual and cross-lingual stance detection; (ii) we propose a semi-automatic method to collect and annotate a corpus of tweets based on a categorization of Twitter users. This method partially alleviates the huge effort of manually annotating the corpus tweet by tweet; (iii) we report new state-of-the-art results on the TW-10 dataset of IberEval 2018 (Taulé et al., 2018); (iv) comparison between results using our new corpus and the TW-10 dataset shows the benefits of providing a balanced multilingual corpus, and (v) both the datasets and code are made public to facilitate future research and reproducibility of results[3].

2. Related Work

The state of the art is divided into two main approaches. First, those that rely on *traditional* machine learning models combined with hand-engineered features (Mohammad et al., 2017) or vector-based word representations (word embeddings) (Bøhler et al., 2016). In particular, (Mohammad et al., 2017) obtained the best results for the supervised setting of the SemEval 2016 task using a SVM classifier to learn word n-grams (1-, 2-, and 3-gram) and character n-grams (2-, 3-, 4-, and 5-gram) features, outperforming deep learning approaches (Zarrella and Marsh, 2016; Wei et al., 2016).

Among the deep learning systems published, the pkudblab system (Wei et al., 2016) proposed a Convolutional Neural Network (CNN) architecture combined with a voting scheme to guide the predictions instead of generating them based on the accuracy obtained in the validation set. The MITRE team (Zarrella and Marsh, 2016) employed two recurrent RNN classifiers: the first was trained to predict task-relevant hashtags on a large unlabeled Twitter corpus which was then used to initialize a second RNN to be trained on the SemEval 2016 training set. (Du et al., 2017) proposed a neural network-based model to incorporate target-specific information by means of an attention mechanism. Finally, (Sun et al., 2018) proposed a hierarchical attention network to weigh the importance of various linguistic information, and learn the mutual attention between the document and the linguistic information.

It should be said that neural network approaches have been more successful so far for the SemeEval 2016 Task B (weakly-supervised setting). Apart from the previously mentioned systems (Wei et al., 2016), (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. Another interesting work is that of (Rajadesingan and Liu, 2014) who tried to determine stance at user level. Their assumption was that if many users retweeted a particular pair of tweets in a short time, then it is likely that this pair of tweets had something in common and share the same opinion on the topic.

As far as we know, most approaches to stance detection are developed for English, with the few exceptions that use the Catalan and Spanish data from IberEval 2017 and 2018 (Taulé et al., 2017; Taulé et al., 2018) or the work of (Mohtarami et al., 2019) using the Arabic corpus provided by (Baly et al., 2018).

With respect to the "MultiModal Stance Detection in tweets on Catalan #1Oct Referendum" task at IberEval 2018 (MultiStanceCat), the best results for Spanish were obtained by the uc3m team (Segura-Bedmar, 2018). They presented a system based on bag-of-words with TF-IDF vectorization. They evaluated several of the most commonly used classifiers, obtaining a final 28.02 F1 macro score in the Spanish test data. The best result in Catalan subset was obtained by the CriCa team (Cuquerella and Rodríguez, 2018). Their approach consisted of combining the Spanish and Catalan subsets 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. Since Spanish and Catalan share many words, stemming helped to generalize. Additionally, it is quite common to encounter tweets containing words and expressions in both languages. Their final F1 macro was 30.68.

3. Experimental Setup

The development of the Catalonia Independence Corpus was motivated by the experiments performed on the IberEval TW-10 data. The result of those experiments showed that, due to the highly imbalanced nature of the TW-10 corpus, any comparison of systems across languages were not particularly meaningful. In this section we will summarize the setup for the experiments performed on both datasets, TW-10 and our new Catalonia Independence Corpus.

Apart from the data pre-processing described in Section 3.1., we experimented with four different system architectures: (i) TF-IDF vectorization with a SVM classifier; (ii) a SVM trained with fastText word embeddings (Grave et al., 2018) for the representation of tweets; (iii) the fastText text classification system (Joulin et al., 2017) with fastText word embeddings and, finally (iv) the Flair system (Akbik et al., 2018), which implements a Recurrent Neural Network (RNN) for text classification that can be combined with static and context-based string embeddings. In the following, we describe the pre-processing and each of the architectures tested in both the TW-10 and the Catalonia Independence Corpus (CIC).

3.1. Data Pre-processing

Since each tweet in the TW-1O dataset is given in context, with the previous and the next tweet, we use them to obtain longer and richer texts for classification.

Normalization: We believe that normalization helps to reduce the number of features for TF-IDF feature representation and to maximize the number of words that correspond with the vocabulary of pre-trained word vector mod-

[3] https://github.com/ixa-ehu/catalonia-independence-corpus

els. First, we remove all punctuation and any expression starting with "@", "RT", URLs and numbers. The next step is lowercasing and normalization of spelling: we remove repeated characters with one and replacing common shortened words to their normal form. For example, *holaaaaaaa* is converted to *hola*. However, we leave untouched consonants composed of two characters (*tt, ll, rr*). Finally, diacritics are systematically removed.

Lemmatization: Next, we apply a simplified version of lemmatization consisting of replacing the word form with its lemma via dictionary look-up [4]. If a word is not found we leave it in its original form. Note that this method is not capable of resolving ambiguities. For example, the Spanish preposition *para* ("for") and the verb *para* ("stop") will be mapped to the same lemma, namely, *parar* (the infinitive "to stop" in Spanish). Furthermore, named entities are sometimes wrongly lemmatized. To reduce the error rate, we manually edited the list of lemmas, and deleted the less frequent ambiguous words. In any case, our experiments showed that this type of lemmatization reduces dramatically the number of features helping to improve results for every experimental setting. In addition, it allows to deal with unseen words. For example, if the Spanish word *andando* (walking) does not appear in the training corpus but another form of its lemma does, then both words will be recognized as having the same lemma, namely, the Spanish verb *andar* (to walk).

Tokenization: We perform whitespace tokenization, also removing stopwords (auxiliary verbs, prepositions, articles, pronouns and the most frequent words) and words shorter than three characters.

3.2. SVM+TF-IDF

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.

Information Gain is used for feature selection (Cover and Thomas, 2006). Information Gain provides a method to calculate the mutual information between the features and the classification labels. According to (Aggarwal and Zhai, 2012), mutual information is defined on the basis of the level of co-occurrence between the label and word. In other words, it represents the predictive power of each feature, and measures the number of bits of information obtained for prediction of a class in terms of the presence or absence of a feature in a document. 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 performed for hyper-parameter optimisation. The grid-search results are measured by 5-fold cross-validation on the training set. To reduce the cost of the grid-search process, we select two of the SVM (RBF kernel) parameters, namely, C and gamma.

3.3. SVM+fastText Embeddings

Word embeddings encode words as continuous real-valued representations in a low dimensional space. Word embeddings are trained over large corpora and are able to capture semantic and syntactic similarities based on co-ocurrences. Word embeddings allow to build rich representations of text and have enabled improvements across most NLP tasks.

To the best of our knowledge, the only publicly available pre-trained models for both Catalan and Spanish are those distributed by fastText (Grave et al., 2018). Initial experimentation showed that the Common Crawl[5] models performed better for our particular 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 2M 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 et al., 2016), which is calculated as follows:

$$V(t) = \frac{1}{n}\sum_{i=1}^{n} W_i$$

where *V(t)* is the vector representing a tweet, *n* is the number of words and *W* the vector for each word. In order to facilitate the look-up into the pre-trained word embedding model, the pre-processing described in the previous section is modified, leaving untouched the diacritics and the stopwords.

3.4. 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. A first 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 et al., 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.

We use the fastText system in its default parameters, with the following exceptions: (i) instead of training the word embeddings online, we provide as input the pre-trained fastText word embedding models for Catalan and Spanish described in the previous section and, (ii) we use bag of bigrams and trigrams as additional features with the aim of capturing word order information.

[4] https://github.com/michmech/lemmatization-lists

[5] http://commoncrawl.org/

3.5. Neural Architecture

Flair refers to both a deep learning system and to a specific type of character-based contextual word embeddings. While fastText generates static word embeddings, generating a unique vector-based representation for a given word independently of the context, contextual word embeddings aimed to generate different word representations depending on the context in which the word occurs. Examples of such contextual representations are ELMo (Peters et al., 2018) and Flair (Akbik et al., 2018), which are built upon LSTM-based architectures and trained as language models.

The Flair toolkit (Akbik et al., 2019) allows to train sequence labelling and text classification models based on neural networks. Flair provides a common interface to use and combine different word embeddings, including both Flair and fastText embeddings. For text classification the computed word embeddings are fed into a BiLSTM to produce a document level embedding which is then used in a linear layer to make the class prediction. For best results, we follow their advice of combining in a stack the contextual Flair embeddings for Spanish with the fastText embeddings (Akbik et al., 2018). Every result reported with Flair is the average five training runs initialized at random.

3.6. Evaluation

The models are tuned via cross-validation for the TW-10 dataset. The Catalonia Independence Corpus provides a development set which is used for tuning the models during training. The metric used by the organizers of SemEval 2016 (Mohammad et al., 2016) and IberEval 2018 (Taulé et al., 2018) reported the F1 macro-average score of two classes: FAVOR and AGAINST, although the NONE class is also represented in the test data. We use the provided evaluation script [6] that calculates the final F1 macro score:

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

4. TW-1O Referendum Dataset

The TW-10 for IberEval 2018 dataset was collected using the hashtags #1oct, #1O, #1oct2017 and #1octl6 to obtain the tweets from Twitter (Taulé et al., 2018). These hashtags were widely used in the debate on the right to hold a referendum on Catalan independence on the 1st of October 2017. A total of 87,449 tweets in Catalan and 132,699 tweets in Spanish were collected between the 20th and 30th of September. The final dataset consists of 11,398 tweets: 5,853 written in Catalan (the TW-1O-CA corpus) and 5,545 in Spanish (the TW-1O-ES corpus). The dataset was annotated manually by three experts. Also, each tweet is given together with its previous and next tweets as context. Table 1 shows the average length of tweets after concatenating the tweet with its context.

Table 2 illustrates the imbalanced nature of the Catalan subset, which makes it difficult to built and compare models for Catalan and across languages. Thus, while for Spanish the distribution of classes is quite similar, in Catalan the FAVOR class occurs 35 times more than AGAINST, and 8 times more than NONE.

[6] http://alt.qcri.org/semeval2016/task6/data/uploads/eval_semeval16_task6_v2.zip

TW-10 corpus	Catalan	Spanish
Average tweet length (tokens)	37.69	38.86

Table 1: Average length of tweets plus their context in the TW-1O corpus.

Label	Catalan	Spanish
Against	120	1785
Favor	4085	1680
None	479	972
Total	4684	4437

Table 2: Distribution of classes in the TW-1O trainset.

Tables 3 and 4 reports our results for Catalan and Spanish respectively. It is clear that the Catalan subset makes it very difficult to perform any meaningful experiments given its class distribution. While the best approach for Catalan is SVM+TDF-IDF, it is clear that the results are heavily influenced by the under-represented AGAINST class.

System	$F1_{against}$	$F1_{favor}$	$F1_{macro}$
SVM+TF-IDF	22.86	94.68	**58.77**
SVM+FTEmb	0.00	93.88	46.94
fastText+FTEmb	12.90	94.60	53.78
Flair+FTEmb	14.79	94.40	54.59
Baseline			
(Cuquerella and Rodríguez, 2018)	-	-	30.68

Table 3: Results on the TW-1O Catalan testset.

System	$F1_{against}$	$F1_{favor}$	$F1_{macro}$
SVM+TF-IDF	68.50	64.53	66.52
SVM+FTEmb	63.65	58.85	61.25
fastText+FTEmb	69.58	65.37	**67.48**
Flair+FTEmb	60.23	52.44	56.34
Baseline			
(Segura-Bedmar, 2018)	-	-	28.02

Table 4: Results on the TW-10 Spanish testset.

The results for Spanish are a little bit more interesting. First, we can see that the fastText linear classifier combined with fastText embeddings (fastText+FTEmb) obtains much better results than SVM+FTEmb. As the document representation is the same, that means that the fastText classifier (Joulin et al., 2017) improves over the performance of SVM. Finally, our results provide a significant improvement over previous state-of-the-art in this dataset for both languages.

Nonetheless, motivated by the results obtained for Catalan, we decided to propose a new multilingual corpus for stance detection with a better distribution of classes.

5. Catalonia Independence Corpus 2019

During the process of developing the Catalonia Independence Corpus (CIC) we tried to address the main shortcomings of the TW-10 dataset, as it has been described in Section 4..

We had at our disposal a collection of tweets from 12 days during February and March of 2019 posted in Barcelona

and during September 2018 posted in the town of Terrassa, Catalonia, prepared for commercial research in stance detection and political ideology (left-right) prediction. The collection process was performed by crawling with full access to the Twitter API, obtaining messages of up to the official limit of 240 characters. We decided to use it to create a new dataset for academic research. In order to do so, first we separated them by language[7] and obtained 680000 tweets in Catalan and 2 million tweets in Spanish. We then processed each set separately. We discarded tweets with identical messages and tweets containing less than three words.

Annotation was performed using the same three labels and guidelines as the previously described datasets (SemEval 2016 and TW-10). Thus, FAVOR will state a positive stance towards the independence of Catalonia, AGAINST the opposite, and NONE will express neither a negative nor a positive stance, or simply that it is not possible to reach a conclusion.

5.1. User Categorization

Unlike previous approaches, we do not annotate manually each tweet. Instead, the annotation process is based on classifying users. 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. Secondly, we extracted the most retweeted tweets and categorized their authors manually by checking their Twitter accounts. The assumption was that for a person it is easier to annotate a whole Twitter account rather than the text of a single tweet without context. The decision about their stance was also made taking into account 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 or a Spanish or Republican Catalan flag, etc.), or by the Bio section. We follow this process to assign a FAVOR, AGAINST or NEUTRAL stance to each user.

Furthermore, we extracted the relations between users based on their retweets (Otte and Rousseau, 2002). Assuming that all those who make a retweet share the author's opinion, we categorized these users with the same label as the author of the retweeted message. While this method may introduce some noise, it allowed us to quickly obtain a large amount of annotated data quite cheaply.

In total, 25,510 users were categorized. We do not distinguish between Catalan and Spanish tweets because most of the active users in Catalonia are bilingual and can write in both languages. Table 5 reports the distribution of the categorized users. The final set contains 131022 unique tweets in Catalan and 202645 unique tweets in Spanish.

Label	Count
Favor	22247
Against	3091
Neutral	176

Table 5: Distribution of the categorized users.

[7]https://code.google.com/archive/p/language-detection/

5.1.1. Topic Detection

We annotated the corpus assigning the stance classes to usernames. However, this does mean that we can use every tweet from the users, given that many messages may not be related to the independence of Catalonia. In order to address this issue we performed the following steps:

Hashtags and keywords: 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 added keywords in both languages, 25 in total. We marked 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 the hashtags and keywords filter.

Label	Catalan	Spanish
Against	1476	8267
Favor	23030	11843
Neutral	986	497

Table 6: Distribution of tweets obtained by hashtags and keywords related to "independence".

Topic modelling: We can see in Table 6 that the vast majority of the tweets are labelled as FAVOR. In order 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 et al., 2003) as a kind of basic target detection algorithm to the corpus of categorized users described in Table 5. The objective was to obtain more relevant tweets for our under-populated classes (AGAINST and NEUTRAL in Catalan and NEUTRAL in Spanish). We manually revised the obtained topics and selected only those tweets which were clustered within the "independence" topic.

CIC Corpus	Catalan	Spanish
Average tweet length (tokens)	27.17	30.31

Table 7: Average tweet length in the Catalonia Independence Corpus.

Finally, we selected approximately 10,000 tweets (excluding those shorter than 4 words) per language 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 (in Table 7) is slightly shorter than the average in the TW-1O dataset (see Table 1) given that our corpus does not include the previous and next tweets as context. However, our corpus is larger than previous works (Mohammad et al., 2016; Taulé et al., 2018) and presents a more balanced distribution of classes, as shown by Table 8.

Finally, here we can see an example from the Catalonia Independence Corpus.

> **Tweet:** *Puigdemont visitarà el dia 13 de febrer la Universitat de Groningen dels Països Baixos*

Label	Catalan	Spanish
Against	3988	4105
Favor	3902	4104
Neutral	2158	1868
Total	10048	10077

Table 8: Distribution of classes in the Catalonia Independence Corpus.

i presentará La crisi catalana, una oportunitat per Europa. És un goig veure com ens reben els països democràtics https://t.co/O38mDKwwn3

Stance: FAVOR

Language: Catalan

Translation: *Puigdemont will visit on February 13th the University of Groningen, Netherlands, and present The Catalan Crisis, An Opportunity For Europe. It's a pleasure to see how democratic countries are receiving us https://t.co/O38mDKwwn3*

5.2. Results

This section reports on the results obtained by the systems presented in Section 3..

System	$F1_{against}$	$F1_{favor}$	$F1_{macro}$
SVM+TF-IDF	68.89	72.91	70.90
SVM+FTEmb	59.43	64.46	61.95
fastText+FTEmb	70.73	72.21	**71.47**
Flair+FTEmb	59.08	58.08	58.96

Table 9: Results on the Catalan testset of the Catalonia Independence Corpus (CIC-CA).

System	$F1_{against}$	$F1_{favor}$	$F1_{macro}$
SVM+TF-IDF	70.67	71.50	71.09
SVM+FTEmb	64.24	62.51	63.38
fastText+FTEmb	73.20	71.13	**72.43**
Flair+FTEmb	61.76	54.84	58.29

Table 10: Results on the Spanish testset of the Catalonia Independence Corpus (CIC-ES).

It is clear that the results for both Catalan (Table 9) and Spanish (Table 10) are higher across languages and systems than those obtained on the TW-10 dataset. This means that the semi-automatic method for the annotation of tweets presented in this paper is quite effective and provides good quality annotated data. Furthermore, there is a consistency in the behaviour of the systems across both languages, which allows to compare their performance in multilingual settings. These results are also consistent with the TW-1O Spanish subset, where the fastText+FTEmb system also obtained the best scores. Finally, the deep learning approach from Flair seems to lag behind linear classifiers. While a bit surprising, this is also coherent with the results obtained in English with the SemEval 2016 dataset, as explained in the Related Work section. Our hypothesis is that the short length of the tweets make it more difficult to generate good contextual-based word representations. However, further experimentation is required to clarify this issue.

6. Error Analysis

In order to perform an analysis of the quality of the annotations obtained by our semi-automatic method (as described in Section 5.), we took a sample of 100 tweets per language from the training sets. This sample was manually revised by three human annotators. We found out that the error rate in the Spanish sample was around 5%, whereas for the Catalan sample was slightly higher, around 15%. It should be noted that those error rates are approximate because the three human 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 underspecified. Second, many tweets use figurative language such as sarcasm and irony. Finally other tweets referred to the topic in a indirect manner. Below it can be found a couple of examples of contentious tweets in which it is not really clear which of the annotations are the correct one, namely, the one provided by our method (semi-automatic user-based) or the manual one. In Tweet 1, we can see a seemingly neutral message, but the author uses anti-independence slogan. In Tweet 2, although it seems to be neutral, the interpretation depends on the context of the message, where the annotator should know the details of the case.

Tweet 1: *Arrimadas irá a Waterloo este domingo para recordar a Puigdemont que la república no existe https://t.co/6luAEAj2UD*

Our method: NEUTRAL

Manual annotation: AGAINST/NEUTRAL

Language: Spanish

Translation: *Arrimadas will go to Waterloo this Sunday to remind Puigdemont that the republic does not exist https://t.co/6luAEAj2UD*

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

Manual annotation: FAVOR/NEUTRAL

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.*

Manual annotation of stance in tweets is a difficult task for humans, partly because it depends greatly on the annotator's background knowledge and intuition. Furthermore, annotating tweets one by one, as opposed to user-based annotation, albeit automatic, suffers from a lack of context.

7. Concluding Remarks

In this paper we provide a new dataset for stance detection in Catalan and Spanish. The objective is two-fold: (i) to promote research on stance detection in other languages different to English and, (ii) to facilitate experimentation in multilingual and cross-lingual settings. We show that the methodology used to build the Catalonia Independence Corpus generates good quality annotated data without having to manually annotate tweet by tweet. Most importantly, it also helps to alleviate the imbalance in the classes distribution. Our experimental results confirm these considerations as the tested systems exhibit consistent behaviour across languages. We believe that our methodology can help to obtain larger annotated datasets from limited resources while making the annotation process cheaper and faster.

Additionally, we establish new state-of-the-art results on the TW-1O dataset for both Catalan and Spanish. Our hypothesis to explain the large difference with previous work is the more exhaustive pre-processing performed, apart from the use of the fastText word embeddings to obtain the tweets representation. The results also show the superior performance of the fastText linear classifier over SVM or RNN approaches on both datasets. These results are somewhat similar to those obtain for English with the SemEval 2016 data, where linear classifiers still are competitive or outperform newer deep learning approaches.

We publicly distribute the datasets and code to facilitate further multilingual and cross-lingual research on stance detection[8].

8. Acknowledgements

This work has been funded by the Spanish Ministry of Science, Innovation and Universities under the project DeepReading (RTI2018-096846-B-C21) (MCIU/AEI/FEDER, UE) and by the BBVA Big Data 2018 "BigKnowledge for Text Mining (BigKnowledge)" project. The second author is funded by the Ramon y Cajal Fellowship RYC-2017-23647. We also acknowledge the support of the Nvidia Corporation with the donation of a Titan V GPU used for this research.

9. Bibliographical References

Aggarwal, C. C. and Zhai, C. (2012). *Mining Text Data*. Kluwer Academic Publishers, Boston/Dordrecht/London.

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

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., and Vollgraf, R. (2019). FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota, June. Association for Computational Linguistics.

[8] https://github.com/ixa-ehu/catalonia-independence-corpus

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

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

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March.

Bøhler, H., Asla, P., Marsi, E., and 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)*, pages 445–450, San Diego, California, June. Association for Computational Linguistics.

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

Cuquerella, C. A. and Rodríguez, C. C. (2018). Crica team: Multimodal stance detection in tweets on catalan 1oct referendum (multistancecat). In *IberEval@SEPLN*.

Derczynski, L., Bontcheva, K., Liakata, M., Procter, R., Wong Sak Hoi, G., and 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)*, pages 69–76, Vancouver, Canada, August. Association for Computational Linguistics.

Du, J., Xu, R., He, Y., and Gui, L. (2017). Stance classification with target-specific neural attention networks. In *Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI 2017*.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and 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.*

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., and 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, Short Papers*, pages 427–431, Valencia, Spain.

Kenter, T., Borisov, A., and 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.

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Mikolov, T., Chen, K., Corrado, G., and 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*.

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

Mohammad, S. M., Sobhani, P., and Kiritchenko, S. (2017). Stance and sentiment in tweets. *ACM Trans. Internet Technol.*, 17(3):26:1–26:23, June.

Mohtarami, M., Glass, J., and 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)*, pages 4441–4451, Hong Kong, China, November. Association for Computational Linguistics.

Otte, E. and Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. *Journal of Information Science*, 28(6):441–453.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *NAACL*.

Rajadesingan, A. and Liu, H. (2014). Identifying users with opposing opinions in twitter debates. In *Social Computing, Behavioral-Cultural Modeling and Prediction - 7th International Conference, SBP 2014, Washington, DC, USA, April 1-4, 2014. Proceedings*, pages 153–160.

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), Sevilla, Spain, September 18th, 2018.*, volume 2150, pages 180–186.

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

Taulé, M., Martí, M. A., Rangel, F., Rosso, P., Bosco, C., , and 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)*, pages 158–177, Murcia, Spain.

Taulé, M., Rangel, F., Martí, M. A., and 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*, pages 149–166, Sevilla, Spain.

Wei, W., Zhang, X., Liu, X., Chen, W., and 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)*, pages 384–388, San Diego, California, June. Association for Computational Linguistics.

Zarrella, G. and 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)*, pages 458–463, San Diego, California, June. Association for Computational Linguistics.