SardiStance @ EVALITA2020: Overview of the Task on Stance Detection in Italian Tweets

Alessandra Teresa Cignarella[1,2], **Mirko Lai**[1], **Cristina Bosco**[1], **Viviana Patti**[1] **and Paolo Rosso**[2]
1. Dipartimento di Informatica, Università degli Studi di Torino, Italy
2. PRHLT Research Center, Universitat Politècnica de València, Spain
{lai,cigna,bosco,patti}@di.unito.it, prosso@dsic.upv.es

Abstract

English. *SardiStance* is the first shared task for Italian on the automatic classification of stance in tweets. It is articulated in two different settings: A) *Textual Stance Detection*, exploiting only the information provided by the tweet, and B) *Contextual Stance Detection*, with the addition of information on the tweet itself such as the number of retweets, the number of favours or the date of posting; contextual information about the author, such as follower count, location, user's biography; and additional knowledge extracted from the user's network of friends, followers, retweets, quotes and replies. The task has been one of the most participated at EVALITA 2020 (Basile et al., 2020), with a total of 22 submitted runs for Task A, and 13 for Task B, and 12 different participating teams from both academia and industry.

1 Introduction/Motivation

The interest towards detecting people's opinions towards particular targets, and towards monitoring politically polarized debates on Twitter has grown more and more in the last years, as it is attested by the proliferation of questionnaires and polls online (Küçük and Can, 2020). In fact, through the constant monitoring of people's opinion, desires, complaints and beliefs on political agenda or public services, policy makers could better meet population's needs.

In the fields of Natural Language Processing and Sentiment Analysis, this translates into the creation of a specifically dedicated task, namely: *Stance Detection* (SD), which is defined as the task of automatically determining from the text whether the author of a given textual content is in favor of, against, or neutral towards a certain target. Research on this topic, beyond mere academic interest, could have an impact on different aspects of everyday life such as public administration, policy-making, marketing or security strategies.

Although SD is a fairly recent research topic, considerable effort has been devoted to the creation of stance-annotated datasets. In their recent survey on this topic, Küçük and Can (2020) describe the existence of a variety of stance-annotated datasets (different text types such as tweets, posts in online forums, news articles, or news comments) for at least eleven languages.

The first shared task on SD was held for English at SemEval in 2016, i.e. *Task 6 "Detecting Stance in Tweets"* (Mohammad et al., 2016b) for detecting stance towards 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 proposed at *IberEval 2017* for both Catalan and Spanish (Taulé et al., 2017) where the target was only one, i.e. "Independence of Catalonia". A re-run was proposed the following year at the evaluation campaign *IberEval 2018* regarding the target "*Catalan first of October Referendum*" encouraging furthermore an exploration of multimodal expressions such as audio, videos and images (Taulé et al., 2018).

SardiStance@EVALITA2020 is the pioneer task for SD in Italian tweets. The motivation behind the proposal of this task is multi-faceted. On the one hand, we aimed at the creation of a new annotated dataset for SD in Italian which would enrich the panorama of available resources for this language, such as CONREF-STANCE-ITA (Lai et al., 2018)

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

and X-STANCE (Vamvas and Sennrich, 2020). On the other hand, the organization of this task allows us a deeper investigation of SD at a *contextual* level, by encouraging the participants and the research community to follow this research line that has proved promising in previous work, see e.g. Lai et al. (2019), Lai et al. (2020) and Del Tredici et al. (2019). In fact, with the data distributed in Task B different types of social network communities, based on friendships, retweets, quotes, and replies could be investigated, in order to analyze the communication among users with similar and divergent viewpoints.
The efficacy of approaches based on contextual features paired with textual information has been widely attested in literature on SD (Magdy et al., 2016; Rajadesingan and Liu, 2014) and additionally confirmed by the results obtained in this shared task, especially by those teams who participated to both settings (see Section 5).

2 Definition of the Task

With this task proposal, we wanted to invite participants to explore features based on the textual content of the tweet, such as structural, stylistic, and affective features, but also features based on contextual information that does not emerge directly from the text, such as knowledge about the domain of the political debate or information about the user's community. For these reasons, we proposed two different settings:

• Task A - Textual Stance Detection:

The first task was a three-class classification task where the system had to predict whether a tweet is in FAVOUR, AGAINST or NONE towards the given target, exploiting only textual information, i.e. the text of the tweet.

From reading the tweet, which of the options below is most likely to be true about the tweeter's stance towards the target? (Mohammad et al., 2016a)

1. **FAVOUR**: 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.

• Task B - Contextual Stance Detection:

The second task was the same as the first one: a three-class classification task where the system had to predict whether a tweet is in FAVOUR, AGAINST or NONE towards the given target. Here participants had access to a wider range of contextual information based on the **post** such as: the number of retweets, the number of favours, the number of replies and the number of quotes received to the tweet, the type of posting source (e.g. iOS or Android), and date of posting. Furthermore we shared (and encouraged its exploitation) contextual information related to the **user**, such as: number of tweets ever posted, user's bio, user's number of followers, user's number of friends. Additionally we shared users' contextual information about their **social network**, such as: friends, replies, retweets, and quotes' relations. The personal ids of the users were anonymized but their network structures were maintained intact.
Participants could decide to participate to both tasks or only to one. Although they were encouraged to participate to both.

3 Data

We chose to gather the data from the social networking Twitter due to the free availability of a huge amount of users' generated data and because it allowed us to explore different types of relations among the users involved in a debate.

3.1 Collection and annotation of the data

We collected around 700K tweets written in Italian about the "Movimento delle Sardine" (*Sardines movement*[1]), retrieving tweets containing the keywords "sardina", "sardine", and the homonymous hashtags. Furthermore, we collected all the conversation threads in which the said tweet belongs, iteratively following the reply's tree. We also collected the quoted tweets and the list of all the retweets of each previously recovered tweet, obtaining about 1M tweets. Finally, we collected the friend list of all the users included in the annotated dataset.

The tweets were gathered between the 46th week of 2019 (November) and the 5th week of 2020 (January), corresponding to a 12 weeks timewindow. Through the experience matured as participants in previous shared tasks of SD, and in or-

[1] https://en.wikipedia.org/wiki/Sardines_movement.

der to reduce noise in text, we collected data taking into account the following constraints: only one tweet per author for each week, no retweets, no replies, no quotes, no tweets containing URLs, no tweets containing pictures or videos.

Then, we included only Italian tweets posted using a limited number of "sources" (utilities used to post the tweet, such as iOS, Android, etc...) in order to avoid to include pre-written tweets posted using a *Tweet button*.[2] Furthermore, we validated that all the collected tweets presented a *Jaccard similarity coefficient* < 0.8. From about 25K filtered tweets, we finally randomly selected around 300 tweets for each week (only the first week of 2020 does not reach 300 tweets), thus obtaining 3,600 tweets in total.

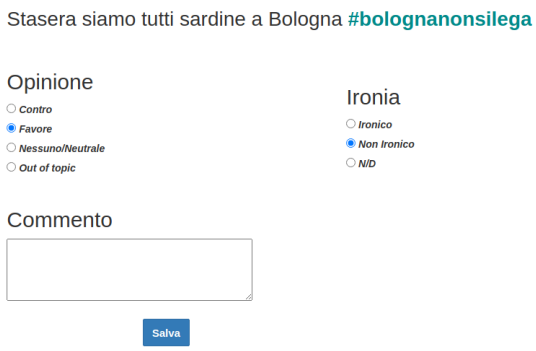

Figure 1: Platform for the annotation of tweets.

We created a web platform for annotation purposes, see Figure 1, in order to facilitate the labelling task to the annotators, unifying the visualization mode and shuffling the tweets in a random order.[3] 12 different native Italian speakers with an interest for news and politics were involved in the annotation, according to detailed guidelines we provided with examples for annotation and examples in their native language. We randomly shuffled the annotators and matched them into 66 pairs in which each pair would annotate 55 tweets. As a result, each annotator labelled 605 tweets independently and each tweet was annotated by two annotators, who had to choose among four different labels: AGAINST, FAVOUR, NONE/NEUTRAL and OUT OF TOPIC.

[2] https://developer.twitter.com/en/docs/twitter-for-websites/tweet-button/overview.

[3] In this way, each annotator was surely seeing emojis – which, we believe are essential in order to understand the correct stance– in the same way of the other annotators independently of the device used.

Furthermore, as it can also be seen in Figure 1 (*Tonight we are all sardines in Bologna #bolognanonsilega*), we asked the annotators to mark whether, in their opinion, the tweet was IRONIC or NOT IRONIC. Finally, we were not able to obtain satisfactory results on this end, so we did not include it in the task.

3.2 Analysis of the annotation

At the end of a first phase of annotation, which lasted more or less a month, we obtained 2,256 tweets in agreement, with a clear decision on one of the three main classes. Other 917 tweets presented a *light disagreement* (i.e. FAVOUR vs. NEUTRAL or AGAINST vs. NEUTRAL), and the remaining 457 tweets were discarded because the majority of annotators considered them out of topic or were in *strong disagreement* (i.e. FAVOUR vs. OUT OF TOPIC).

We then proceeded in the resolution of those 917 tweets, whose disagreement was deemed "light" in order to obtain a bigger dataset. We resorted once again to the annotation platform used in the first phase, we revised the annotation guidelines and asked the annotators to label the tweets again. In this phase, we paid attention that the tweets in disagreement were not assigned to the same pair of annotators that had previously labelled them, and furthermore we chose to show the two annotations in contrast, along with any comment - if present - to the annotator that had to solve the disagreement.

After the second phase, we computed the inter-annotator agreement (IAA) through Cohen's kappa coefficient (over the three main classes) resulting in $\kappa = 0.493$ (weak agreement). The same coefficient was also used to compute the IAA among annotators over the two most significant classes (AGAINST and FAVOUR, excluding the NEUTRAL class), resulting in a higher score: $\kappa = 0.769$ (moderate agreement). Notably, we observed that the IAA significantly changes depending on the observed pair of annotators (it ranges from 0.873 to 0.473) in the first phase of the annotation. We also noticed that the average IAA, computed through the sum of each IAA between any annotator and the remaining 11 annotators, can significantly change (ranging from 0.704 to 0.609). In other words, some annotators tend to strongly agree with all the other ones, while others tend to disagree with the majority. As future work,

we aim to shed more light on this phenomena exploring the background of the annotators and the social relationship among them.

3.3 Composition of the dataset

After the second round of annotation we were finally able to create the official dataset for the *SardiStance* shared task. It is composed by a total of 3,242 tweets, 1,770 of which belong to the class AGAINST, 785 to the class FAVOUR, and 687 to the class NONE. In Table 1 we show the distribution of such instances accordingly to the training set and the test set and in Table 2 we report tweet as example for each class.

TRAINING SET			TEST SET		
AGAINST	FAVOUR	NONE	AGAINST	FAVOUR	NONE
1,028	589	515	742	196	172
2,132			**1,110**		

Table 1: Distribution of tweets.

text	label
LE SARDINE IN PIAZZA MAGGIORE NON SONO ITALIANI SE LO FOSSERO NON SI METTEREBBERO CONTRO LA DESTRA CHE AMA L'ITALIA E VUOLE RIMANERE ITALIANA *THE SARDINES IN PIAZZA MAGGIORE ARE NOT ITALIAN IF THEY WERE THEY WOULD NOT GO AGAINST THE RIGHT THAT LOVES ITALY AND WANTS TO REMAIN ITALIAN*	AGAINST
Non ci credo che stasera devo andare in teatro e non posso essere fra le #Sardine #Bologna #bolognanonsilega *I can't believe that I have to go to the theater tonight and I can't be among the #Sardines #Bologna #bolognanonsilega*	FAVOUR
Mi sono svegliato nudo e triste perché a Bologna, tra salviniani e antisalviniani, non mi ha cagato nessuno. *I woke up naked and sad because in Bologna, between Salvinians and anti-Salvinians, nobody paid me attention.*	NONE

Table 2: Examples from the dataset.

3.4 Data Release

We shared data following the methodology recommended in (Rangel and Rosso, 2018) in order to comply to GDPR privacy rules and Twitter's policies. The identifiers of tweets and users have been anonymized and replaced by unique identifiers. We exclusively released the emojis eventually contained in the location and description user's biography, in order to make very hard to trace users and to preserve everybody's privacy.

Task A

The training data (TRAIN.csv) was released in the following format:

```
tweet_id  user_id  text  label
```

where `tweet_id` is the Twitter ID of the message, `user_id` is the Twitter ID of the user who posted the message, `text` is the content of the message, `label` is AGAINST, FAVOUR or NONE.

Task B

In order to participate to Task B, we released additional contextual information.

• the file TWEET.csv, containing contextual information regarding the tweet, with the following format:

```
tweet_id  user_id  retweet_count
favorite_count  source  created_at
```

where `tweet_id` is the Twitter ID of the message, `user_id` is the Twitter ID of the user who posted the message, `retweet_count` indicates the number of times the tweet has been retweeted, `favorite_count` indicates the number of times the tweet has been liked, `source` indicates the type of posting source (e.g. iOS or Android), and `created_at` displays the time of creation according to a yyyy-mm-dd hh:mm:ss format. Minutes and seconds have been encrypted and transformed to zeroes for privacy issues.

• the file USER.csv, containing contextual information regarding the user. It was released in the following format:

```
user_id  statuses_count  friends_count
followers_count  created_at  emoji
```

where `user_id` is the Twitter ID of the user who posted the message, `statuses_count`, `friends_count` indicates the number of friends of the user, `followers_count` indicates the number of followers of the user, `created_at` displays the time of the user registration on Twitter, and `emoji` shows a list of the emojis in the user's bio (if present, otherwise the field is left empty).

• The files FRIEND.csv, QUOTE.csv, REPLY.csv and RETWEET.csv containing contextual info about the social network of the user. Each file was released in the following format:

```
Source  Target  Weight
```

where `Source` and `Target` indicate two nodes of a social interaction between two Twitter users. More specifically, the source user performs one of the considered social relation towards the target user. Two users are tied by a friend relationship if the source user follows the target user (friend relationship does not have a weight, because it is either present or absent); while two users are tied by a quote, retweet, or reply relationship if the source user respectively quoted, retweeted, or replied the target user. Table 4 shows some metrics about the shared networks.

	nodes	**edges**
friend	669,817	3,076,281
retweet	110,315	575,460
quote	2,903	7,899
reply	14,268	29,939

Table 4: Networks metrics.

`Weight` indicates the number of interactions existing between two users. Note that this information is not available for the friend relation (hence, this column was not present in the FRIEND.csv file) due to the fact that it is a relationship of the type present/absent and cannot be described through a weight. In all the files, users are defined by their anonimyzed User ID.

Regrettably, we did not think to anonymize the screen names contained in the text of the tweets (with the same numeric string used to anonymize users), for allowing to match it with the users' ids and allowing the exploration of the network based on mentions. We will surely take it into account in our future works.

4 Evaluation Measures

Each participating team was allowed to submit a maximum of 4 runs for each sub-task: two constrained runs and two unconstrained runs. Submitting at least a constrained run was anyway compulsory. We decided to provide two separate official rankings for Task A and Task B, and two separate ranking for constrained and unconstrained runs. Systems have been evaluated using F1-score computed over the two main classes (FAVOUR and AGAINST). Therefore, the submissions have been ranked by the averaged F1-score over the two classes, according the following equation: $F1_{avg} = (F1_{favour} + F1_{against})/2$.

4.1 Baselines

We computed a baseline using a simple machine learning model, for Task A: a Support Vector Classifier based on token uni-gram features. A second baseline we computed for Task B is a system based on our previous work on Stance Detection: a Logistic Regression classifier paired with token n-grams features (unigrams, bigrams and trigrams), plus features based on a binary one-hot encoding representation of the communities extracted from the network of retweets and the network of friends (see the best system for Italian, in Lai et al. (2020)).

5 Participants and results

A total of 12 teams, both from academia and industry sector participated to at least one of the two tasks of SardiStance. In Table 3 we provide an overview of the teams in alphabetical order.

Teams were allowed to submit up to four runs (2 constrained and 2 unconstrained) in case they implemented different systems. Furthermore, each team had to submit at least a constrained run. Participants have been invited to submit multiple runs to experiment with different models and architectures. However, they have been discouraged from

team name	**institution**	**report**	**task**
deepreading	UNED, Spain	(Espinosa et al., 2020)	A, B
GhostWriter	You Are My Guide, Italy	(Bennici, 2020)	A, B
IXA	UPV/EHU, Spain	(Espinosa et al., 2020)	A, B
MeSoVe	ISASI, Italy	-	A
QMUL-SDS	QMUL-SDS-EECS, UK	(Alkhalifa and Zubiaga, 2020)	A, B
SSN_NLP	CSE Department/SSNCE, India	(Kayalvizhi et al., 2020)	A
SSNCSE-NLP	SSN College of Engineering, India	(Bharathi et al., 2020)	A, B
TextWiller	UNIPD, Italy	(Ferraccioli et al., 2020)	A, B
UNED	UPV/EHU and UNED, Spain	(Espinosa et al., 2020)	B
UninaStudents	UNINA, Italy	(Moraca et al., 2020)	A
UNITOR	UNIROMA2, Italy	(Giorgioni et al., 2020)	A
Venses	UNIVE, Italy	(Delmonte, 2020)	A

Table 3: Participants and reports.

submitting slight variations of the same model. Overall we have 22 runs for Task A and 13 runs for Task B.

5.1 Task A: Textual Stance Detection

Table 5 shows the results for the textual stance detection task, which attracted 22 total submissions from 11 different teams. Since the only two systems in an unconstrained setting were submitted by the same team we decided not to create a separate ranking for them, but rather to include them in the same ranking, and marking them with a different color (gray in Table 5).

team name	run	F1-score			
		AVG	AGAINST	FAVOUR	NONE
UNITOR	1	**.6853**	.7866	**.5840**	.3910
UNITOR	1	**.6801**	.7881	**.5721**	.3979
UNITOR	2	.6793	**.7939**	.5647	.3672
DeepReading	1	.6621	.7580	.5663	**.4213**
UNITOR	2	.6606	.7689	.5522	.3702
IXA	1	.6473	.7616	.5330	.3888
GhostWriter	1	.6257	.7502	.5012	.3810
IXA	2	.6171	.7543	.4800	.3675
SSNCSE-NLP	2	.6067	.7723	.4412	.2113
DeepReading	2	.6004	.6966	.5042	.3916
GhostWriter	2	.6004	.7224	.4784	.3778
UninaStudents	1	.5886	.7850	.3922	.2326
baseline		*.5784*	*.7158*	*.4409*	*.2764*
TextWiller	1	.5773	.7755	.3791	.1849
SSNCSE-NLP	1	.5749	.7307	.4192	.3388
QMUL-SDS	1	.5595	.7091	.4099	.2313
QMUL-SDS	2	.5329	.6478	.4181	.3049
MeSoVe	1	.4989	.7336	.2642	.3118
TextWiller	2	.4715	.6713	.2718	.2884
SSN_NLP	1	.4707	.5763	.3651	.3364
SSN_NLP	2	.4473	.6545	.2402	.1913
Venses	1	.3882	.5325	.2438	.2022
Venses	2	.3637	.4564	.2710	.2387

Table 5: Results Task A.

The best results are achieved by the UNITOR team that, with an unconstrained, ranked as 1st position with $F1_{avg}$ = 0.6853. The best result for the constrained runs is achieved once again by the UNITOR team with $F1_{avg}$ = 0.6801.

The best results for the two main classes AGAINST and FAVOR are obtained by the three best systems of the ranking, which are all submissions by the team UNITOR. On the other hand, though, the Deepreading team, ranking as 4th, has obtained the best F1-score for the NONE class, with $F1_{none}$ = 0.4213.

Among the 12 participating teams, at least 6 show an improvement over the baseline, which was computed using an SVM paired with token unigrams as unique feature, resulting an already strong result to beat ($F1_{avg}$ = 0.5784).

5.2 Task B: Contextual Stance Detection

Table 6 shows the results for the contextual stance detection task, which attracted 13 total submissions from 7 different teams.

team name	run	F1-score			
		AVG	AGAINST	FAVOUR	NONE
IXA	3	**.7445**	**.8562**	**.6329**	.4214
TextWiller	1	.7309	.8505	.6114	.2963
DeepReading	1	.7230	.8368	.6093	.3364
DeepReading	2	.7222	.8300	.6143	**.4251**
TextWiller	2	.7147	.8298	.5995	.3680
QMUL-SDS	1	.7088	.8267	.5908	.1811
UNED	2	.6888	.8175	.5600	.2455
QMUL-SDS	2	.6765	.8134	.5396	.1553
SSNCSE-NLP	2	.6582	.7915	.5249	.3691
SSNCSE-NLP	1	.6556	.7914	.5198	.3880
baseline		*.6284*	*.7672*	*.4895*	*.3009*
GhostWriter	1	.6257	.7502	.5012	.3810
GhostWriter	2	.6004	.7224	.4784	.3778
UNED	1	.5313	.7399	.3226	.2000

Table 6: Results Task B.

The best scores are achieved by the IXA team that with a constrained run obtained the highest score of $F1_{avg}$ = 0.7445. The best F1-score for the main classes AGAINST and FAVOUR is achieved by the team ranked 1st, IXA, team with $F1_{against}$ = 0.8562, and $F1_{favour}$ = 0.6329, respectively. Once again, the Deepreading team, ranking 3rd and 4th, has obtained the best F1-score for the NONE class, with $F1_{none}$ = 0.4251.

Almost all participating systems show an improvement over the baseline, which was computed using a Logistic Regression classifier paired with token n-grams features (unigrams, bigrams and trigrams), features based on the network of retweets, and features based on the network of friends (Lai et al., 2020).

6 Discussion

In this section we compare the participating systems according to the following main dimensions: system architecture, features, use of additional annotated data for training, and use of external resources (e.g. sentiment lexica, NLP tools, etc.). We also operate a distinction between runs submitted in Task A and those submitted in Task B. This discussion is based on the participants' reports and the answers the participants provided to a questionnaire proposed by the organizers. Two teams, namely TextWiller and Venses wrote a

joint report, overlapping between this task and the *HaSpeeDe 2* task (Sanguinetti et al., 2020), as they participated in both competitions. The three following teams, Deepreading, IXA, and UNED, also wrote a unique report as the participants, belong to the same research project and wanted to compare their three different approaches.

6.1 Systems participating to Task A

System architecture. Among all submitted runs we counted a great variety of architectures, ranging from classical machine learning classifiers, to recent state-of-the-art approaches, and statistically-based models. For instance, regarding the use of classical ML, the team UninaStudents used a SVM, and the team MeSoVe used Logistic Regression in one run. Regarding the use of neural networks, the QMUL-SDS team used bidirectional-LSTM, a CNN-2D, and a bi-LSTM with attention. Also SSN_NLP exploited the LSTM neural network.

Four teams exploited different variants of the BERT model: Ghostwriter used AlBERTo trained on Italian tweets, IXA used GilBERTo and UmBERTo[4], while UNITOR adopted only this latter model. Finally the Deepreading team made use of transformers such as BERT XXL and XML-RoBERTa, paired together with linear classifiers. TextWiller is the only team to have exploited the *xg-boost* algorithm, and ItVenses relied on supervised models, based on statistics and semantics. The UNED team proposed instead a voting system among the output of different models.

Features. Besides having explored a variety of system architectures, the teams participating in Task A, also used many different textual features, in the most of cases based on n-grams or char-grams. MeSoVe and TextWiller additionally engineered features based on emoticons. The team UNED, in one of their runs, proposed a system relying on psychologycal and social features, while UninaStudents proposed features of uni-grams of hashtags. Interestingly, UNITOR added special tags to the texts, which are the result of a classification with respect some so-called "auxiliary task". In particular, they trained three classifiers based respectively on SENTIPOLC 2016 (Barbieri et al., 2016) for sentiment analysis classification, on HaSpeeDe 2018 (Bosco et al., 2018) for hate speech detection, and on IronITA 2018 (Cignarella et al., 2018) for irony detection; and they added three tags to each instance of the SardiStance datasets with respect to these three dimensions: sentiment, hate and irony. ItVenses proposed features collected automatically from a unique dictionary list, frequency of occurrence of emojis and emoticons, and semantic features investigating propositional level, factivity and speech act type.

Additional training data. The only team who participated to the unconstrained setting of *SardiStance* is UNITOR. They proposed two unconstrained runs in addition to other two constrained ones. For the unconstrained setting, they downloaded and labeled about 3,200 tweets using distant supervision and used the additional data to train their systems. In particular they created the following subsets:
- 1,500 AGAINST: tweets from 2019 containing the hashtag: #gatticonsalvini;
- 1,000 FAVOUR: tweets from 2019 containing the hashtags: #nessunotocchilesardine, #iostoconlesardine, #unmaredisardine, #vivalesardine and #forzasardine;
- 700 NONE/NEUTRAL: texts derived from news titles. These were retrieved by querying to Google news with the keyword "sardine".

Other resources. Five teams declared to have used also other resources such as lexica, word embeddings, or others. In particular, GhostWriter used grammar model to rephrase the tweets. MeSoVe exploited SenticNet (Cambria et al., 2014) and the "Nuovo vocabolario di base della lingua italiana".[5] QMUL-SDS took advantage of temporal embeddigns and FastText, while only one team, UninaStudents, used a sentiment lexicon: AFINN (Nielsen, 2011). Lastly, Venses used a proprietary lexicon of Italian, enriched with conceptual, semantic and syntactic information; and similarly TextWiller approach relies on a self-created vocabulary and trained word-embeddigs on the corpus PAISÀ (Lyding et al., 2014).

6.2 Systems participating to Task B

Seven teams participated in Task B submitting a total of 13 runs. Most teams extensively explored the additional features available for Task B; GhostWriter, on the contrary, proposes the same

[4] https://huggingface.co/Musixmatch/umberto-commoncrawl-cased-v1.

[5] https://dizionario.internazionale.it.

two approaches presented in Task A. Notably, the three runs with a score lower than the baseline do not have benefited from any features based on the users' social network.

System architecture. Most teams enriched the models they submitted in Task A by taking advantage of contextual information available in Task B. UNED, DeepReading, and TextWiller exploited the *xg-boost* algorithm selecting different features from contextual data. The language model BERT was used in different variants by SSNCSE-NLP, DeepReading, and IXA. In particular, the last two teams proposed three voting based ensemble methods that use two or more models that exploit the *xg-boost* algorithm. Furthermore, the neural network framework proposed by QMUL-SDS exploits and combine four different embedding methods into a dense layer for generating the final label using a *softmax* activation function.

Features. Not every team took full advantage of contextual information. For example, SSNCSE-NLP only exploits the number of friends in run 1, and the number of quotes and friends in run 2. In its run 1 UNED also exploited some features based on the tweets in addition to the psychological and emotional ones, using the *xg-boost* algorithm. The other teams exploited different approaches for learning vector representations of the nodes of the available networks. DeepReading, IXA, and UNED proposed a feature that computes the mean distances of each user to the rest of users whose stance is known. TextWiller experimented a multi-dimensional scaling (MDS) for retaining the first and second dimension for each of the four networks instated. *Node2vec* and *deepwalk* for learning a vector representation of the nodes of the networks were used respectively in QMUL-SDS's runs 1 and 2.

The comparison between the approaches respectively used for dealing with Task A and Task B, clearly highlights the benefits of exploiting information from different and heterogeneous sources. In particular, it is interesting to observe that all the teams that participated to both tasks, also produced better results in the second setting. Experimenting with different classifiers trained with the textual content of the tweets as well as with features based on contextual information (additional info on the tweets, on users, or their social networks) seems therefore to allow to obtain overall better results.

In particular, among the 6 teams that participated to both tasks, only 4 fully explored the social network relations of the author of the tweet. The only two runs that overcome the baseline without investigating the structures of the social graphs are those submitted by the SSNCSE-NLP team. Only one team participated to both tasks exploiting the same architecture. This, allowed us to compare the F1-scores obtained in the first setting with those obtained in the second, highlighting that adding contextual features could increase performance of +0.2432, in terms of $F1_{avg}$.

Additionally, we calculated the increment in performance between the score obtained by the run ranked as 1st position in Task A (UNITOR, F_{avg} = 0.6853) and the score of the run ranked as 1st position in Task B (IXA, F_{avg} = 0.7445), showing that taking advantage of contextual features could increase performance up to 8,6% in terms of $F1_{avg}$.

7 Conclusions

We presented the first shared task on Stance Detection for Italian, discussing the development of the datasets used and the participation. A great panel for discussions about techniques and state-of-the-art approaches has been opened which can be used for investigating future research directions.

Acknowledgments

The work of C. Bosco, M. Lai and V. Patti is partially funded by the project "Be Positive!" (under the 2019 "Google.org Impact Challenge on Safety" call). The work of C. Bosco and V. Patti is also partially funded by Progetto di Ateneo/CSP 2016 *Immigrants, Hate and Prejudice in Social Media* (S1618_L2_BOSC_01). The work of P. Rosso is partially funded by the Spanish MICINN under the research projects MISMIS-FAKEnHATE on Misinformation and Miscommunication in social media: FAKE news and HATE speech (PGC2018-096212-B-C31) and PROMETEO/2019/121 (DeepPattern) of the Generalitat Valenciana.

A special mention also to the people who helped us with the annotation of the dataset. In random order: Matteo, Luca, Ylenia, Simona, Elisa, Sebastiano, Francesca, Simona, Komal and Angela, thank you very much for your great help.

References

Rabab Alkhalifa and Arkaitz Zubiaga. 2020. QMUL-SDS @ SardiStance: Leveraging Network Interactions to Boost Performance on Stance Detection using Knowledge Graphs. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*. CEUR-WS.org.

Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. Overview of the EVALITA 2016 SENTIment POLarity Classification task. In *Proceedings of the 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2016)*. CEUR-WS.org.

Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro. 2020. EVALITA 2020: Overview of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR-WS.org.

Mauro Bennici. 2020. ghostwriter19 @ SardiStance: Generating new tweets to classify SardiStance EVALITA 2020 political tweets. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*. CEUR-WS.org.

B. Bharathi, J. Bhuvana, and Nitin Nikamanth Appiah Balaji. 2020. SardiStance@EVALITA2020: Textual and Contextual stance detection from Tweets using machine learning approach. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*. CEUR-WS.org.

Cristina Bosco, Felice Dell'Orletta, Fabio Poletto, Manuela Sanguinetti, and Maurizio Tesconi. 2018. Overview of the Evalita 2018 Hate Speech Detection Task. In *Proceedings of 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018)*. CEUR-WS.org.

Erik Cambria, Daniel Olsher, and Dheeraj Rajagopal. 2014. SenticNet 3: a Common and Commonsense Knowledge Base for Cognition-driven Sentiment Analysis. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI 2014)*.

Alessandra Teresa Cignarella, Simona Frenda, Valerio Basile, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2018. Overview of the EVALITA 2018 task on Irony Detection in Italian Tweets (IronITA). In *Proceedings of 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018)*. CEUR-WS.org.

Marco Del Tredici, Diego Marcheggiani, Sabine Schulte im Walde, and Raquel Fernández. 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 2019)*. ACL.

Rodolfo Delmonte. 2020. Venses @ HaSpeeDe2 & SardiStance: Multilevel Deep Linguistically Based Supervised Approach to Classification. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*. CEUR-WS.org.

Maria S. Espinosa, Rodrigo Agerri, Alvaro Rodrigo, and Roberto Centeno. 2020. DeepReading @ SardiStance: Combining Textual, Social and Emotional Features. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*. CEUR-WS.org.

Federico Ferraccioli, Andrea Sciandra, Mattia Da Pont, Paolo Girardi, Dario Solari, and Livio Finos. 2020. TextWiller @ SardiStance, HaSpeede2: Text or Con-text? A smart use of social network data in predicting polarization. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*. CEUR-WS.org.

Simone Giorgioni, Marcello Politi, Samir Salman, Danilo Croce, and Roberto Basili. 2020. UNITOR@Sardistance2020: Combining Transformer-based architectures and Transfer Learning for robust Stance Detection. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*. CEUR-WS.org.

S. Kayalvizhi, D. Thenmozhi, and Chandrabose Aravindan. 2020. SSN_NLP@SardiStance : Stance Detection from Italian Tweets using RNN and Transformers. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*. CEUR-WS.org.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys*, 53(1):1–37.

Mirko Lai, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. 2018. Stance evolution and Twitter interactions in an Italian political debate. In *Proceedings of the 23rd International Conference on Natural Language & Information Systems (NLDB 2018)*. Springer.

Mirko Lai, Marcella Tambuscio, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. 2019. Stance polarity in political debates: A diachronic perspective of network homophily and conversations on twitter. *Data & Knowledge Engineering*, 124:101738.

Mirko Lai, Alessandra Teresa Cignarella, Delia Irazú Hernández Farías, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2020. Multilingual stance detection in social media political debates. *Computer Speech & Language*, 63(101075).

Verena Lyding, Egon Stemle, Claudia Borghetti, Marco Brunello, Sara Castagnoli, Felice Dell'Orletta, Henrik Dittmann, Alessandro Lenci, and Vito Pirrelli. 2014. The PAISA' Corpus of Italian Web Texts. In *Proceedings of the 9th World Archaeological Congress (WAC-9) @ the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*. ACL.

Walid Magdy, Kareem Darwish, Norah Abokhodair, Afshin Rahimi, and Timothy Baldwin. 2016. #isisisnotislam or #deportallmuslims?: Predicting unspoken views. In *Proceedings of the 8th ACM Conference on Web Science (WebSci 2016)*. ACM.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016a. A Dataset for Detecting Stance in Tweets. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. ELRA.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016b. SemEval-2016 Task 6: Detecting Stance in Tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. ACL.

Maurizio Moraca, Gianluca Sabella, and Simone Morra. 2020. UninaStudents @ SardiStance: Stance detection in Italian tweets - Task A. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*. CEUR-WS.org.

Finn Årup Nielsen. 2011. AFINN. *Richard Petersens Plads, Building*, 321.

Ashwin Rajadesingan and Huan Liu. 2014. Identifying users with opposing opinions in Twitter debates. In *Proceedings of the 7th Social Computing, Behavioral-Cultural Modeling and Prediction International Conference (SBP-BRiMS 2014)*. Springer.

Francisco Rangel and Paolo Rosso. 2018. On the implications of the general data protection regulation on the organisation of evaluation tasks. *Language and Law / Linguagem e Direito*, 5(2):95–117.

Manuela Sanguinetti, Gloria Comandini, Elisa Di Nuovo, Simona Frenda, Marco Stranisci, Cristina Bosco, Tommaso Caselli, Viviana Patti, and Irene Russo. 2020. HaSpeeDe 2@EVALITA2020: Overview of the EVALITA 2020 Hate Speech Detection Task. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*. CEUR-WS.org.

Mariona Taulé, M. Antònia Martí, Francisco M. Rangel Pardo, Paolo Rosso, Cristina Bosco, and Viviana Patti. 2017. Overview of the Task on Stance and Gender Detection in Tweets on Catalan Independence. In *Proceedings of the 2nd 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.

Mariona Taulé, Francisco M. Rangel Pardo, M. Antònia Martí, and Paolo Rosso. 2018. Overview of the Task on Multimodal Stance Detection in Tweets on Catalan #1Oct Referendum. In *Proceedings of the 3rd 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.

Jannis Vamvas and Rico Sennrich. 2020. X-Stance: A Multilingual Multi-Target Dataset for Stance Detection. In *Proceedings of the 5th Swiss Text Analytics Conference (SwissText 2020) & 16th Conference on Natural Language Processing (KONVENS 2020)*. CEUR-WS.org.