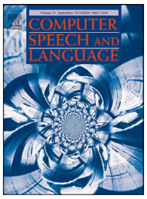
Computer Speech & Language 63 (2020) 101075

Contents lists available at ScienceDirect

Computer Speech & Language

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



Multilingual stance detection in social media political debates

Mirko Lai[a,1,*], Alessandra Teresa Cignarella[a,b,1], Delia Irazú Hernández Farías[c], Cristina Bosco[a], Viviana Patti[a], Paolo Rosso[b]

[a] *Dipartimento di Informatica, Università degli Studi di Torino, Italy*
[b] *PRHLT Research Center, Universitat Politècnica de València, Spain*
[c] *División de Ciencias e Ingenierías, Universidad de Guanajuato Campus León, Mexico*

ARTICLE INFO

Article History:
Received 18 January 2018
Revised 30 December 2019
Accepted 4 February 2020
Available online 12 February 2020

Keywords:
Stance detection
Multilingual
Contextual features
Political debates
Twitter

ABSTRACT

Stance Detection is the task of automatically determining whether the author of a text is in favor, against, or neutral towards a given target. In this paper we investigate the portability of tools performing this task across different languages, by analyzing the results achieved by a Stance Detection system (i.e. MultiTACOS) trained and tested in a multilingual setting.
First of all, a set of resources on topics related to politics for English, French, Italian, Spanish and Catalan is provided which includes: novel corpora collected for the purpose of this study, and benchmark corpora exploited in Stance Detection tasks and evaluation exercises known in literature. We focus in particular on the novel corpora by describing their development and by comparing them with the benchmarks. Second, MultiTACOS is applied with different sets of features especially designed for Stance Detection, with a specific focus to exploring and combining both features based on the textual content of the tweet (e.g., style and affective load) and features based on contextual information that do not emerge directly from the text. Finally, for better highlighting the contribution of the features that most positively affect system performance in the multilingual setting, a features analysis is provided, together with a qualitative analysis of the misclassified tweets for each of the observed languages, devoted to reflect on the open challenges.

1. Introduction

Detecting stance of people towards specific targets is a field of Natural Language Processing (NLP) research that is currently collecting an increasing interest. It has been defined in literature as Stance Detection (SD), that is the task of automatically determining whether the text's author is in favor, against, or neutral towards a statement or targeted event, person, organization, government policy, movement, etc. (Mohammad et al., 2017).

Like Sentiment Analysis, also SD has been applied in several domains to discover the reputation of an enterprise, what is the general public thinks of a political reform, if costumers of a fashion brand are happy about the customer service etc. Nevertheless, whereas the aim of Sentiment Analysis is at categorizing texts according to a notion of polarity, i.e. as positive, negative or

[*] Corresponding author.
E-mail addresses: mirko.lai@unito.it, mirla@doctor.upv.es (M. Lai).
[1] The first two authors equally contributed to this work.

https://doi.org/10.1016/j.csl.2020.101075
0885-2308/© 2020 Elsevier Ltd. All rights reserved.

neutral, that of SD consists in classifying texts according to the attitude they express towards a given target of interest. This difference between sentiment polarity and stance can be observed for instance in the following tweet[i]

> **Target of interest:** Climate change is a real concern
>
> @RegimeChangeBC @ndnstyl It's sad to be the last generation that could change but does nothing. #Auspol
> **Polarity:** NEGATIVE
> **Stance:** FAVOR

where the opinion expressed by the user includes a negative polarity contrasting with the stance expressed in favor of the target of interest. This support the idea that SD deserves to be treated as a singular classification task and needs to be distinguished from classical Sentiment Analysis tasks focused on polarity.

Several shared tasks already took place for promoting Sentiment Analysis (Bethard et al., 2016; Mohammad et al., 2016b; Nakov et al., 2013; Rosenthal et al., 2015), but only recently SD has been acknowledged as an independent task having its own characteristics, peculiarities and benchmark datasets. The first shared task on SD was indeed held for English at SemEval in 2016, i.e. Task 6 “Detecting Stance in Tweets” (Mohammad et al., 2016b). It consisted in detecting the orientation in favor or against six different targets of interest: “Hillary Clinton”, “Feminist Movement”, “Legalization of Abortion”, “Atheism”, “Donald Trump”, and “Climate Change is a Real Concern”. A more recent evaluation for SD systems was instead proposed at IberEval 2017 for both Catalan and Spanish (Taulé et al., 2017) where the target was only one, i.e. “Independence of Catalonia”.

Not surprisingly in all cases the SD task was based on data extracted from social media, namely Twitter, and about politics and public life topics. On the one hand, social media are contexts where people spontaneously express opinions, desires, complaints, beliefs and outbursts. On the other hand, politics and public life are among the topics mainly discussed by users in social media. In these choices the possible relevance of SD techniques for policy makers and public administrators is also mirrored, e.g. for better meeting population's needs and preventing feelings of dissatisfaction and extreme reactions of hostility and anger. A further motivation for collecting texts from social media, in particular Twitter, is that this is a great source of freely available data.

Given the increasing interest for the task of automatically detecting stance in political polarized debates on Twitter and the recent development of SD benchmark corpora in several languages, our main focus in this paper is investigating the portability of tools performing SD in a multilingual setting. The languages that we work with in this research are five, namely: English, Spanish, French, Catalan and Italian but a manifold of corpora in other languages is available to the scientific community: Arabic (Magdy et al., 2016), Chinese (Yuan et al., 2019), Russian (Vychegzhanin and Kotelnikov, 2019), and Turkish (Küçük and Can, 2019).

Our starting point is a model, called MultiTACOS, for addressing SD as a classification problem, which exploits a multifaceted set of features especially designed for identifying stance in Twitter. In particular, we are interested in exploring and combining features based on the *textual content of the tweet*, such as structural, stylistic, and affective features, but also features based on *contextual information* that do not emerge directly from the text, such as e.g. knowledge about the domain of the political debate or about the user community. Some of these typologies of features have been successfully exploited and evaluated in SD tasks in monolingual settings, but their combination and contribution, and their potential in a multilingual setting have never been studied. We are in particular interested in finding answers for research questions about:

- RQ1. What is the contribution that different typologies of features (and specific features within them) can provide for accomplishing a SD task in different languages and domains?
- RQ2. How much are these features portable across different languages and domains?

Especially focusing on the impact of combining content and contextual information with information based on social network community, our main contribution for exploring SD is twofold: on the one hand to release novel resources for SD for the French and Italian languages; on the other hand, to provide experimental evidences supporting hypotheses about the portability of the MultiTACOS SD model across five languages, namely Catalan, English, French, Italian and Spanish.

Considering that the resources annotated for stance currently existing, i.e. those cited above as benchmarks for English and Spanish-Catalan respectively, are not enough for working in a multilingual perspective, for the purpose of this study we provide indeed two newly annotated corpora, respectively for French and Italian, to be exploited in the experiments together with the benchmarks cited above. Like the latter datasets, the novel ones include a comparable amount of texts retrieved from Twitter and about similar political topics and debates, where targets are political opinions, like in referendums, or politicians, like in electoral campaigns. For instance, referring to the portion of the English dataset focused on the targets related to political elections in the USA (“Hillary Clinton” and “Donald Trump”), we collected for French tweets about “Emmanuel Macron” and “Marine Le Pen”, i.e. the last two candidates of the run-off of the political elections in France in 2017. For what concerns instead Italian, we retrieved tweets about the constitutional referendum held in 2016, which mirrors the features of the debate about the “Independence of Catalonia” which is the target of the Spanish-Catalan corpus. These novel corpora not only extend the scenario of resources available for the community working on SD, but also pave the way for multilingual studies and experiments.

In order to provide answers to our research questions, we propose a battery of experiments where we apply the MultiTACOS machine learning model for SD in a multilingual setting. Our experiments confirm that better results across languages and

[i] Example taken from Mohammad et al. (2017).

political domains can be achieved by joining features related to tweet textual contents with those related to context, especially exploiting features based on information about the social network structure of user communities.

The paper is organized as follows. In Section 2 we discuss the related work. Section 3 is devoted to the description of the datasets, showing the features of the benchmarks known in literature and those of the novel datasets, their similarity and difference for what concerns both composition and annotation schema and procedure. In Section 4 the experiments performed applying the SD system MultiTACOS are described. Finally in Section 5, by following a multilingual perspective, we discuss our results analysing also the errors detected (Section 5.4). Section 6 concludes the paper.

2. Related work

Social media texts provide an interesting source of information for investigating people's opinion on controversial topics. They are considered as especially useful for monitoring political sentiment with possible different focuses and levels of granularity of the sentiment analysis (Bosco and Patti, 2017): detecting users stance, detecting the polarity of messages expressing opinions about candidates in political elections, forecasting the outcome of elections or referendums (Celli et al., 2016) and so on.

Among the areas that can be of some interest with respect to SD and its application in a multilingual perspective, in this section, we mainly analyze the contribution of NLP and computational social science.

2.1. Stance detection in social media contents

Within the NLP, SD is commonly interpreted as the task of detecting from textual contents the users stance towards a given target of interest. To the best of our knowledge, (Somasundaran and Wiebe, 2009) were the first one in this community to focus on detecting the stance towards a target rather than the polarity of a sentence. They presented (in a unsupervised framework) a stance recognition method for debate-side classification (i.e. recognizing which stance a person is taking) from web blogs. The method is based on the association among preferences with opinions towards different aspects. The first specific shared task on SD in Twitter was organized in 2016 in the framework of the SemEval evaluation campaign (Mohammad et al., 2016b) as part of the *Sentiment Analysis track*: "Task 6: Detecting Stance in Tweets" (henceforth SemEval-2016 Task 6). The organizers provided an English dataset annotated with stance considering six commonly known targets in the United States (Mohammad et al., 2016a). The participating systems were asked to determine the stance contained in each given tweet towards a target entity among them. The task 6 was moreover organized around two subtasks:

- *Task A. Supervised Framework*. Annotated datasets are provided for both training and testing about five different targets: "Atheism", "Climate Change is a Real Concern", "Feminism Movement", "Hillary Clinton", and "Legalization of Abortion".

- *Task B. Weakly Supervised Framework*. Unlike in Task A, for this task only data for testing were released, and they were only about one target, "Donald Trump".

A total of nineteen systems participated in Task A while only nine in the Task B. The evaluated systems in the shared task used widely applied features in text classification such as n-grams and word vectors together with information extracted from sentiment lexicons. Furthermore, word embeddings and deep neural networks were also exploited. It is worth to be mentioned that as stated in Mohammad et al. (2016b), for what concerns Task A, considering the scores over all targets, none of the participating systems surpassed a baseline SVM classifier that uses word and character n-grams as features (F-score: 68.98). The highest score in Task A achieved an F-measure of 67.82 and of 56.28 for Task B, confirming the additional difficulty related to the weakly supervised setting. Further information about the task and the participating systems can be found in Bethard et al. (2016).

Mohammad et al. (2017) investigated the importance of exploiting the sentiment expressed in a given text in order to improve SD. The dataset of the SemEval-2016 Task 6 was annotated with the overall sentiment expressed in each instance without considering the target. The features exploited in their system include n-grams, char-grams, sentiment features coming from different lexica such as EmoLex (Mohammad and Turney, 2013), Hu and Liu lexicon (Hu and Liu, 2004), and MPQA Subjectivity Lexicon (Wilson et al., 2005). Besides, they also considered the presence/absence of the target of interest in the tweet, the frequency of part-of-speech tags, emoticons, hashtags, uppercase characters, elongated words, and punctuation marks. The combination of these features together with a support vector machine classifier allowed they to outperform the scores achieved by all the participating systems in SemEval-2016 Task 6.

Focusing on the Twitter dataset released for SemEval2016-Task 6, Lai et al. (2017b) proposed an approach for detecting stance that relies on the knowledge of the domain and of the context surrounding a target of interest. The approach was evaluated selecting two targets from the original dataset, i.e. Hillary Clinton and Donald Trump. Three groups of features were considered: *Structural* (hashtags, mentions, punctuation marks, etc.), *Sentiment* (a set of four lexica to cover different facets of affect ranging from prior polarity of words such as AFINN (Nielsen, 2011) and Hu and Liu Lexicon, to fine-grained emotional information such as LIWC (Pennebaker et al., 2001) and the Dictionary of Affect in Language (Whissell, 2009)), and *Context-based* (with the attempt to capture the information surrounding a given target, the authors used two concepts: "friends" and "enemies" as the entities related to the target, defining a set of relationships between the target and the entities around it). Furthermore, Lai et al. (2017b) also exploited the additional annotation carried out in Mohammad et al. (2017) on the dataset of the shared task. The proposed approach outperforms the state-of-the-art results, showing that information about enemies and friends of politicians help in detecting stance towards them.

It is also worth mentioning that several systems which participated to the SemEval-2016 Task 6 competition involved the use of deep learning techniques, such as Wei et al. (2016) who used a Convolutional Neural Network (CNN) combined with a voting scheme based on the concept of “divide and conquer”, and Zarrella and Marsh (2016) who exploited a Recurrent Neural Network (RNN) with four layers containing 128 Long Short Term Memory (LSTM) units. In the last years, after the end of the contest, the dataset released for the SemEval-2016 Task 6 has been considered as a benchmark and therefore exploited to carry on research regarding SD in English tweets by several research groups (Augenstein et al., 2016; Dey et al., 2018; Wei et al., 2018; Zhou et al., 2019; Del Tredici et al., 2019). Among them, let us focus on the ones where a score on the specific stance targets addressed in this paper (Hillary Clinton or Donald Trump) is reported: Augenstein et al. (2016) who proposed a neural approach based on bidirectional conditional encoding, Dey et al. (2018) who implemented a two-phase LSTM using attention, Wei et al. (2018) who explored the performances of a biderectional Long Short-Term Memory neural network (biLSTM), and Zhou et al. (2019) who used a condensed CNN with attention over self-attention.

In the StanceCat shared task[ii] Taulé et al. (2017) held within the evaluation campaign IberEval 2017, the Independence of Catalonia was chosen as the target of stance in tweets written in Spanish or Catalan. Well-known approaches for classification, such as SVM (Support Vector Machine), and novel techniques, such as deep learning, were applied by the ten different teams participating in the shared task for detecting stance (in favor, against or neutral) towards the target of interest in the annotated dataset provided by the organizers for both languages. For Catalan and Spanish both ITACOS Lai et al. (2017a) resulted the best performing system, which consists in a supervised approach based on three groups of features: *Stylistic* (bag of: n-grams, char-grams, part-of-speech labels, and lemmas), *Structural* (Hashtags, mentions, uppercase characters, punctuation marks, and the length of the tweet), and *Context* (the language of each tweet and information coming from the URL in each tweet). These results validate the relevance of contextual information in SD.

2.2. Multilingual sentiment analysis

Given the close relationship between the two tasks highlighted in the introduction, many works in the field of multilingual Sentiment Analysis result to be useful source of inspiration for our work as they tackle the pointy issue of multilinguality. In Denecke (2008) the authors exploit SentiWordNet to explore sentiment in a multilingual perspective (training on fifteen languages and testing on German), which proves to be a useful resource. Other researchers exploit supervised learning for a few languages (French, German and Spanish) and machine translation techniques (to obtain data in English) (Balahur and Turchi, 2014) while others simply resort to classical machine learning techniques (Boiy and Moens, 2009) applied on English, Dutch and French. In Tromp and Pechenizkiy (2011) the authors propose a pipeline for English and Dutch (SentiCorr) based on four steps including: language identification for short texts, part-of-speech tagging, subjectivity detection and polarity detection, but more interestingly, they tested it on three different datasets extracted from three different social media and personal correspondence (i.e. e-mails), obtaining good performances.

2.3. Social network analysis

Several scholars also investigated SD in a social network perspective, leveraging methods from the computational social science research field.

Lai et al. (2017c) analyzed the role of social relations together with the users’ stance towards the BREXIT referendum. Furthermore, taking into account that people may change their stance after some particular event, happening when the debate is still active, they also explore stance from a diachronic perspective. The authors collected a set of English tweets containing the hashtag #brexit, and provided an annotated corpus where diachronic triplets of tweets posted by 600 users active in the debate have been annotated for stance. The results show two main results that may be of particular interest for addressing SD: that users sharing the same stance towards a particular issue tend to belong to the same social network community, and users’ stance diachronically evolves.

A similar experiment has been performed by Lai et al. (2018) analyzing the political debate on Twitter about the Italian Constitutional referendum held in 2016. The authors analyzed both the diachronical evolution of the stance and the online social relations of the users involved in the debate. Interestingly, the typology of the relations used for creating the network (retweets, replies, and quotes) highly affect the performance of the SD system.

The effects of online social network interactions on future attitudes are examined in Magdy et al. (2016), focusing on how a content generated by a user and network dynamics can be used to predict future attitudes and stances in the aftermath of a major event. The authors explored the effectiveness of three types of features for the prediction, namely content features (i.e., the body of the tweets from a user), profile features (i.e., user-declared information such as name, location, and description), and network features (i.e., user interactions with the Twitter community, through mentions, retweets, and replies).

Concerning SD in tweets, in Rajadesingan and Liu (2014) implement a semi-supervised framework coupled with a supervised classifier to identify users with differing opinions. The authors exploit a retweet-based label propagation, based on the observation that if many users retweet a particular pair of tweets within a reasonably short period of time, then it is highly likely that the two tweets are similar in some aspect. In their work, they label tweets either as “for” or as “against” on the basis of the similarity with the values of the labels surrounding each tweet.

[ii] Detecting the gender of the author of a given tweet was also a sub-task to be addressed in the shared task.

Similarly, in the work of Raghavan et al. (2007), a label propagation algorithm is used for community detection. Their approach is particularly simple and efficient, in fact, in their iterative algorithm each node adopts the label that most of its neighbors currently have and it seems to work really well in unsupervised contexts.

An interesting work regarding the concept of "hompohily", i.e. the tendency of individuals to associate and bond with similar others, is that of DellaPosta et al. (2015). Their work, although describes opinions and aggregating circles from a sociological perspective is very much connected with the world of SD. In fact, the authors propose computational experiments on a case study taking into account the political and the ideological alignments. Their aim is to analyze how homophily and influence lead to the stereotyped perception of the world.

3. Datasets

The first step for building the scenario for testing a SD system in a multilingual perspective consists in collecting the datasets necessary for this task. The existing benchmarks for SD, respectively released for English for SemEval 2016 (Bethard et al., 2016) and for Spanish and Catalan for IberEval 2017 (Taulé et al., 2017), include texts about political topics. In order to extend the variety of languages for our study, we enriched this former collection with two sets of tweets of similar topics in other two languages, Italian and French, thus generating a data repository where five different languages are represented.

In particular, to collect the tweets in French and Italian for the creation of the two brand-new datasets, in order to improve the homogeneity of the collections and enhancing the possibility of comparison among the five datasets involved in this study, we strictly respected the same criteria applied for the retrieval of the two benchmark datasets (i.e. the English and the Spanish-Catalan one).

For instance, we discarded all the retweets (RTs) like in the retrieval of the English dataset collected for SemEval 2016 Task 6 (Mohammad et al., 2016b).

This strategy, which can be in principle seen as causing the loss of information to be usefully exploited for detecting the social network underlying the data, is in line with a NLP viewpoint. In this field of research retweets are indeed often considered as a redundant piece of text that could bias automatic systems (Mohammad et al., 2017). As far as the topics are concerned, collecting data from an election campaign in French and a referendum in Italian, we can draw also for the novel resources the same distinction in sub-topics that we have seen in the benchmarks. The motivation of such a subdivision inside the more general topic of politics lies in the belief that both language and attitude of users are different when it comes to the election or when they have to deal with a yes/no choice as the one presented in referendums. Therefore, we believe that an automatic system could take advantage of a fine-grained selection of features, depending on the sub-topic that is involved. A similar intuition can be found in the work of West (1991).

As a matter of fact, it may be observed that the tweets collected in Spanish - Catalan not properly refer to a referendum. They refer to the "Independence of Catalonia", a subject that has been thoroughly discussed within the 2015 Catalan regional election that was held on Sunday, 27 September 2015, electing the 11th Parliament of the Autonomous Community of Catalonia. An unofficial poll on the same topic, ruled illegally by the Constitutional Court, has been previously held (in November 2014), achieving a large majority of votes rooting for independence. According to the view of the secessionists, Catalan regional elections held in September 2015 have been considered a *de facto* referendum on the matter of independence. In our work, following the considerations of Bosco et al. (2016) and the groundwork suggested in the shared task StanceCat at IberEval 2017 (Taulé et al., 2017), we also consider the Spanish - Catalan tweets as a kind of referendum towards the target "Independence of Catalonia"[iii]

Finally, Table 1 resumes our collection of corpora showing for each of them the topic (election vs referendum), language (English, Spanish - Catalan, French, and Italian) and size. We also introduce a label for each corpus, which will be used in the rest of the paper for referring to each specific dataset. In Section 3.1 we describe in detail the four Twitter datasets, mentioning source, techniques, pre-processing, filtering and dimensions. In Section 3.2, on the other hand, we focus our attention on the annotation procedure, guidelines and Inter-Annotator Agreement (IAA).

3.1. Data Collection

In this section we first describe the collection of the SemEval 2016 and IberEval 2017 benchmark datasets and subsequently the collection of the two novel datasets created for this work.

Benchmark Datasets

English Dataset (E-USA). The English dataset is extracted from the prior dataset released by the organizers of the first shared task for SD at SemEval 2016 (Mohammad et al., 2016b). At SemEval 2016 (Mohammad et al., 2017), the organizers gathered tweets using query hashtags concerning the topic of the 2016 United States presidential primaries for the Democratic and Republican parties main candidates, i.e. Hillary Clinton and Donald Trump, such as: *#Hillary4President, #Trump2016, #WhyIAmNotVotingForHillary, #Hillary2016, #WakeUpAmerica*. They discarded retweets (RTs) and tweets with URLs and kept only those where the query hashtags appeared at the end of the tweet. Finally, they removed the query hashtags from each post. From this collection they randomly sampled 2000 tweets regarding the two candidates that were left after the described pre-processing filtering. See Mohammad et al. (2016a) for more details about how the dataset was constructed.

[iii] Therefore, the abbreviated label for this dataset was highlighted with an asterisk symbol after the letter "R".

Table 1
Overview of datasets.

Dataset	*Type*	*Label*	*Topic*	*Language*
Benchmark	*Election* *Referendum*	*E-USA* *R*-CAT*	*Hillary Clinton Donald Trump* *Independence of Catalonia*	English Spanish Catalan
New	*Election* *Referendum*	*E-FRA* *R-ITA*	*Emmanuel Macron Marine Le Pen* *Constitutional Reform*	French Italian

Spanish-Catalan Dataset (R*-CAT). StanceCat dataset was released during the *Stance and Gender Classification Task* that took place as part of IberEval 2017 (Taulé et al., 2017). Organizers of the shared task used Twitter API in order to gather all the tweets, excluding RTs, in Spanish or Catalan containing the hashtags *#Independencia* (#Independence) or *#27S*[iv], within the months of September and December 2015. In total, 10,800 tweets were gathered and annotated (5,400 written in Catalan and 5,400 written in Spanish). See Taulé et al. (2017) for more details about how the dataset was constructed.

New Datasets

French Dataset (E-FRA). We created the French dataset for the present research. It consists of tweets concerning the French presidential elections held in 2017 between the two opponents, i.e. Emmanuel Macron and Marine Le Pen. We used the Twitter Stream API in order to gather about 2,8M tweets (no RTs) over the two weeks preceding and following the second turn of the French presidential elections (held on May 6/7, 2017). The following keywords were used: *macron, #presidentielles2017, lepen*, and *le pen*. Finally, we randomly selected a sample of 2,000 tweets regarding the figures of Emmanuel Macron and Marine Le Pen.

Italian Dataset (R-ITA). This corpus includes tweets about the topic of the Referendum held in Italy on December 4, 2016, a reform of the Italian Constitution. On Sunday 4 December 2016, Italians were asked whether they approve a constitutional law that amends the Constitution to reform the composition and powers of the Parliament, the division of powers between the State, the regions, and other administrative entities. 59.11% of voters rejected the reform causing the resignation of Matteo Renzi, the Prime Minister that assumed full responsibility for the referendum defeat. We used the Twitter API to gather Italian tweets (no RTs) about the debate on this topic, and therefore containing the hashtags *#referendumcostituzionale*, generated by users during the month before the referendum (November 2016), obtaining 6M tweets. Afterwards, we randomly sampled 1,000 tweets.

The new resources for Italian and French complete the test bed for our experiments about SD. The four datasets are indeed featured by comparable topics and size. Nevertheless, the size of the R*-CAT dataset is much bigger than the other three ones. An enormous effort has been spent by the organizers of the shared task for building it: it comprise 10,800 annotated tweets in both languages.

3.2. Data annotation

As far as the annotation schema, all four datasets have undergone an annotation which follows the same guidelines initially proposed for the English dataset in Mohammad et al. (2016a). Nevertheless, the intrinsic nature of each language and dataset has determined the application of some minor change in the annotation phase, as we will comment in the following paragraphs.

In particular, for what concerns the labels of the schema and the criteria to be followed by the annotators for selecting among them in the annotation of each tweet, they are summarized in the following box as reported in Mohammad et al. (2016a).

From reading the tweet, which of the options below is most likely to be true about the tweeter’s stance or outlook towards the target?

1. **FAVOR**: We can infer from the tweet that the tweeter supports the target.
2. **AGAINST**: We can infer from the tweet that the tweeter is against the target.
3. **NONE**: We can infer from the tweet that the tweeter has a neutral stance towards the target or there is no clue in the tweet to reveal the stance of the tweeter towards the target (support/against/neutral) (this label was previously divided in NEUTRAL and NO STANCE) (Mohammad et al., 2016a).

In the rest of this section we first focus on the two benchmark datasets, and then on the two novel ones, showing the peculiarities of the annotation procedure, guidelines and IAA.

Benchmark Datasets

English dataset (E-USA). The 2,000 tweets of these datasets, 1,000 for each of the two targets of stance, have been uploaded on the Crowdflower platform[v] to be annotated by annotators previously evaluated against a small gold standard set of annotated posts and achieving an accuracy higher than 70%.

[iv] The 2015 Catalan regional election that was held on Sunday, 27 September 2015.

[v] Now Figure Height: https://www.figure-eight.com

Table 2
Label distribution in the E-USA dataset.

Hillary Clinton				Donald Trump			
FAVOR	*AGAINST*	*NONE*	*TOTAL*	*FAVOR*	*AGAINST*	*NONE*	*TOTAL*
163	565	256	984	299	148	260	707

The four labels of the originally proposed annotation schema (i.e. FAVOR, AGAINST, NEUTRAL, and NO STANCE), after the manual annotation took place, have been reduced to three, encompassing NEUTRAL and NO STANCE labels into one unique category, named NONE (neither favor nor against), since less than 0.1% of the data received the NEUTRAL label. After the annotation of each post made by at least eight independent annotators, a corpus including 984 tweets for “Hillary Clinton” and 707 for “Donald Trump” has been released including only tweets having an IAA greater than 60%.

The detailed scores of the IAA for the two targets we are interested in (“Donald Trump” and “Hillary Clinton”) were not published by the authors. As for what concerns the measures of IAA in the English dataset, as in SemEval-2016 Task 6 comprehended other four targets[vi] in addition to “Donald Trump” and “Hillary Clinton”, the agreement was calculated over all topics and targets (score of 73.11%). In fact, the IAA in Mohammad et al. (2017) was calculated as the average percentage of times two annotators agreed with each other, with a metric that is not compatible with the most common Fleiss’ Kappa coefficient used at IberEval 2017 (Taulé et al., 2017).

In Table 2, which shows the label distribution for each target, in particular, we can see that for the target “Hillary Clinton” a significant unbalanced distribution skewing towards the label AGAINST is present. Whereas the label distribution for “Donald Trump” skewing towards the label FAVOR. The following two tweets (examples 1 and 2) are extracted from the E-USA dataset.

1. @realDonaldTrump A man who isn’t afraid to speak the truth is a man who I’ll vote for! Take it home, Mr. Trump! #FuturePresident #SemST
   English: “Donald Trump” → FAVOR

2. Use your brain, keep Hillary out of the White House.Clinton2016
   English: “Hillary Clinton” → AGAINST

In Example 1, the target, i.e. Donald Trump, is mentioned through the @ (at) symbol, as it is in use on Twitter to mention other users. Also, in this tweet, the author makes a clear statement about his favorite candidate. In Example 2, the target is instead Hillary Clinton, and the user manifests her/his stance against the democratic candidate by using a strong rhetoric.

Spanish - Catalan Dataset (R*-CAT). For building the dataset R*-CAT, released for the IberEval shared task on SD, 5,400 tweets were selected for Catalan and the same amount for Spanish. For this resource the annotation schema is the same based on three labels and proposed in Section 3.2 (Taulé et al., 2017) for the English corpus. The annotation process involved three trained annotators. As first step they tagged stance in 500 tweets in each of the two languages of the corpus and then discussed the annotation in order to achieve agreement and shared guidelines. After that, the three annotators went on to independently annotate the whole corpus. In the released gold resource, one of the labels among AGAINST, FAVOR or NONE was assigned to a tweet only when proposed by at least two annotators. By contrast, for the tweets on which the three annotators disagreed, the annotation has been discussed until a consensus is achieved at least from two annotators over three. It is important to underline that within this procedure no tweets had been discarded.

The IAA on 10,800 tweets was calculated through Fleiss’ Kappa coefficient reaching a value of $\kappa = 0.60$ in both sub-corpora. The results obtained show a moderate agreement, demonstrating the complexity of the task.

Table 3 shows the label distribution over the two languages for the “Independence of Catalonia” target. As we can appreciate from the numbers shown, a prevalence of the tag NONE features the Spanish posts. On the contrary, tweets written in Catalan have an evident preference for the tag FAVOR. It is also worth mentioning the scarce presence of Catalan tweets AGAINST the target “Independence of Catalonia” (only 163 tweets, i.e. 3% of the Catalan sub-corpus). This does not necessarily mean that the majority of Catalan people are in FAVOR of the independence, although Twitter users from Catalonia are. Below we show two tweets extracted from the R*-CAT dataset.

3. Vamos!! Sal a votar y anima a todos a ir a votar.No a la independencia y sí a laconvivencia. #Iceta27S @miqueliceta http://t.co/0wSHlb5cCb
   Let’s go!! Go out to vote and convince everyone to go voting. No to independence and yes to living together. #Iceta27S @miqueliceta http://t.co/0wSHlb5cCb
   Spanish: “Independence of Catalonia” → AGAINST

4. Avui tenim una doble victória: ha guanyat el sí i ha guanyat la democrácia! Catalunya sí vol votar!!! #27S
   Today we have a double victory: the yes won and also democracy won! Catalonia wants to vote! #27S
   Catalan: “Independence of Catalonia” → FAVOR

[vi] The other four targets were: “Feminist Movement”, “Legalization of Abortion”, “Atheism’ and “Climate Change is a Real Concern”.

Table 3
Label distribution in the R*-CAT dataset.

Independence of Catalonia (Spanish)				Independence of Catalonia (Catalan)			
FAVOR	AGAINST	NONE	TOTAL	FAVOR	AGAINST	NONE	TOTAL
419	1,807	3,174	5,400	3,311	163	1,926	5,400

In Example 3, written in Spanish, the target "Independence of Catalonia" is explicitly mentioned and with it an encouragement of the user to other people to go out and vote no, perpetrating ideals of coexistence and sharing. On the other hand, in Example 4, written in Catalan, the user cheers for the victory of the yes within the context of the referendum. S/he explicitly states that Catalonia wants to vote.

New Datasets

French dataset (E-FRA). In the dataset E-FRA we collected tweets in French with the target "Emmanuel Macron" or "Marine Le Pen". The same annotation schema applied for the other datasets has been exploited, but we provided improved guidelines for the label NONE, which has been perceived as especially hard to be annotated. In particular, we detailed the directive for this label as follows: *We can infer from the tweet that the tweeter has a neutral stance towards the target, or there is no clue in the tweet to reveal the stance of the tweeter towards the target (support/against/neutral), or the tweeter considers the target to be the least bad choice.*

The first step of the annotation process consists in the creation by a domain expert of a 100 tweets gold standard for each of the targets. Then were recruited on Crowdflower native French speakers living in France and achieving an accuracy near to 70% when evaluated against this gold standard. 1,000 tweets for each target are then independently annotated for SD by three annotators on CrowdFlower, following the improved guidelines.

The IAA has been separately calculated for each on the two targets. The Fleiss' Kappa coefficient was $\kappa = 0.47$ on tweets targeting "Emmanuel Macron", and $\kappa = 0.44$ on those targeting "Marine Le Pen". Considering this IAA too low, we decided to discard all tweets in which an agreement was not reached by all three. The remaining tweets were 530 for the target "Emmanuel Macron" and 586 for the target "Marine Le Pen".

Table 4 shows the label distribution over the French dataset for both "Emmanuel Macron" and "Marine Le Pen" targets. As we can notice, the label distribution for both targets is skewing towards the label AGAINST. Following, we present two tweets extracted from the E-FRA dataset.

5. Je suis sarkoziste á 200% j ai vote Fillon mais Macron jamais alors pour la 1ére fois je voterais marine
   I am 200% sarkozist I voted Fillon but never Macron then for the first time I will vote for marine
   French: "Emmanuel Macron" → AGAINST

6. Je combats tout des idées de Madame Le Pen. Elle est déterminée; elle n'a pas compris que je l'étais encore plus q... https://t.co/70MUj74Ltm
   I fight every idea of Madame Le Pen. She is determined; she hasn't comprehended yet that so am I, even more... https://t.co/70MUj74Ltm
   French: "Marine Le Pen" → AGAINST

In Example 5 the user presents an AGAINST stance towards Macron, in fact, the author states that s/he will never vote for the candidate, but s/he would likely vote for his opponent. Also Example 6 presents a tweet labeled as AGAINST. In this case the target is "Marine Le Pen" whose ideas the user is in strong disagreement with.

Italian dataset (R-ITA). In the dataset R-ITA, the target of interest is the "Constitutional Referendum", and all the tweets are written in Italian. We applied the same annotation process exploited for developing E-FRA, but recruiting native Italian speakers that live in Italy rather than the French ones.

The IAA calculated with Fleiss' Kappa coefficient is $\kappa = 0.81$ and demonstrates a substantial agreement (almost perfect) among annotators. The released dataset includes only the 833 tweets obtained by discarding all those not featured by an agreement among all the annotators.

Table 5 shows the label distribution over the R-ITA dataset for the target "Constitutional Referendum". As we can notice, the label distribution is skewing towards the label AGAINST. The tweet below was extracted from the R-ITA dataset.

Table 4
Label distribution in the E-FRA dataset.

Emmanuel Macron				Marine Le Pen			
FAVOR	AGAINST	NONE	TOTAL	FAVOR	AGAINST	NONE	TOTAL
91	308	131	530	65	466	55	586

Table 5
Label distribution in the R-ITA dataset.

Constitutional Referendum			
FAVOR	AGAINST	NONE	TOTAL
163	486	184	833

7. 4 milioni di euro buttati nel cesso da #Renzi #bastaunNo per mandarlo a casa #IoVotoNO #referendumcostituzionale... https://t.co/jQ061sdfa0
   4 million euros thrown down the toilet by #Renzi #justNo to send him home #IVoteNO #constitutionalreferendum... https://t.co/jQ061sdfa0
   Italian: "Constitutional Referendum" → AGAINST

In Example 7 the author is AGAINST the "Constitutional Referendum", in fact, s/he states that s/he will never vote yes and s/he wants to "send Renzi home" (i.e. the Prime Minister who organized and promoted the Referendum).

To wrap up what described so far, in Table 6 we report an overview of the datasets and the distribution of labels and targets over the tweets. The table contains the number of tweets that overcame all phases of annotation that, as we explained in Section 3.2, were not discarded during the process. This is the multilingual test bed we provided for carring out the experiments described in the following sections.

4. Automatic stance classification

In the present research, we address SD as a classification task aiming at investigating it in a multilingual perspective. For this purpose we apply MultiTACOS, which is the extension of iTACOS, a system we successfully exploited in past experiments about SD for Spanish and Catalan only (Lai et al., 2017a,b). We propose novel experiments for developing an in-depth investigation of several supervised learning methods that seemed more promising in our previous work: SVM, Naïve Bayes (NB), and Logistic Regression (LR)[vii] We ran tests with the three methods for each target dataset and in each language as we will describe in detail in Section 5.1. The features we exploited for classifying the tweets with MultiTACOS are instead described in the following section.

We exploited four groups of features, namely Stylistic, Structural, Affective and Contextual. Provided that the first three groups are widely explored and well-known in literature, we will mainly focus on the last ones, which can be a novel contribution for the research area.

STYLISTIC FEATURES. First, we pre-processed all the tweets in order to have a lowercase version of them. Then, four different text representations were used:

- **Bag of Words** (*BoW*). We considered unigrams, bigrams and trigrams with binary representation.
- **Bag of Part-of-Speech**. The labels (*BoP*) extracted by TreeTagger[viii] were used in order to create a binary representation of unigrams, bigrams and trigrams of labels.
- **Bag of Lemmas** (*BoL*). The lemmas extracted by TreeTagger were used in order to create a binary representation of unigrams, bigrams and trigrams of lemmas.
- **Bag of Char-grams** (*BoC*). We exploited a binary representation of chars considering 2, 3, 4, and 5 char n-grams. We included all types of chars, also spaces, dots, commas, etc...

STRUCTURAL FEATURES. We also explore the use of structural characteristics in a similar way that Lai et al. (2017b).

- **Bag of Twitter Marks** (*BoTM*). We exploited the unigrams binary representation of the Bag of Words considering only the words extracted from multi-word Twitter Marks (hashtags and mentions).
- **Bag of Hashtags** (*BoH*). We considered the hashtags as terms for building a vector with binary representation of unigrams (Bag of Words).
- **Bag of Hashtags Plus** (*BoHplus*). We considered the tokens contained in the hashtags as terms for building a vector with unigrams binary representation (Bag of Words). In this case, we split the hashtag into tokens by capital letters or considering the tokens present in a dictionary. For choosing the tokens, we use a greedy algorithm considering, as optimal solution, the highest value of the average length of the tokens. We created a dictionary for each language considering the words present in the Wikipedia pages of each election/referendum event.[ix]

[vii] The scikit-learn implementation of the machine learning methods was used (scikit-learn.org).

[viii] TreeTagger (http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/). Schmid (1994, 1995) was used for extracting both part-of-speech and lemmas. We considered unigrams, bigrams and trigrams with binary representations of part-of-speech.

[ix] es.wikipedia.org/wiki/Proceso_participativo_sobre_el_futuro_político_de_Catalunya_de_2014, ca.wikipedia.org/wiki/Consulta_sobre_la_independéncia_-de_Catalunya, it.wikipedia.org/wiki/Referendum_costituzionale_del_2016_in_Italia, en.wikipedia.org/wiki/United_States_presidential_primary, en.wikipedia.org/wiki/Democratic_Party_presidential_primaries,_2016, en.wikipedia.org/wiki/Republican_Party_presidential_primaries,_2016, fr.wikipedia.org/wiki/élection_présidentielle_française_de_2017.

Table 6
Overview of label distribution across all datasets.

Language	Target	Label distribution			
		FAVOR	AGAINST	NONE	TOTAL
ENGLISH	Hillary Clinton	163	565	256	984
	Donald Trump	299	148	260	707
SPANISH	Independence of Catalonia	419	1,807	3,174	5,400
CATALAN		3,311	163	1,926	5,400
FRENCH	Emmanuel Macron	91	308	131	530
	Marine Le Pen	65	466	55	586
ITALIAN	Constitutional Referendum	163	486	184	833

- **Bag of Mention** (*BoM*). We considered the mentions as terms for building a unigrams binary representation (Bag of Words).
- **Frequency of Hashtags** (*freqHash*). We considered the number of hashtags present in the text as the only attribute for the vector representation.
- **Frequency of Mentions** (*freqMention*) We considered the number of mention tags present in the text as the only attribute for the vector representation.
- **Uppercase Words** (*UpW*). This feature refers to the amount of words starting with a capital letter. It consist of an only numerical attribute for the vector representation.
- **Punctuation Marks** (*PM*). We took into account the frequency of dots, commas, semicolons, exclamation, question marks and the frequency of all punctuation marks. The feature consists of a vector representation that contains six numerical attributes, one for each considered frequency.
- **Length** (*Length*). Three different numerical attributes were considered to build a vector of three elements: 1) number of words, 2) number of characters, and 3) the average of the word length in each tweet.

AFFECTIVE FEATURES. As it has been investigated in Lai et al. (2017b); and Mohammad et al. (2016a) SD is strongly related to Sentiment Analysis. Attempting to take advantage of this, we decided to exploit a set of features related to the affective content present in tweets. In doing so, we used different lexical resources defining different kinds of affective information, ranging from overall sentiment to finer-grained aspects. Below, we introduce the features we exploited:

- SENTIMENT-RELATED RESOURCES

  - ***AFINN***. AFFINN (Nielsen, 2011) is a lexical resource composed by almost 2,500 English words manually annotated with a polarity value in a range from −5 up to +5. It contains a set of words commonly used on the Internet as well as slang acronyms such as LOL (laughing out loud). We used the sum of the numerical values associated at the word contained in the text for calculating the total polarity of the tweet. The total polarity has been considered as the only attribute for the vector representation of the text.
  - ***HU&LIU***. Hu and Liu (2004) proposed two lists of terms related to sentiment (2,006 positive and 4,783 negative words) for opinion mining. According to this widely used resource we assigned the numerical value +1 to each positive word and −1 to each negative word. The total polarity of each tweet is obtained by summing the values associated to all words. The total polarity has been considered as the only attribute for the vector representation of the text.

- EMOTION-RELATED RESOURCES

  - ***LIWC***. The Linguistic Inquiry and Word Counts (LIWC) is a dictionary developed by Pennebaker et al. (2001). It contains more than four thousands words distributed in several categories for analyzing psychological aspects in written texts. Two categories related to emotions are included in this resource i.e, “posemo” and “negemo”. We assigned the numerical value +1 to each word categorized as “posemo” and −1 to each word categorized as “negemo”. We used the sum of the numerical values associated to the words contained in the text for calculating the total polarity of the tweet. The total polarity has been considered as the only attribute for the vector representation of the text.
  - ***DAL***. Whissell (2009) developed the Dictionary of Affect in Language (DAL) which contains 8,742 words annotated on a scale ranging from 0 up to 3 along three dimensions: Pleasantness, Activation, and Imagery. We used the numerical values associated to the words contained in the text for calculating the sum and the average ratings separately for each dimension. The sum and the average calculated for each dimension has been considered as the six attributes for the vector representation of the text.

All the resources described above for the affective features are developed for English. In order to exploit the same set of features in the other languages involved in our experiments (Spanish, Catalan, Italian and French), we applied them a translation via Google Translate APIs. This is a methodology commonly followed for languages other than English, in absence of any other language-tailored resource, although sometimes automatic translations are not precise and fully satisfying (Agarwal et al., 2011).

CONTEXTUAL FEATURES. Attempting to take advantage of contextual information, three features were included in this group. This kind of information has already proven to be useful in previous SD tasks (Lai et al., 2017b):

- **Language** (*Lan*). Due to the nature of the target of interest, the language could be used as a particular insight on user's position towards it. Here, we can use this feature only towards the target "Catalan Independence" due to the nature of the debate characterized by a request for independence of an autonomous community with a very high percentage of people understanding and speaking both Spanish and Catalan. We created a binary vector of two attributes exploiting the labels ES for Spanish and CA for Catalan provided by the organizer.
- **URL** (*Url*). We observed that tweets containing a URL are common in the datasets. We decided to take advantage of this by considering different pieces of information extracted from the short URL. Firstly, we identified whether the web address of reference was reachable or not. Second, we retrieved the original web address and we split it into tokens by dot. We finally build a binary bag-of-words vector representation of the tweet using the only tokens extracted from the URLs contained in the text. Unfortunately, it has not been possible to apply the same procedure to the English dataset, because as explained in Mohammad et al. (2016a), the tweets containing URLs were discarded in a pre-processing phase.
- **Domain Knowledge** (*Domain*). Lai et al. (2017b) explored domain knowledge in English tweets concerning Democratic and Republican Parties presidential primaries considering the type of relation among the involved politicians and parties. This feature encodes the types of relationship linking the targets, "Hillary Clinton" and "Donald Trump", and the other politicians and parties. We divided the types of relation in the following categories:

  - "Target": it identifies the explicit presence of the target (considering the target "Hillary Clinton" the examined keywords were *Hillary* and *Clinton*).
  - "Pronouns": the dataset was created considering only tweets referred to the target, so we considered the presence of a masculine or feminine pronoun as a reference to the target (considering e.g. the target "Hillary Clinton" we looked for the keywords *she* and *her*).
  - "Target's party": the feature identifies the presence of the party that supports the target (for example, the keyword *democratic* for the target "Hillary Clinton" and the keyword *republican* for the target "Donald Trump").
  - "Target's opponent in target's party": the primaries consist in a confrontation between candidates from the same party. The feature identified the presence of at least one member of the target's party (provided that *Bernie Sanders* was candidate against *Hillary Clinton* for the presidency of the democratic party, for this politician we considered the presence of the keywords *bernie* and *sanders*).
  - "Target's opponent in other parties": it considered the candidates for the presidential primaries in the opposite party (for example, provided that *Donald Trump* and *Ted Cruz* were both Republican Party candidates, and that a tweet in FAVOR of a Republican candidate was also against the target "Hillary Clinton", that is Democratic, we considered the presence of at least one keyword among *donald, trump, ted*, and *cruz*).

  In this research, we also need to represent and take into consideration the difference of the datasets' domains, i.e. presidential primaries elections and referendums. Therefore, we proposed a modified general set of features verifying the presence of involved entities in the text divided in the following categories:

  - "Target": the presence of the target (i.e. if the target is "Emmanuel Macron", the presence of the keyword *macron* and *emmanuel* was considered; in the case of "Independence of Catalonia" and "Constitutional Referendum", the keyword *referendum* was considered).
  - "Target's supporters": the presence of a supporter of the target was considered (e.g. in the case of "Emmanuel Macron" the keyword *brigitte*, Macron's wife; in the case of "Constitutional Referendum" the keywords related to politicians that promoted the reform, like *Renzi* or *Boschi*, were considered).
  - "Target's parties supporters": the presence of parties or movements that support the target was considered (i.e. for the target "Catalan indipendence" the presence of keywords referring to the Catalan independence coalition *Junts pel SÃ* was considered).
  - "Target's opponent": the presence of the target opponents (considering the target "Emmanuel Macron" the keywords related to opposition candidates were considered like e.g. *le pen* and *lepen* were considered).
  - "Target's parties opponent": In the last category the presence of the target's opponent party is considered (e.g. provided the target "Constitutional Referendum", the keywords related to the party *Movimento 5 Stelle*, which was against the reform, like *movimento 5 stelle* or *M5S*, were considered).

  We considered the mention of a list of entities for evaluating the presence of a specific type of relation in text. We finally used each type of relation as an attribute for realizing a binary vector representation of the text.

  The full list of keywords for each category and for each target, which was created by a domain expert for each topic, is freely available.[x]

[x] https://github.com/mirkolai/MultilingualStanceDetection

Table 7
Network information.

Target	Following-based network			Retweet-based network		
	Edges	Nodes	Detected communities	Edges	Nodes	Detected communities
Emmanuel Macron	532,637	256,359	33	6,582,849	869,390	9,809
Constitutional Referendum	897,545	514,660	27	588,132	111,094	2,251

- **User Community Knowledge (*Community*)**
Several works explore the social network structure of the relationships among users for improving the Sentiment Analysis classification of their posts (Xu et al., 2011; Deitrick and Hu, 2013). According to *Networks Science*, the entities involved in the network relationships are usually called *nodes*, while the relations among the nodes are usually called *edges*. A measure of the strength can be also assigned to each edge, which assumes the same value for all nodes in *unweighted* networks.
Two recent works focuses on SD using a network representation of the relationships among Twitter users involved in two different debates considering the networks based on following (Lai et al., 2017c) and on retweet (Lai et al., 2018) relationships.
In this work, we represent the relationships among Twitter users involved in the different debates in the form of graphs based on both following and retweet relationships. We extracted social media network communities from each graph using the Louvain Modularity algorithm (Blondel et al., 2008) and we used the communities as a binary feature, i.e. given the vector containing one instance for each community, the value will be 1 respecting to the community to which the author of the tweet belongs, and 0 otherwise.
Unfortunately, it has not been possible to apply the same process to the two benchmark datasets (i.e. English, Spanish-Catalan), because the datasets released by the organizers only contain the textual content of tweets without any information about the author. Then, we gathered the structure of following and retweet networks for both “Emmanuel Macron” (we do not explore the feature for the target “Marine Le Pen” due to it deals with a semi-supervised task) and the “Constitutional Referendum”. For what concerns the gathering of the following relationships, we take advantages of the GET friends/ids Twitter’s API for gathering the friend list of the author of each tweet contained in E-FRA (target ‘Emmanuel Macron”) and R-ITA.[xi]
Table 7 show the size of each network and the number of the communities retrieved by the Louvain Modularity algorithm.

As we will see in the following sections, combining features about textual contents, investigated within the NLP community, with those about social network structure, will allow us to achieve better results.

Summarizing, beside the first three groups of features described above, i.e. Stylistic, Structural and Affective, which are more canonical and widely exploited in NLP works, in the present research we highlight the importance of a fourth group of features, i.e. Contextual, as a novel way to exploit knowledge related to the tweets context. In the work of Lai (2019) both viewpoints (NLP and Network Analysis) co-exist and it is shown that merging the two approaches is indeed effective for gaining better performances in several tasks of SA, including SD. Inspired by those findings, even though our research develops in the frame of NLP, we elaborated on some key concepts from Network Science, by including in our model a novel set of *User Community Knowledge* features.

5. Results for stance classification

5.1. Evaluation metrics

We aim to investigate the portability of SD techniques across different languages, domains and machine learning algorithms, and mainly, to investigate the relevance of the different kinds of features.

We performed the training of all evaluated models with 3 different supervised learning algorithms[xii] using a combination of the 4 groups of features described in Section 4 such as *Stylistic, Structural, Affective*, and *Contextual*. Specifically, we trained one model for each combination of group of features for each proposed machine learning method. For each dataset is provided a 80−20% split between training and test sets. In particular, in the two benchmark datasets (E-USA and R*-CAT) the training set and the test set were released directly from the organizers of the shared tasks, while for the two new datasets (E-FRA and R-ITA) the splitting is randomly performed, maintaining the same ratio of 80−20% between training and test sets.

The macro-average of the F1-score metric (F_{avg} between f-AGAINST and f-FAVOR) proposed at Semeval 2016 (Mohammad et al., 2016b) and used also at IberEval 2017 (Taulé et al., 2017) was employed to evaluate the prediction of each trained model over the test set.

At SemEval 2016 the baselines were: (1) Majority class: a classifier that simply labels every instance with the majority class (‘favor’ or ‘against’) for the corresponding target; (2) SVM-unigrams: five SVM classifiers (one per target) trained on the corresponding training set for the target using word unigram features; (3) SVM-ngrams: a SVM classifier trained using word n-grams (1-, 2-, and 3-gram) and character n-grams (2-, 3-, 4-, and 5-gram) features; (4) SVM-ngrams-comb: a SVM classifier trained on

[xi] https://developer.twitter.com/en/docs/accounts-and-users/follow-search-get-users/api-reference/get-friends-ids
[xii] We reported only the results obtained with SVM, and LR because always better than those with NB.

Table 8
The highest F_{avg} values on E-USA dataset.

Target	LSTM one-hot	biLSTM one-hot	CNN one-hot	Classifier	uni-gram	Stylistic	Structural	Affective	Contextual	F_{avg}
Hillary Clinton	42.72	55.00	47.31	*LR*	58.18	✓	✓	✓		60.95
				SVM	58.51	✓		✓	✓	64.51
Donald Trump	27.87	29.40	26.19	*LR*	21.04		✓	✓	✓	55.74
				SVM	21.06		✓	✓	✓	55.42

the combined (all 5 targets proposed in the SemEval-2016 Task 6) training set using word n-grams (1-, 2-, and 3-gram) and character n-grams (2-, 3-, 4-, and 5-gram) features (Mohammad et al., 2016b).

At the StanceCat shared task of IberEval 2017 the baselines were Majority class and Low Dimensionality Representation (LDR) Taulé et al. (2017). Moreover, we compared each result obtained from proposed features with a model that used the same machine learning algorithm, but that was trained whit a simple baseline feature such as the binary uni-gram (uni-gram).

In order to further extend the comparison with other approaches, we also considered deep learning methods and we implemented three different neural architectures. In particular we centered our focus on those architectures that are exploited by state-of-the-art approaches, and which proved to obtain competitive results in SD: a Long Short-Term Memory neural network (LSTM) (used by MITRE (Zarrella and Marsh, 2016)), a biderectional Long Short-Term Memory neural network (biLSTM) (exploited Wei et al., 2018) and a CNN (exploited by Wei et al., 2016 and Zhou et al., 2019). The three of them have been implemented with the *one hot encoding* scheme. In the tables in the following paragraphs we report the achieved results.

5.2. Experimental results

We will now describe the experimental phase of our work, firstly comparing the results we obtained with the benchmark results obtained by the best teams competing in each task (SemEval 2016 for E-USA and StanceCat 2017 for R*-CAT (Mohammad et al., 2016b; Taulé et al., 2017). Subsequently we will explore and comment on the experimental results obtained on the new datasets (E-FRA and R-ITA).

Several experiments have been conducted across all datasets, comparing the results in each phase. Our goal was to explore the significance of features in different environments and to test whether the results obtained could be considered language-independent or topic-independent.

Benchmark Datasets

English Dataset (E-USA). We conducted the experiments over the E-USA dataset under a supervised framework for the target "Hillary Clinton" and under a semi-supervised framework for the target "Donald Trump". As we can see from Table 8, the best result for "Hillary Clinton" is obtained with a model that exploits SVM as machine learning algorithm trained with *Stylistic, Affective*, and *Contextual* features. We trained the model for "Donald Trump" with the tweets about the target "Hillary Clinton" due to the fact that no training set exits for "Donald Trump". The best model for "Donald Trump" exploits LR, but similar results are obtained using SVM. This setting also explains the big difference in the results obtained with all three neural models regarding Clinton and Trump, for which the biggest difference between the two targets is recorded with biLSTM (Δ = 25.06).

Both the best results with LR and SVM were obtained training the models with *Structural, Affective*, and *Contextual* features. As we can notice, both the best performing models (results in bold) exploit *Affective* and *Contextual* features.

In Table 9 we compare the results obtained by MultiTACOS with the official results achieved at SemEval-2016 Task 6. MultiTACOS obtains very competitive results (64.51 vs 67.12 and 55.74 vs 56.28).

Table 9
Our result compared with official results at SemEval-2016 Task 6.

Hillary Clinton		
Baselines		
	Majority class	36.83
	SVM-unigrams	57.02
	SVM-ngrams	58.63
	SVM-ngrams-comb	56.50
Participating Teams		
Rank	*Team*	*Result*
1	TakeLab	67.12
	MultiTACOS	64.51
2	pkudblab	64.41
3	pkulcwm	62.26
4	uwb	59.82
5	idi@ntnu	57.89

Donald Trump		
Baselines		
	Majority class	29.72
	SVM-ngrams-comb	28.43
Participating Teams		
Rank	*Team*	*Result*
1	pkudblab	56.28
	MultiTACOS	55.74
2	LitisMind	44.66
3	inf-ufrgs-opinion-mining	42.32
4	uwb	42.02
5	ecnu	34.08

Table 10
Results achieved in further works that used the E-USA dataset.

Hillary Clinton		Donald Trump	
Model	*Result*	*Model*	*Result*
TGMN-CR (Wei et al. (2018))	66.21	conditional bi-LSTM (Augenstein et al. (2016))	49.01
TAN (Dey et al. (2018))	65.38		
AT-biGRU (Zhou et al. (2019))	57.94		

We draw a distinction between the results obtained by our system on the tweets concerning the target of “Hillary Clinton”, for which we scored 64.51 F_{avg}, and on the tweets concerning the target “Donald Trump”, for which the score is 55.74 F_{avg}. In line with the results achieved by all other participating teams, in our scores a difference of almost 10 points can be noted. This can be explained by observing that the tweets of the training set used for the SD on the target “Donald Trump” were about another target entity. In the same table we also report the scores of the baselines of the shared task: Majority class, SVM-unigrams, SVM-ngrams, and SVM-ngrams-comb.

In Table 10, for completeness, we also report the results of some works who exploited the English benchmark dataset for SD after the SemEval-2016 competition ended. In particular we report the results of those works that provided their values with the F_{avg} measure and in which the targets of “Hillary Clinton” and/or “Donald Trump” were the main focus, in order to easily compare them within our framework. Several other researchers exploited the same dataset, years after the competition, but either they worked on other topics such as “Legalization of Abortion” and “Climate Change is a Real Concern”, or they reported their results referring to other metrics. These conditions made those works not useful in our setting.

The majority of the reported works used deep learning models. LSTM was explored in both supervised and semi-supervised tasks (Augenstein et al., 2016; Dey et al., 2018). In particular Dey et al. (2018) proposed a model based on both RNN and LSTM taking advantage of the target specific attention. RNN-based method was also proposed by Zhou et al. (2019). The best result was reached by TakeLab (Tutek et al., 2016) (supervised approach) using an ensemble of learning algorithms tuned with a genetic algorithm and by pkudblab (Wei et al. 2016) (semi-supervised approach) using keyword rules. Our method, based on machine learning algorithms combined with *ad hoc* features, proves to be competitive when compared with the state of the art.

Spanish-Catalan Dataset (R*-CAT). We conducted the experiments over the R*-CAT dataset under the same supervised framework for both languages, training the classifiers on a training set constituted by tweets in both languages.

As we can see from Table 11, the best result for the target “Independence of Catalonia” in Spanish is obtained with a model that exploits SVM as machine learning algorithm trained with *Stylistic, Structural* and *Affective* features the best result that our system obtains in Catalan is 48.05 using LR combined with *Structural* and *Affective* features, but it is not enough to reach the results obtained exploiting a system that uses LR trained with the UNI-GRAM baseline which is 50.97. The low results do not come as a surprise, in fact, in StanceCat at IberEval 2017 (Taulé et al., 2017), for the sub-task concerning tweets in Catalan, only one system outperformed the proposed Majority Class baseline[xiii] As we can notice, the two best performing models exploit *Affective* and *Structural* features. Additionally the only time that *Contextual* features are used, is for combination with LR in tweets in Spanish.

In Table 12 we compare the results obtained by MultiTACOS with the official results in StanceCat at IberEval 2017. As we can see MultiTACOS obtained top scores both in Spanish and Catalan.

The results obtained with MultiTACOS, developed within the present research, are lower than the ones obtained with the original system iTACOS due to the fact that we considered features in an aggregated way in order to have more advantages in a multilingual perspective and better explore the diverse characteristics of the different groups of features. On the other hand the results of the iTACOS systems are higher because the set of features that we exploited in Lai et al. (2017a) were specifically tailored for the StanceCat task.[xiv]

The majority of teams used SVM, but also neural networks, deep learning, and RBF kernels-based approaches (Taulé et al., 2017). iTACOS (Lai et al., 2017a) obtains the highest results for both Spanish and Catalan sub-tasks experimenting with SVM, logistic regression, decision trees, random forest and multinomial NB.

Few authors explored other approaches after the IberEval-2017 shared task. In particular Respall and Derczynski (2017) improved the results of the participating teams using an SVM-based approach, but only for the Spanish sub-task. After IberEval-2017 a second shared task on SD in Catalan and Spanish tweets was held at IberEval-2018 (Taulé et al., 2018). In this contest the target for SD was the “Independence of Catalonia” and a newly released corpus of the two languages tweets was provided by the organizers. In order to propose a multi-modal perspective, the textual content and the information included in the URLs and images both were made available to be taken into account for determining the stance of the posts. Considering the textual perspective only (without using links and images), all four participating teams took advantage of SVM using tf-idf weight and both character and word n-grams in at least one of the runs they submitted. Only one team used word embeddings and CNNs.

[xiii] See Lai et al. (2017a), and Taulé et al. (2017).

[xiv] In the shared StanceCat task at IberEval 2017 we submitted five runs for SD in both languages, i.e. five models for Catalan and five models for Spanish. In Table 12 they are listed as iTACOS.1, iTACOS.2, etc....

Table 11
The highest F_{avg} values on R*-CAT dataset.

Target	LSTM ONE-HOT	BILSTM ONE-HOT	CNN ONE-HOT	Classifier	UNI-GRAM	Stylistic	Structural	Affective	Contextual	F_{avg}
Catalan Independence (Spanish)	41.14	39.45	38.76	*LR*	44.94	✓	✓	✓	✓	47.78
				SVM	42.01	✓	✓	✓		**48.30**
Catalan Independence (Catalan)	43.38	47.21	**51.64**	*LR*	50.97		✓	✓		48.05
				SVM	46.84	✓	✓	✓		45.89

Table 12
Our result compared with official results at IberEval 2017.

Catalan Indipendence (Spanish)		
Baselines	*Majority class*	44.79
	LDR	41.35
Participating Teams		
Rank	*Team*	*Result*
1	iTACOS.1	48.88
	MultiTACOS	48.05
2	LTRC_IIITH.system1	46.79
3	LTRC_IIITH.system4	46.40
4	ELIRF-UPV.1	46.37
5	ELIRF-UPV.2	46.37

Catalan Indipendence (Catalan)		
Baselines	*Majority class*	48.82
	LDR	43.75
Participating Teams		
Rank	*Team*	*Result*
1	iTACOS.2	49.01
2	iTACOS.1	48.85
	MultiTACOS	48.30
3	iTACOS.3	46.85
4	LTRC_IIITH.system1	46.75
5	ARA1337.s1	46.59

The highest result for Spanish was achieved by Cuquerella and Rodríguez (2018) using a model based on SVM trained with bag-of-words weighted with tf-idf. The highest result for Catalan was achieved by Segura-Bedmar (2018) training an SVM with a bag-of-words representation with the stem of the words and weighted with tf-idf.

New Datasets

French Dataset (E-FRA). We carried out the experiments over the E-FRA dataset under a supervised framework for the target "Emmanuel Macron" and under a semi-supervised framework for the target "Marine Le Pen" with the aim of emulating a procedure similar to the one we used for the E-USA dataset.

As we can see from Table 13, the best result for both "Emmanuel Macron" and "Marine Le Pen" is obtained with a model that exploits LR as machine learning algorithm trained with *Affective*, and *Contextual* features. The same is valid also for the best performing model with SVM.

We trained the model for "Marine Le Pen" with the tweets about the target "Emmanuel Macron". We decided to not create a training set for "Marine Le Pen" as well as no training set exits for "Donald Trump" in the E-USA dataset and we wanted to maintain coherence among datasets of the same typology. The best model for "Marine Le Pen" exploits LR trained with *Affective*, and *Contextual* features.

We operate a distinction between the results obtained by our system on the tweets concerning the target of "Emmanuel Macron", for which we scored 68.65 F_{avg} (trained with LR) and the results obtained by our system on the tweets about the target "Marine Le Pen" for which the score is 48.57 F_{avg} (trained with LR). The difference of almost 20 points it is not surprising because all the models for the target "Marine Le Pen" were trained with a training set of tweets concerning the other target. Let us highlight the fact that the models trained for the target "Marine Le Pen" could not take advantage of the feature based on the information about the author's community; we could exploit this kind of contextual feature only in the supervised framework.

As it happened in the English scenario, also in the French dataset we can see how the results of the biLSTM approach worsen in the unsupervised framework (target "Marine Le Pen") with respect to the performances obtained in the supervised scenario "Emmanuel Macron". While, surprisingly the approaches based on LSTM an CNN do not seem to be strongly influenced by these

Table 13
The highest F_{avg} values on E-FRA dataset.

Target	LSTM ONE-HOT	BILSTM ONE-HOT	CNN ONE-HOT	Classifier	UNI-GRAM	Stylistic	Structural	Affective	Contextual	F_{avg}
Emmanuel Macron	55.08	61.59	55.64	*LR*	51.69			✓	✓	**68.65**
				SVM	52.57			✓	✓	67.41
Marine Le Pen	50.61	39.80	**52.97**	*LR*	38.63			✓	✓	48.57
				SVM	34.52			✓	✓	45.58

Table 14
The highest F_{avg} values on R-ITA dataset.

Target	LSTM one-hot	biLSTM one-hot	CNN one-hot	Classifier	uni-gram	Stylistic	Structural	Affective	Contextual	F_{avg}
Constitutional Referendum	89.56	76.66	**97.13**	*LR*	94.17	✓			✓	95.93
				SVM	95.11	✓			✓	95.57

Table 15
The highest F_{avg} values on R-ITA dataset removing polarized hashtags and all hashtags.

Removing	LSTM one-hot	biLSTM one-hot	CNN one-hot	Classifier	uni-gram	Stylistic	Structural	Affective	Contextual	F_{avg}
Polarized Hashtags	61.62	75.85	71.56	*LR*	72.33		✓	✓	✓	**90.64**
				SVM	73.04		✓		✓	87.62
All Hashtags	51.72	67.21	63.43	*LR*	56.43				✓	**86.81**
				SVM	61.49			✓	✓	86.36

dimensions. In any case, also in French, our approach based on classical machine learning models is competitive with regard to the approaches based on deep models.

Italian Dataset (R-ITA). We conducted the experiments over the R-ITA dataset under a supervised framework. As we can see from Table 14, the best result for the target "Constitutional Reform" in Italian is obtained with a model that exploits LR as machine learning method trained with *Stylistic*, and *Contextual* features. Surprisingly, *Affective* and *Structural* features do not appear in neither of the two best results that we report. Our intuition behind this situation lies in the fact that we believe the Italian dataset to be particularly *sui generis* when compared with the other three. The exploitation of hashtags is wide and coherent in the whole corpus. For instance the hashtags #iovotosí (#Ivoteyes) and #iovotono (#Ivoteno) have been exploited almost in each tweet that we took into consideration, and we believe that just their presence (as boolean value) already is a clear manifestation of stance. For this reason *Stylistic* features such as Bag of Words and *Contextual* are already sufficient to reach extremely high F-scores (95.93 F_{avg}).

The approaches based on deep learning architectures obtain results comparable to our models, being the one based on a CNN the one that performs best (97.13F_{avg}) overall also surpassing our model's performances. While the other two, based on LSTM and biLSTM are lower.

In order to explore the importance of some features and in particular, those who exploit the use of hashtags, we performed a separate experiment removing the polarized hashstag #iovotosí (#Ivoteyes), #iovotono (#Ivoteno), #hovotatosi (#Ivotedyes), #votiamono (#wevoteno) etc. from the text of the R-ITA tweets[xv] After this operation, as showed in Table 15, LR achieved the highest result (90.64 F_{avg}) using *Structural, Affective*, and *Contextual* features. *Contextual* features gain a particular significance for SD when explicit information derived from hashtagging the tweet goes missing or, in this case, is explicitly removed.

It is important to note that also when completely removing all hashtags[xvi], LR trained with the *Contextual* feature achieved a high F-measure (86.81 F_{avg}).

It is interesting to see how the performance of CNN, in both the settings in which the hashtags are removed drops significantly (Δ = 25.57 removing only polarized hashtags and Δ = 33.70 removing all hashtags). While our systems' performances, based on classical machine learning algorithms, but combined with *ad hoc* SD features do not drop so drastically (Δ = 9.57 ± 4.93).

A general conclusion of the analysis of the results is that removing hashtags, obviously decreases the quality of results, but at the same time sheds some light on the importance of *Contextual* features in SD, as already explored in Lai et al. (2017b, 2017c, 2018).

5.3. Feature analysis discussion

The experiments we performed allowed us to focus on the behaviour of the groups of features both in the different five languages, i.e. English, Spanish, Catalan, French and Italian, and in the different political domains, i.e. elections and referendums. Hypothesizing that they may mirror the differences in users' styles for communicating stance towards target entities, we detected the contribution provided by each feature and by combinations of them by performing a features analysis whose results are reported in Tables 16 and 17. In both tables are displayed the five best-performing features concerning each target, in combination either with LR or SVM, on the election datasets (E-USA and E-FRA) and on those about referendum (R-ITA, R*-CAT).

Among the different perspectives that we can take for analyzing the results of the feature analysis we performed, first of all we observe the results with respect to the groups of features (*Stylistic, Structural, Affective* and *Contextual*) and then with respect to languages (English, Spanish, Catalan, French and Italian) and political domains (referedums and elections), also considering the algorithms applied (SVM and LR).

[xv] We used the following regular expressions not distinguishing between letters that only differ in case #([a-z]{0,}vot[a-z]{1,}) for removing polarized hashtags.

[xvi] We used the following regular expressions #(w+) for removing all hashtags.

Table 16
The ranking of the results obtained with LR and SVM on English (E-USA) and French (E-FRA), by separately considering only the top-5 performing features.

Dataset	Target	Algorithm	1°	2°	3°	4°	5°
E-USA (English)	Hillary Clinton	LR	*BoC*	*HU&LIU*	*AFINN*	*BoP*	*BoW*
			57.77	45.12	43.23	41.76	41.46
		SVM	*BoC*	*BoW*	*BoL*	*BoTM*	*BoHplus*
			57.55	46.33	45.43	43.31	40.29
	Donald Trump	LR	*HU&LIU*	*BoP*	*AFINN*	*BoH*	*freqHash*
			33.97	31.76	30.12	29.85	29.72
		SVM	*BoP*	*BoHplus*	*BoH*	*freqHash*	*freqMention*
			31.41	30.36	29.76	29.72	29.72
E-FRA (French)	Emmanuel Macron	LR	*Community Retweet*	*Community Following*	*BoC*	*BoHplus*	*BoP*
			65.16	59.54	57.44	45.80	44.17
		SVM	*Community Retweet*	*Community Following*	*BoC*	*BoP*	*BoHplus*
			68.00	59.07	56.68	53.89	46.29
	Marine Le Pen	LR	*DAL*	*HU&LIU*	*BoC*	*BoP*	*BoL*
			45.30	44.75	44.57	44.30	44.30
		SVM	*BoH*	*BoP*	*BoL*	*freqHash*	*freqMention*
			45.50	44.30	44.30	44.30	44.30

Table 17
The ranking of the results obtained on Spanish and Catalan (R*-CAT) and Italian (R-ITA) by separately considering only the top-5 performing features.

Dataset	Target	Algorithm	1°	2°	3°	4°	5°
R*-CAT	Catalan Independence (Spanish)	LR	*DAL*	*BoC*	*BoTM*	*BoW*	*BoL*
			51.30	49.04	45.91	45.03	44.63
		SVM	*DAL*	*BoC*	*BoTM*	*BoL*	*BoW*
			50.30	48.58	45.27	44.21	43.04
	Catalan Independence (Catalan)	LR	*BoW*	*DAL*	*BoTM*	*BoC*	*BoL*
			49.28	49.3	48.79	48.69	46.31
		SVM	*BoW*	*BoL*	*DAL*	*BoC*	*BoTM*
			49.30	48.30	48.30	44.22	41.23
R-ITA (Italian)	Constitutional Referendum	LR	*BoW*	*BoC*	*BoTM*	*BoHplus*	*BoH*
			95.34	94.36	92.88	92.61	92.61
		SVM	*BoC*	*BoW*	*BoTM*	*BoH*	*BoHplus*
			95.06	94.84	94.20	93.93	93.21

As a general consideration, the approach proposed allows to achieve promising results in a multilingual setting, and especially some of the features we explore seem to work well independently from the target language.

Focusing on the groups of features, we can see that among the features exploited by almost all the best performing models for each of the five languages we can find the *Stylistic* features, such as *BoW*, *BoC*, and *BoP*. This underlays hypotheses, supported by many related work, about the representativeness of social media data of these straightforward features. In particular, the *Stylistic* feature *BoC* performs well in all five languages, always ranking in the 5 best-performing features as we can see from Tables 16 and 17.

For what concerns the presence of Twitter marks (feature *BoTM*), we can also observe that it seems not more influenced by the language or by whether the approach is supervised or semi-supervised, than by the typology of target or the nature of the dataset. The presence or absence of Twitter marks, which is especially noticeable in the Spanish, Catalan (R*CAT) and Italian (R-ITA) datasets, supports indeed the inference that the use of hashtags and mentions is wider in campaigns for referendum than in those for political elections.

Among the *Affective* features, we can observe that *HU&LIU* alone obtains really good results in the election datasets (E-USA and E-FRA) leading to the insight that an affective lexicon might prove more useful when the target of interest are people, as in the case of political elections, and less useful when the target is a referendum or a reform. The contribution of *HU&LIU* seems moreover relatively independent from the language involved, regardless on the fact that this is a resource developed for English and only available as (non manually revised) translation for the other languages. We can indeed observe that also for the target ‘Emmanuel Macron” the F_{avg} for the feature *HU&LIU*, even if not scored in the best five positions (and not included in Table 16) and outperformed by *Contextual* features like *Community Retweet* and *Community Following*, are still quite high and well comparable to those achieved for the other targets (LR = 42.35, SVM = 37.65).

In the R*CAT dataset, furthermore, it is interesting to notice how an *Affective* resource, such as *DAL* alone obtains very good results in Spanish (LR = 51.30 and SVM = 50.30), also outperforming simple approaches such as *BoW*, *BoC* and *BoL*. The resource *DAL* has been exploited in at least two different contexts: in the supervised dataset E-FRA (target “Emmanuel Macron”) and in the R*-CAT dataset (Spanish portion of the data), underlining how affective resources could be of great help in different tasks, domains and applied to different targets.

The *Contextual* features and in particular the *Community Retweet* and the *Community Following* perform really well on the R-ITA dataset (*Community Retweet*: LR = 84.96 and SVM 84.52. *Community Following*: LR = 57.67 and SVM = 59.52), their values are not reported in Table 17 simply because they do not rank in the best five performing ones.

Table 18
The combinations of the three best-scored features on the E-USA and the E-FRA datasets. The features not used in at least one of the best combinations are not shown, and "-" indicates unavailable features (in benchmark datasets released in the context of evaluation campaigns).

Dataset	Target		$\mathbf{F_{avg}}$	BoP	BoL	BoC	BoH	BoM	freqHash	PM	AFINN	DAL	*HU&LIU*	Domain	Community Following	Community Retweet
E-USA	Hillary Clinton	LR	62.17			✓				✓	✓				-	-
		SVM	62.69		✓								✓	✓	-	-
	Donald Trump	LR	49.90					✓					✓	✓	-	-
		SVM	49.69					✓					✓	✓	-	-
E-FRA	Emmanuel Macron	LR	71.20									✓	✓			✓
		SVM	70.95	✓											✓	✓
	Marine Le Pen	LR	50.83						✓				✓	✓	-	-
		SVM	51.94		✓		✓						✓		-	-

As previously said, the *"Community Features"* could be only applied in the case of the R-ITA dataset and E-FRA dataset (i.e. those we created), while the benchmark datasets, distributed within evaluation campaigns, did not contain metadata enabling the collection of social network information, but only provided the textual content of the tweet.

Finally for what concerns *Structural* features, the hashtags play a really important role in the Italian "Constitutional Referendum", allowing to reach surprisingly high F_{avg} in particular with the features *BoH* (LR= 92.61 SVM = 93.21) and *BoHplus* (LR = 92.61, SVM = 93.21). The exploitation of hashtags is indeed wide and coherent in this whole corpus, see e.g. *#iovotosí* (#Ivoteyes) and *#iovotono* (#Ivoteno), which have been exploited almost in each tweet that belong to the R-ITA dataset, and we believe that just their presence already is a clear manifestation of stance and helps the automatic system, as already commented under Table 15.

If we assume languages and domains as our main reference for the analysis of the results presented in Tables 16 and 17, we can see that the results seem more influenced by domains than by language. In particular *Affective* and *Contextual* features are relatively language-independent and in general they produce better results over datasets in which the target is a person (i.e. election datasets: E-USA, E-FRA). Moreover, an ablation test conducted on the *Contextual* group of features demonstrated that the feature *Common Knowledge* is more relevant in supervised contexts where the target is indeed a person.

On the other hand, the *Language* feature is particularly discriminating with the target "Independence of Catalonia" where nationalist feelings play a big role and the Catalan language itself is exploited to convey Catalan independentist attitude.

Fur further investigating the contribution of the features with respect to political domains, we provide Tables 18 and 19, where are shown the results obtained with the best combinations of a maximum of three features, respectively on the election datasets (E-USA and E-FRA) and on the referendum datasets (R*-CAT and R-ITA).

Comparing Tables 18 and 19 we can see the relevance of the *Affective* feature *HU&LIU*, already cited above, regardless of the algorithm applied in the election datasets, but not in the referendum ones.

Another feature that is well scored (in all the cases where it is available) in both election and referendum datasets is *Community Retweets*, a *Contextual* feature, which in the R-ITA dataset combined with *BoW* and *Community Following* leads to F-score of 98.49 with Logistic Regression algorithm. Similarly, for the target "Emmanuel Macron" in the E-FRA dataset, the feature *Community Retweets* combined with *BoW* and *Community Following* leads to 70.95 with the SVM algorithm. As previously said, the *"Community Features"* could be only applied in the case of the R-ITA dataset and E-FRA dataset (i.e. those we created), while the benchmark datasets, distributed within evaluation campaigns, didn't contain metadata for collecting it.

In Table 19, we can see how the most simple combination of three features is that obtained with LR algorithm onto the Catalan subset of the R*-CAT dataset. Here only really straightforward *Stylistic* features have been used: *BoW, BoL* and *UpW*. The following two important features for the Catalan language (even if it is not displayed in the table) are the use of hashtags and mentions (*BoTW*) and the use of the Catalan language itself (*Lang*), which alone already obtains an F-score of 38.02 with both LR and SVM.

In conclusion, the more interesting finding of all the experimental settings is the good results obtained by the *Contextual* features, in particular *Community Retweet* and *Community Following*. The results of several experiments and tests with different types of features, confirm moreover the contribution of the *Stylistic* features in all supervised contexts and their lower contribution in semi-supervised contexts. The same happens when we tested *Structural* features, which perform better in supervised contexts, especially thanks to features connected with Twitter Marks (hashtags and mentions) often exploited by users for expressing the stance in a debate. On the other hand, in the semi-supervised contexts, the best results are obtained using models which exploit *Affective* and *Contextual* features.

5.4. Error analysis

In this section we provide a qualitative analysis of errors in SD occurred in the different experiments presented above for the four debates at issue. In particular, we examined several failure cases to identify possible causes of errors, with the twofold aim to identify error classes, on the one hand, by analysing the specific language-debate settings, and on the other hand, by trying to discover error patterns which are occurring in different languages and political debates. For each language and target, each tweet in the set of the misclassified tweets was individually annotated with possible causes of errors by at least one of annotator, and the results were collectively discussed to identify potential reasons and error patterns. In the following, we report and discuss notable error classes resulting from our analysis.

Error Patterns. As a general consideration, it is interesting to notice that in almost all the debates considered in the different languages, the most frequent kind of 'total' stance misclassification error, i.e. when a classifier assigns the opposite stance with respect to what is expected according to the gold standard, is the following: the classifiers interpreted a stance as being "in favor" when the real value was "against" (F → A). See Table 20.

Only in the case of Catalan, the A → F error rate is higher. Our hypothesis is that this is due to a bias resulting from the difference in the number of tweets classified according to the stance expressed: there is a considerable higher number of tweets in favor of the target (independence of Catalonia) in the Catalan dataset, compared with the amount of tweets against the target. No dataset for the other languages is so unbalanced towards the "favor" class.

Notable error classes included:

- **Sarcasm, metaphors, and other figurative language devices (sarcasm)**. Occasionally, in all languages tweets contain sarcasm, metaphors, or other figurative devices, such as *rethorical questions*, that can be difficult for the model to properly

Table 19
The combinations of the three best-scored features on the R*-CAT and the R-ITA datasets. The features not used in at least one of the best combinations are not shown, and “-” indicates unavailable features (in benchmark datasets released in the context of evaluation campaigns).

Dataset	Target		$\mathbf{F}_{\mathbf{avg}}$	BoW	BoP	BoL	BoC	BoH	BoHplus	UpW	AFINN	DAL	Url	Community Following	Community Retwert
R*-CAT	Independence of Catalonia (Spanish)	LR	51.20									✓		-	-
		SVM	50.43			✓	✓		✓					-	-
	Independence of Catalonia (Catalan)	LR	53.64	✓		✓				✓				-	-
		SVM	52.74			✓		✓					✓	-	-
R-ITA	Constitutional Referendum	LR	98.49	✓										✓	✓
		SVM	97.45		✓						✓				✓

Table 20
Error percentages: gold standard stance → predicted stance.

Language	Target	Error	
		A → F	F → A
ENGLISH	Hillary Clinton	6.67%	36.00%
	Donald Trump	8.79%	25.08%
SPANISH	Independence of Catalonia	2.50%	8.10%
CATALAN		4.20%	2.70%
FRENCH	Emmanuel Macron	3.23%	19.35%
	Marine Le Pen	19.75%	25.93%
ITALIAN	Constitutional Referendum	14.29%	0%

comprehend. For example, see the following example for Italian, which is a clear case of ‘polarity reversal’ (one says something good to mean something bad). Here the correct label is ‘against’, but the system misclassified it as ‘in favor’, probably because the system did not address figurative language:

R-ITA. Target: “Constitutional Referendum”

@serracchiani @bastaunsi questo è far decidere liberamente gli #Italiani? al #ReferendumCostituzionale
@serracchiani @bastaunsi is this letting #Italians freely decide? #ReferendumCostituzionale

Also in the following French posts, the system did not correctly recognize the negative stance towards Macron because of the presence of a *rhetorical question*:

E-FRA. Target: “Emmanuel Macron”

Prélèvement à la source : la première catastrophe industrielle du président Macron ?
Tax withholding at source: the first industrial catastrophy of president Macron?

In the following tweet we can observe the presence of irony towards the target “Marine LePen” and also the exploitation of an emoji:

E-FRA. Target: “Marine Le Pen”

Marine le Pen c’était celle qui copiait sur toi et qui a une meilleure note que toi 😂
Marine Pen was the one who copies you and has a better rating than you 😂

In the Spanish tweet below, from the R*-CAT dataset, the stance is misjudged probably due to the presence of an analogy between Romeva’s speech and Lewis Carrol’s masterpiece ‘Alice in Wonderland’.

R*-CAT. Target: “Independence of Catalonia”

Puedo entender el deseo de muchos independentistas pero el discurso de Romeva es el nuevo Alicia en el país de las maravillas. #27S
I can understand the hope of several separatists but Romeva’s speech is the new Alice in Wonderland. #27S

Instead, in the following tweet from the E-USA dataset our system did not recognize the stance in favor of Donald Trump, because of the subtle word pun, based on the figurative use of ‘trump card’ and on homonyms (in card games like bridge the ‘trump card’ is the most powerful one among the cards of the same suit):

E-USA. Target: “Donald Trump”

@realDonaldTrump You are the trump card of my heart. #SemST

A similar problem is encountered in the following French post, where some fine-grained semantic consideration is necessary for interpreting the meaning.

E-FRA. Target: “Marine Le Pen”

Choisir entre Le Pen et Macron, c comme choisir entre un âne et un poney Tu choisis le poney mais tu sais que ça ne t’emménera pas trés loin
Choosing between Le Pen and Macron, is like choosing between a donkey and a pony You choose the pony but you know that it will not take you very far

The semantics of phrases like the ones mentioned above are likely hard for the model to learn without a variety of similar training examples to consider.

- **Opinions expressed but related to different entities than the target of interest** *(opinions)*. In SD, systems must determine the author's favorability towards a given target. Stance could be inferred also in tweets where the target is not explicitly mentioned. Moreover, the text may express an opinion about some other entity, even if the target is mentioned. In all such cases, stance must be inferred. We observed, instead, that especially in the French dataset, our system sometimes misclassifies the stance. See for instance the following tweet, where the tweet's author is expressing a positive sentiment towards Le Pen, the rival candidate in the presidential election, so the human annotators inferred that the stance towards Macron is 'AGAINST', but our system assigned the uncorrected label 'FAVOR':

  E-FRA. Target: "Emmanuel Macron"

  Le Pen elle est trop forte elle vient de baiser Macron en allant directement à l'usine #Presidentielle2017
  The Pen is too strong she just fucked Macron by going directly to the factory #Presidentielle2017

  We observed various error cases reflecting this pattern in the E-FRA with target "Emanuel Macron": the authors express a sentiment towards Le Pen, the opinion is not explicitly referred to the target and this makes especially difficult for the system to infer the correct stance towards Macron.

  E-FRA. Target: "Marine Le Pen"

  Bon Macron c'est mieux que Le Pen, mais c'est moins bien que Mélenchon, mais c'est mieux que Le Pen...
  Well Macron is better than Le Pen, but it's worse than Mélenchon, but it's better than Le Pen ...

  The same situation applies in the R-ITA dataset, where many times the stance is misjudged towards the target of interest "Constitutional Referendum" due to the presence of opinions towards Matteo Renzi, the Prime Minister who assumed full responsibility for the referendum defeat.

  R-ITA. Target: "Constitutional Referendum"

  C'é cosí tanto #Renzi in tv che sto pensando di chiedergli di contribuire a pagare il canone. #referendumcostituzionale #referendum
  There is so much #Renzi on TV that I'm thinking of asking him to help pay the TV license fees. #referendum #constitutionalreferendum

  In the two following tweets the focus is on a candidate of the Ciudadanos party (Ines Arrimada). Our system is not able to differentiate the concept of target of interest and the focus on the named entity, and therefore, to establish a relationship between the unionists' candidate and her stance towards the matter of Independence of Catalonia.

  R*-CAT. Target: "Independence of Catalonia"

  @InesArrimadas xata, primer apren a comptar, després ja parlarem!!! ???? #27STV3
  @InesArrimadas honey, first learn to count, then we will talk with you! ???? #27STV3

  R*-ESP. Target: "Independence of Catalonia"

  Arrimadas de que quieres que dimita Mas si en estos momentos no tiene ningún cargo? #ciutadans #27s
  Arrimadas from what do you want Mas to resign if at this time he has no political position? #ciutadans #27s

- **Background knowledge and commonsense** *(background)*. In many cases users do not express their stance in an explicit manner. However, an evaluation of it could be inferred by human annotators by relying on *common sense knowledge* or *world knowledge*, as in the following tweet. Here word knowledge is necessary in order to get the sarcastic connotation and to infer the negative stance ('against'):

  R-ITA. Target: "Constitutional Referendum"

  Dall'Europa ci supportano!
  Europe is supporting us!

  In the following tweet, the user makes a specific reference to the episode of the killings in Benghazi through the hashtag #REMEMBERBENGHAZI2016. To correctly infer the stance our system should have *world knowledge* about what happened.

  E-USA. Target: "Hillary Clinton"

  @lylafmills Simple. A revolution Two Independence Days And a clean slate. #2ndamendment #REMEMBERBENGHAZI2016 #PATRIOTSWILLRISE #SemST

  Also the following tweet in French entails some external *world knowledge* to be understood. In fact, we know that some French tabloids have been pushing insinuations on the sexual preferences of candidate Macron.

  R-FRA. Target: "Emmanuel Macron"

Emmanuel Macron : le candidat à la présidentielle préférédes gays
Emmanuel Macron : the candidate favored by the gays

- **Very short tweets** *(short text)*. It has been observed that many tweets that are very short pieces of texts have been misclassified in different languages. Sometimes tweets are composed only by Twitter Marks, URLs and mentions not even overcoming the length of 80 characters on the totality of the 140 available.

- **Noisy texts and incomplete sentences** *(noisy)*. We often observed also the presence of noisy texts in the misclassified tweets (mispellings, abbreviations, new words) and also, especially in the Italian case, the considerable frequency (27.58% of the misclassified posts) of tweets composed of incomplete sentences and characterized by ellipsis and unfinished thoughts followed by three dots, see for instance the following tweet:

  R-ITA. Target: "Constitutional Referendum"

  @beppe_grillo @Mov5Stelle #referendumcostituzionale #Renzi non serve aggiungere altro 4 milioni. di ragioni per...
  @beppe_grillo @Mov5Stelle #referendumcostituzionale #Renzi no need to add anything else 4 millions. of reasons to...

  In particular in the E-USA dataset we can observe an abundance of abbreviations. For instance, in the tweet below, the initials RWNJ stand for "right-wing nut job" but our system is not able to infer it without an extension of the real meaning of the abbreviation that the stance is 'against' conservatives.

  E-USA. Target: "Hillary Clinton"

  While I like Bernie as much as the next liberal, if we nominate him we could actually lose to some RWNJ #SemST

  In the same way, GOP stands for "Grand Old Party", common nickname for the Republican Party of the United States. Without this type of knowledge it is impossible for both humans and our system to detect the proper stance. Furthermore the pronoun *you* is abbreviated in *U* and the verb form *are* is shortened in *r*.

  E-USA. Target: "Donald Trump"

  @ChristieC733 YES DONALD U R 100% CORRECT' as long as u stay the coruse and don't pander to GOP U GOT MY VOTE #SemST

- **Hashtags and mentions included in the syntactic structure of the sentences** *(hashtag)*. We often observed the presence of hashtags included in the syntactic structure of the tweet's sentences. Hashtags are used by Twitter users for accomplishing different linguistic functions, enabling metadiscourse to be embedded in social media communication (Zappavigna, 2015). See for instance the following example for Italian:

  R-ITA. Target: "Constitutional Referendum"

  #referendumcostituzionale #sí o #no ? #Flick #DAlimonte ne parlano domattina 8,15 diretta #streaming...
  #referendumcostituzionale #yes or #no ? #Flick #DAlimonte will talk about it tomorrow morning at 8,15 on air #streaming...

  Tweets can be also linked to other tweeters through the use of at-mentions (e.g. @username). Like hashtags, also at-mentions can be a syntactic part of a sentence or phrase within a tweet, and especially in the English dataset with target 'Donald Trump', we may observe many misclassified cases where the mention *@realDonaldTrump* plays a precise syntactic role in the sentence, which is decisive to interpret the author's stance; see for instance:

  E-USA. Target: "Donald Trump"

  #presidentialelection2016 Make plans to help your future now, so that later you don't regret it, again! Vote @realDonaldTrump #SemST

  See also the following English tweet, where "You are an idiot" is referred to Donald Trump, who is tagged as a at-mention, but is also core part of the syntactic structure of the sentence:

  E-USA. Target: "Donald Trump"

  Dear @realDonaldTrump: You are an idiot. #america #politics #sticktoyourhair #SemST

- **Numericals and percentuage figures** (*numerical*). This feature characterizes a relevant number (22%) of misclassified tweets in the French debate:

  E-FRA. Target: "Emmanuel Macron"

  ...et #Hollande était faible (- de 40%) et #Macron est fort (80% et +) dans les circo bourgeoises de l'ouest parisi...
  ...and #Holland was weak (- of 40%) and #Macron is strong (80% and +) in the bourgeois circus of western Paris...

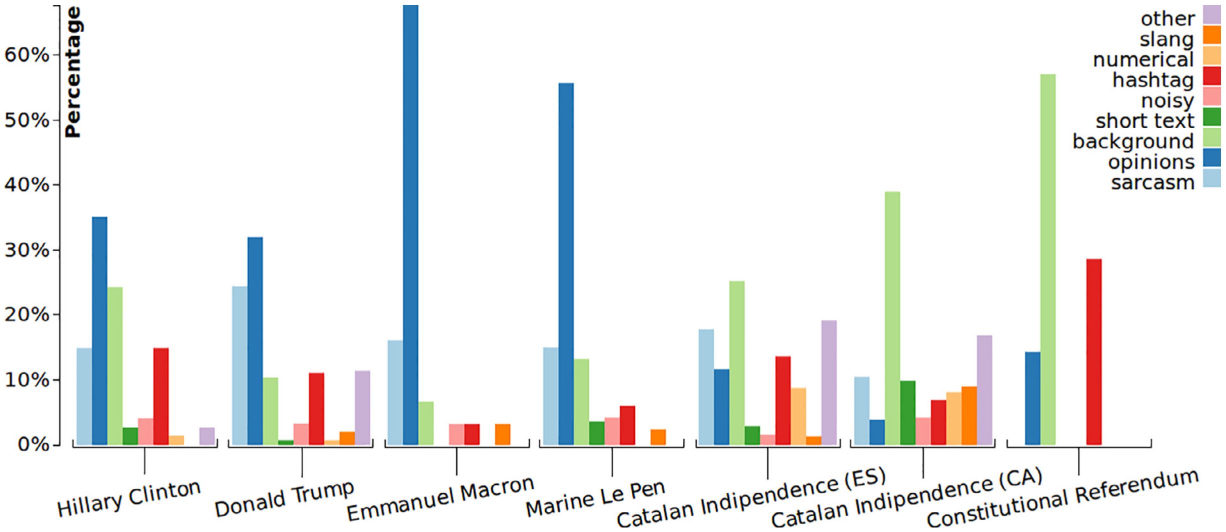


Fig. 1. Distribution of error types for each target.

E-FRA. Target: “Emmanuel Macron”

8 mai : 285K adhérents 10 mai : 310K (314K à 22h) Soutenant le projet de Macron et ses 577 candidats...
May, 8th : 285K subscribers May, 10th 310K (314K around 22h) Supporting Macron’s project and his 577 candidates...

- **Slang and slurs** (*slang*). Misclassified tweets often contain colloquial expressions, slurs, slang words, e.g. “You go get em Donald!”, “giving a big old s/o to Donald Trump”, “Kudos to Donald Trump” (E-USA, Donald Trump), “beur”, as typical in Verlan slang, which stands for a person of North African origin living in France (E-FRA, Le Pen).

In Fig. 1 we show the percentage of error types for each language and stance target in our debates. The percentages shown in this figure have been obtained by manual inspection of the tweets. As labels for the annotation, we used one label for each of the eight error classes described above. We also added an additional label *other*, to include all the unspecified cases.[xvii]

Observing Fig. 1 it is interesting to notice that the expression of opinions not related to the stance target is the most common cause of error in all the election datasets across the different languages (French and English). Indeed, in the French dataset (E-FRA), for both targets (Emmanuel Macron and Marine Le Pen), the highest ratio of errors is that labeled as *opinion* which stands for “Opinions expressed but related to different entities than the target of interest”. The same happens in the E-USA dataset for both targets (Donald Trump and Hillary Clinton), even though with a lower impact.

This is not so surprising, considering that sometimes in the tweet the target of interest is not explicitly mentioned. Or it could even happen that one of the two targets is insulted, and that this does not necessarily mean an opposite stance towards the political competitor.

Concerning the R*-CAT dataset, the two error classes with the higher ratio are *background* and *other* across the two languages. The first one refers to “Background knowledge and commonsense”, that is: in many cases users do not express their stance in an explicit manner. However, it could be inferred by human annotators by relying on *common sense knowledge* or *world knowledge*, a kind of information which is still hard to be fully captured. The label *other* refers to the presence of other type errors, such as for instance Catalan-Spanish code mixing, which is specifically featuring the data on Catalan Independence.

For what concerns the Italian dataset (R-ITA), the error analysis has been performed only on 7 tweets, due to the fact the best system reaches a really high F-score and the dataset is not very big. We report the values in Fig. 1 but they are not statistically significant.

Finally, among the noticeable findings, we can also observe a high rate of figurative devices in the misclassified tweets across all the languages and stance targets: *sarcasm* (the error class referring to “Sarcasm, metaphors, and other figurative language devices”) is indeed always among the top three error classes.

6. Conclusion

In the present work we investigated SD from a multilingual perspective focusing on datasets centered on four different political debates in five different languages: English, Spanish, Catalan, French and Italian. Two datasets were already available as benchmarks developed within the context of recent evaluation campaigns, i.e E-USA (developed for SemEval-2016 Task 6) and

[xvii] Notice that multiple error categories sometimes were selected because of the co-occurrence of difficulties that can be responsible for misclassification.

R*-CAT (developed for StanceCat at Ibereval 2017); the other two were created expressly for this study, i.e E-FRA and R-ITA. Among them, two datasets are about elections, i.e. E-USA and E-FRA, while the others two concern a referendum, i.e. R*-CAT and R-ITA.

Our main goal was to apply a machine learning system in a multilingual scenario in order to investigate the portability of SD techniques across different languages. This motivated the selection of the datasets which are featured by the similarity of domains (i.e. politics, electoral campaigns, and referendums). Nevertheless, providing that only a few resources annotated for stance currently exist, the side effect of this research can also be seen in the enlargement of the language scenario available to the community research working in this area. The new datasets annotated for stance (French and Italian) are available to the research community.[xviii]

We conducted several experiments using different classical machine learning methods and exploiting four groups of features: *Stylistic, Structural, Affective* and *Contextual* for testing the portability of these features across different languages and domains.

We observed that *Stylistic* features obtained fairly good results in all supervised contexts ("Hillary Clinton" for English, Spanish, Catalan and Italian) with the exception of the target "Emmanuel Macron" in French. Moreover, the specific *Stylistic* feature *BoC* obtains high results in all five languages independently from the target and the type of debate.

Also *Structural* features performed better in supervised contexts, especially thanks to features connected with Twitter Marks (hashtags and mentions) with which, users normally express their stance in a debate. On the other hand *Affective* and *Contextual* features are mostly exploited from models that are trained on a dataset in which the target for SD is not present (semi-supervised framework). Additionally, *Affective* features obtain higher scores when the target for SD is a person (in the case of election datasets: E-USA and E-FRA). Conversely, *Contextual* features are helpful when the target for SD is a referendum.

One of the most interesting finding of all the experimental settings is the evidence that *Contextual* language independent features perform well on the task of SD across language and domains. In particular *Community Retweet* and *Community Following*, are also influential independently from the type of the target of interest. Moreover, let us recall that the highest results have been obtained through the combination of content and contextual information. Thus, underlying the importance of merging methods from NLP and Network Science fields to improve SD.

To further extend the comparison with other approaches, we also considered deep learning methods and we implemented three different simple neural architectures (LSTM, biLSTM and CNN) by centering our focus on those ones that were recently exploited by state-of-the-art approaches on stance detection and were obtaining the most promising results. Even if our main focus in this work is on exploring the contribution that different typologies of features can give to the detection of stance across different languages and political domains, it is worth to be mentioned that over all languages and domains, our classical machine learning approaches proved to be competitive with respect to the neural models considered.

Additionally, from the rich experimental setting proposed – that explores both classical machine leaning approaches and neural models – we were able to learn a great lesson. Indeed, we were able to compare the performances of the three straightforward deep architectures we implemented (LSTM, biLSTM, and CNN) with the more elaborate and finer-grained based on classical classification algorithms (SVM and LR). From this comparison we have been able to verify that although neural models, with no need to engineer any kind of feature, prove to be strong in various scenarios, classical approaches dedicated to a specific task such as the method we propose in this research are effective and obtain competitive results.

In fact, our research has shed some light on the importance of different groups of features in a new task such as that of automatic SD, in relation to the complex domain of politics. Additionally, we performed a manual error analysis of the misclassified tweets across all languages and types of political debate. Beyond the suggestions for the further tuning and development of the SD classifier (MultiTACOS), several lessons can be learned by the error analysis and, in particular, some hints, which will inspire future development of our research. Among the future directions inspired by the error analysis, we seek to explore the contribution of the application of some form of syntactic analysis. See for instance the hints that can be extracted from the recent advancement a novel work by Sanguinetti et al. (2017), concerning the application of Universal Dependencies to social media texts. This research is a starting point for deeper investigation regarding the important role of the syntactic (and semantic) representation of mentions, hashtags and paratactical structures in social media texts. Furthermore, it will be interesting to extend the multilingual analysis also addressing new languages, for instance non-indoeuropean languages.

Acknowledgments

Cristina Bosco and Viviana Patti are partially supported by Progetto di Ateneo/CSP 2016 (*Immigrants, Hate and Prejudice in Social Media*, S1618_L2_BOSC_01). The work of Paolo Rosso was partially funded bythe Spanish MICINN under the research project MISMIS-FAKEnHATE on MISinformation and MIScommunication in social media: FAKE news and HATE speech (PGC2018-096212-B-C31).

References

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., Passonneau, R., 2011. Sentiment analysis of twitter data. In: Proceedings of the Workshop on Languages in Social Media. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 30–38.

[xviii] https://github.com/mirkolai/MultilingualStanceDetection

Augenstein, I., Rocktäschel, T., Vlachos, A., Bontcheva, K., 2016. Stance detection with bidirectional conditional encoding. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Austin, Texas, pp. 876–885. https://doi.org/10.18653/v1/D16-1084.

Balahur, A., Turchi, M., 2014. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. Comput. Speech Lang. 28 (1), 56–75.

Bethard, S., Cer, D.M., Carpuat, M., Jurgens, D., Nakov, P., Zesch, T., 2016. In: Proceedings of the 10Th International Workshop on Semantic Evaluation. The Association for Computer Linguistics.

Blondel, V.D., Guillaume, J.-L., Lambiotte, J.-L., Lefebvre, E., 2008. Fast unfolding of communities in large networks. J. Stat. Mech: Theory Exp. 10, 10008–10020. https://doi.org/10.1088/1742-5468/2008/10/P10008.

Boiy, E., Moens, M.-F., 2009. A machine learning approach to sentiment analysis in multilingual web texts. Inf. Retr. Boston 12 (5), 526–558.

Bosco, C., Lai, M., Patti, V., Rangel Pardo, F.M., Rosso, P., 2016. Tweeting in the Debate about Catalan Elections. In: Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (Eds.), LREC workshop on Emotion and Sentiment Analysis Workshop (ESA), LREC-2016. European Language Resources Association (ELRA), Portorož, Slovenia, pp. 67–70.

Bosco, C., Patti, V., 2017. Social media analysis for monitoring political sentiment. In: Alhajj, R., Rokne, J. (Eds.), Encyclopedia of Social Network Analysis and Mining. Springer, pp. 1–13.

Celli, F., Stepanov, E., Poesio, M., Riccardi, G., 2016. Predicting Brexit: classifying agreement is better than sentiment and pollsters. In: Proceedings of the Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media (PEOPLES). ACL Anthology, Osaka, Japan, pp. 110–118.

Cuquerella, C.A., Rodríguez, C.C., 2018. Crica team: multimodal stance detection in tweets on catalan 1oct referendum (multistancecat). 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). CEUR-WS.org, pp. 167–172.

Deitrick, W., Hu, W., 2013. Mutually enhancing community detection and sentiment analysis on twitter networks. J. Data Anal. Inf. Process. 1, 19–29.

Del Tredici, M., Marcheggiani, D., Schulte im Walde, S., Fernández, R., 2019. You shall know a user by the company it keeps: dynamic representations for social media users in NLP. 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). Association for Computational Linguistics, Hong Kong, China, pp. 4706–4716. https://doi.org/10.18653/v1/D19-1477.

DellaPosta, D., Shi, Y., Macy, M., 2015. Why do liberals drink lattes? Am. J. Sociol. 120 (5), 1473–1511.

Denecke, K., 2008. Using sentiwordnet for multilingual sentiment analysis. IEEE 24th International Conference on Data Engineering Workshop, 2008. ICDEW 2008. IEEE, pp. 507–512.

Dey, K., Shrivastava, R., Kaushik, S., 2018. Topical stance detection for twitter: a two-phase lstm model using attention. European Conference on Information Retrieval. Springer, pp. 529–536.

Hu, M., Liu, B., 2004. Mining and summarizing customer reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, Seattle, WA, USA, pp. 168–177.

Küçük, D., Can, F., 2019. A tweet dataset annotated for named entity recognition and stance detection. arXiv preprint arXiv:1901.04787. Available at: https://arxiv.org.

Lai, M., 2019. Language and Structure in Polarized Communities. Universitá degli Studi di Torino.

Lai, M., Cignarella, A.T., Hernandez Fariás, D.I., 2017. iTACOS at IberEval2017: detecting stance in Catalan and Spanish tweets. Notebook Papers of 2nd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL),CEUR Workshop Proceedings. CEUR-WS.org, 2017. CEUR-WS.org, Murcia, Spain, pp. 185–192.

Lai, M., Hernández Farías, D.I., Patti, V., Rosso, P., 2017. Friends and enemies of Clinton and trump: using context for detecting stance in political tweets. In: Sidorov, G., Herrera-Alcántara, O. (Eds.), Advances in Computational Intelligence: 15th Mexican International Conference on Artificial Intelligence, MICAI 2016, Cancún, Mexico, October 23−28, 2016, Proceedings, Part I. Springer International Publishing, Cham, pp. 155–168. https://doi.org/10.1007/978-3-319-62434-1_13.

Lai, M., Patti, V., Ruffo, G., Rosso, P., 2018. Stance evolution and Twitter interactions in an italian political debate. In: Silberztein, M., Atigui, F., Kornyshova, E., Métais, E., Meziane, F. (Eds.), Natural Language Processing and Information Systems. Springer International Publishing, Cham, Switzerland, pp. 15–27. https://doi.org/10.1007/978-3-319-91947-8_2.

Lai, M., Tambuscio, M., Patti, V., Ruffo, G., Rosso, P., 2017. Extracting graph topological information and users’ opinion. In: Jones, G.J., Lawless, S., Gonzalo, J., Kelly, L., Goeuriot, L., Mandl, T., Cappellato, L., Ferro, N. (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction: 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11−14, 2017, Proceedings. Springer International Publishing, Cham, pp. 112–118. https://doi.org/10.1007/978-3-319-65813-1_10.

Magdy, W., Darwish, K., Abokhodair, N., Rahimi, A., Baldwin, T., 2016. #isisisnotislam or #deportallmuslims?: predicting unspoken views. In: Proceedings of the 8th ACM Conference on Web Science. ACM, pp. 95–106.

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., Cherry, C., 2016. A dataset for detecting stance in tweets. In: Chair), N.C.C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (Eds.), Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Paris, France, pp. 3945–3952.

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). Association for Computational Linguistics, San Diego, California, pp. 31–41. https://doi.org/10.18653/v1/S16-1003.

Mohammad, S., Turney, P.D., 2013. Crowdsourcing a word-Emotion association lexicon. Comput. Intell. 29 (3), 436–465.

Mohammad, S.M., Sobhani, P., Kiritchenko, S., 2017. Stance and sentiment in tweets. ACM Trans. Internet Technol. 17 (3), 26:1–26:23. https://doi.org/10.1145/3003433.

Nakov, P., Rosenthal, S., Kozareva, Z., Stoyanov, V., Ritter, A., Wilson, T., 2013. SemEval-2013 task 2: sentiment analysis in twitter. Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013). Association for Computational Linguistics, Atlanta, Georgia, USA, pp. 312–320.

Nielsen, F.Å., 2011. A new ANEW: evaluation of a word list for sentiment analysis in Microblogs. In: Proceedings of the ESWC2011 Workshop on ’Making Sense of Microposts’: Big things come in small packages. CEUR-WS.org, Heraklion, Crete, Greece, pp. 93–98.

Pennebaker, J.W., Francis, M.E., Booth, R.J., 2001. Linguistic inquiry and word count: LIWC 2001. 71.

Raghavan, U.N., Albert, R., Kumara, S., 2007. Near linear time algorithm to detect community structures in large-scale networks. Phys. Rev. E 76 (3), 036106.

Rajadesingan, A., Liu, H., 2014. Identifying users with opposing opinions in twitter debates. International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction. Springer, pp. 153–160.

Respall, V.M., Derczynski, L., 2017. Stance Detection in Catalan and Spanish Tweets. Technical Report. Innopolis University.

Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S., Ritter, A., Stoyanov, V., 2015. SemEval-2015 task 10: sentiment analysis in Twitter. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Association for Computational Linguistics, pp. 451–463.

Sanguinetti, M., Bosco, C., Mazzei, A., Lavelli, A., Tamburini, F., 2017. Annotating Italian social media texts in universal dependencies. In: Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), pp. 229–239.

Schmid, H., 1994. Part-of-Speech Tagging with Neural Networks. In: Proceedings of the 15th conference on Computational linguistics-Volume 1. Association for Computational Linguistics, pp. 172–176.

Schmid, H., 1995. Treetagger| a language independent part-of-Speech tagger. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart 43, 28.

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). CEUR-WS.org, pp. 180–186.

Somasundaran, S., Wiebe, J., 2009. Recognizing stances in online debates. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 226–234.

Taulé, M., Martí, M.A., Pardo, F.M.R., Rosso, P., Bosco, C., Patti, V., 2017. Overview of the task on stance and gender detection in tweets on Catalan independence. In: Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017) co-located with 33th Conference of the Spanish Society for Natural Language Processing (SEPLN 2017). CEUR-WS.org, Murcia, Spain, pp. 157–177.

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: 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). CEUR-WS.org, pp. 149–166.

Tromp, E., Pechenizkiy, M., 2011. Senticorr: multilingual sentiment analysis of personal correspondence. 2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW). IEEE, pp. 1247–1250.

Tutek, M., Sekulić, I., Gombar, P., Paljak, I., Čulinović, F., Boltužić, F., Karan, M., Alagić, D., Šnajder, J., 2016. TakeLab at SemEval-2016 task 6: stance classification in tweets using a genetic algorithm based ensemble. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). Association for Computational Linguistics, San Diego, California, pp. 464–468. https://doi.org/10.18653/v1/S16-1075.

Vychegzhanin, S., Kotelnikov, E., 2019. Stance detection based on ensembles of classifiers. Program. Comput. Softw. 45 (5), 228–240.

Wei, P., Mao, W., Zeng, D., 2018. A target-guided neural memory model for stance detection in twitter. 2018 International Joint Conference on Neural Networks (IJCNN). IEEE, 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.

West, D.M., 1991. Polling effects in election campaigns. Polit. Behav. 13 (2), 151–163. https://doi.org/10.1007/BF00992294.

Whissell, C., 2009. Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural languages. Psychol. Rep. 2 (105), 509–521.

Wilson, T., Wiebe, J., Hoffmann, P., 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 347–354. https://doi.org/10.3115/1220575.1220619.

Xu, K., Li, J., Liao, S.S., 2011. Sentiment community detection in social networks. In: Proceedings of the 2011 iConference. Association for Computing Machinery, New York, NY, USA, pp. 804–805. https://doi.org/10.1145/1940761.1940913.

Yuan, J., Zhao, Y., Xu, J., Qin, B., 2019. Exploring answer stance detection with recurrent conditional attention. In: Proceedings of the AAAI Conference on Artificial Intelligence, 33, pp. 7426–7433.

Zappavigna, M., 2015. Searchable talk: the linguistic functions of hashtags. Soc. Semiot. 25 (3), 274–291. https://doi.org/10.1080/10350330.2014.996948.

Zarrella, G., Marsh, A., 2016. MITRE at SemEval-2016 Task 6: transfer learning for stance detection. In: Proceedings of the International Workshop on Semantic Evaluation. SemEval '16, San Diego, California, June.

Zhou, S., Lin, J., Tan, L., Liu, X., 2019. Condensed convolution neural network by attention over self-attention for stance detection in twitter. 2019 International Joint Conference on Neural Networks (IJCNN). IEEE, pp. 1–8.