SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter

Valerio Basile[♢] **Cristina Bosco**[♢] **Elisabetta Fersini**[♡]
Debora Nozza[♡] **Viviana Patti**[♢] **Francisco Rangel**[♠♣]
Paolo Rosso[♣] **Manuela Sanguinetti**[♢]
[♢] Dipartimento di Informatica, Università degli Studi di Torino (Italy)
[♣] Università degli Studi di Milano Bicocca (Italy)
[♠] Autoritas Consulting (Spain)
[♡] PRHLT Research Center, Universitat Politècnica de València (Spain)
[♢]{name.surname}@unito.it, [♡]{name.surname}@unimib.it,
[♠]francisco.rangel@autoritas.es,
[♣]prosso@dsic.upv.es

Abstract

The paper describes the organization of the SemEval 2019 Task 5 about the detection of hate speech against immigrants and women in Spanish and English messages extracted from Twitter. The task is organized in two related classification subtasks: a main binary subtask for detecting the presence of hate speech, and a finer-grained one devoted to identifying further features in hateful contents such as the aggressive attitude and the target harassed, to distinguish if the incitement is against an individual rather than a group. HatEval has been one of the most popular tasks in SemEval-2019 with a total of 108 submitted runs for Subtask A and 70 runs for Subtask B, from a total of 74 different teams. Data provided for the task are described by showing how they have been collected and annotated. Moreover, the paper provides an analysis and discussion about the participant systems and the results they achieved in both subtasks.

1 Introduction

Hate Speech (HS) is commonly defined as any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics (Nockleby, 2000). Given the huge amount of user-generated contents on the Web, and in particular on social media, the problem of detecting, and therefore possibly contrasting the HS diffusion, is becoming fundamental, for instance for fighting against misogyny and xenophobia.

Some key aspects feature online HS, such as virality, or presumed anonymity, which distinguish it from offline communication and make it potentially also more dangerous and hurtful. Often hate speech fosters discrimination against particular categories and undermines equality, an everlasting issue for each civil society. Among the mainly targeted categories there are immigrants and women. For the first target, especially raised by refugee crisis and political changes occurred in the last few years, several governments and policy makers are currently trying to address it, making especially interesting the development of tools for the identification and monitoring such kind of hate (Bosco et al., 2017). For the second one instead, hate against the female gender is a long-time and well-known form of discrimination (Manne, 2017). Both these forms of hate content impact on the development of society and may be confronted by developing tools that automatically detect them.

A large number of academic events and shared tasks for different languages (i.e. English, Spanish, Italian, German, Mexican-Spanish, Hindi) took place in the very recent past which are centered on HS and related topics, thus reflecting the interest by the NLP community. Let us mention the first and second edition of the *Workshop on Abusive Language*[1] (Waseem et al., 2017), the *First Workshop on Trolling, Aggression and Cyberbullying* (Kumar et al., 2018), that also included a shared task on aggression identification, the tracks on *Automatic Misogyny Identification* (AMI) (Fersini et al., 2018b) and on *Authorship and Aggressiveness Analysis* (MEX-A3T) (Carmona et al., 2018) proposed at the 2018 edition of IberEval[2], the GermEval Shared Task on the *Identification of Offensive Language* (Wiegand et al.,

[1] http://sites.google.com/view/alw2018/
[2] http://sites.google.com/view/ibereval-2018

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 54–63
Minneapolis, Minnesota, USA, June 6–7, 2019.

2018), and finally the *Automatic Misogyny Identification task* (AMI) (Fersini et al., 2018a) and the *Hate Speech Detection task* (HaSpeeDe) (Bosco et al., 2018) at EVALITA 2018[3] for investigating respectively misogyny and HS in Italian.

HatEval consists in detecting hateful contents in social media texts, specifically in Twitter's posts, against two targets: immigrants and women. Moreover, the task implements a multilingual perspective where data for two widespread languages, English and Spanish, are provided for training and testing participant systems.

The motivations for organizing HatEval go beyond the advancement of the state of the art for HS detection for each of the involved languages and targets. The variety of targets of hate and languages provides a unique comparative setting, both with respect to the amount of data collected and annotated applying the same scheme, and with respect to the results achieved by participants training their systems on those data. Such comparative setting may help in shedding new light on the linguistic and communication behaviour against these targets, paving the way for the integration of HS detection tools in several application contexts. Moreover, the participation of a very large amount of research groups in this task (see Section 4) has improved the possibility of in-depth investigation of the involved phenomena.

The paper is organized as follows. In the next section, the datasets released to the participants for training and testing the systems are described. Section 3 presents the two subtasks and the measures we exploited in the evaluation. Section 4 reports on approaches and results of the participant systems. In Section 5, a preliminary analysis of common errors in top-ranked systems is proposed. Section 6 concludes the paper.

2 Data

The data have been collected using different gathering strategies. For what concerns the time frame, tweets have been mainly collected in the time span from July to September 2018, with the exception of data with target women. Indeed, the most part of the training set of tweets against women has been derived from an earlier collection carried out in the context of two previous challenges on misogyny identification (Fersini et al., 2018a,b). Different approaches were employed to collect tweets: (1) monitoring potential victims of hate accounts, (2) downloading the history of identified haters and (3) filtering Twitter streams with keywords, i.e. words, hashtags and stems. Regarding the keyword-driven approach, we employed both neutral keywords (in line with the collection strategy applied in Sanguinetti et al. (2018)), derogatory words against the targets, and highly polarized hashtags, in order to collect a corpus for reflecting also on the subtle but important differences between HS, offensiveness (Wiegand et al., 2018) and stance (Taulé et al., 2017). The keywords that occur more frequently in the collected tweets are: *migrant*, *refugee*, *#buildthatwall*, *bitch*, *hoe*, *women* for English, and *inmigra-*, *arabe*, *sudaca*, *puta*, *callate*, *perra* for Spanish[4].

The entire HatEval dataset is composed of 19,600 tweets, 13,000 for English and 6,600 for Spanish. They are distributed across the targets as follows: 9,091 about immigrants and 10,509 about women (see also Tables 1 for English and 2 for Spanish). Figures 1 and 2 show the distribution of the labels in the training and development set data according to the different targets of hate (woman and immigrants, respectively).

	Training		Test	
Label	**Imm.**	**Women**	**Imm.**	**Women**
Hateful	39.76	44.44	42.00	42.00
Non-Hateful	60.24	55.56	58.00	58.00
Individual Target	5.89	64.94	3.33	80.63
Generic Target	94.11	35.06	96.67	19.37
Aggressive	55.08	30.06	59.84	34.44
Non-Aggressive	44.92	69.94	40.16	65.56

Table 1: Distribution percentages across sets and categories for English data. The percentages for the target and aggressiveness categories are computed on the total number of hateful tweets.

	Training		Test	
Label	**Imm.**	**Women**	**Imm.**	**Women**
Hateful	41.93	41.38	40.50	42.00
Non-Hateful	58.07	58.62	59.50	58.00
Individual Target	13.72	87.58	32.10	94.94
Generic Target	86.28	12.42	67.90	5.06
Aggressive	68.58	87.58	50.31	92.56
Non-Aggressive	31.42	12.42	46.69	7.44

Table 2: Distribution percentages across sets and categories for Spanish data. The percentages for the target and aggressiveness categories are computed on the total number of hateful tweets.

[3] http://evalita.org

[4] The complete set of keywords exploited is available here: https://github.com/msang/hateval/blob/master/keyword_set.md

2.1 Annotation

The data are released after the annotation process, which involved non-trained contributors on the crowdsourcing platform *Figure Eight* (F8)[5]. The annotation scheme applied to the HatEval data is a simplified merge of schemes already applied in the development of corpora for HS detection and misogyny by the organizers (Fersini et al., 2018a,b; Bosco et al., 2018), also in the context of funded projects with focus on the tasks topics[6] (Sanguinetti et al., 2018; Poletto et al., 2017). It includes the following categories:

- **HS** - a binary value indicating if HS is occurring against one of the given targets (women or immigrants): 1 if occurs, 0 if not.
- **Target Range** - if HS occurs (i.e. the value for the feature HS is 1), a binary value indicating if the target is a generic group of people (0) or a specific individual (1).
- **Aggressiveness** - if HS occurs (i.e. the value for the feature HS is 1), a binary value indicating if the tweeter is aggressive (1) or not (0).

We gave the annotators a series of guidelines in English and Spanish, including the definition for hate speech against the two targets considered, the aggressiveness's definition and a list of examples[7]. As requested by the platform, we provided a restricted set of "correct" answers to test the reliability of the annotators. We required to collect at least three independent judgments for each tweet. We adopted the default F8 settings for assigning the majority label (relative majority). The F8 reported average confidence (i.e., a measure combining inter-rater agreement and reliability of the contributor) on the English dataset for the fields HS, TR, AG is 0.83, 0.70 and 0.73 respectively, while for the Spanish dataset is 0.89, 0.47 and 0.47. The use of crowdsourcing has been successfully already experimented in several tasks and in HS detection too, both for English (Davidson et al., 2017) and other languages (Sanguinetti et al., 2018). However, stimulated by the discussion in (Basile et al., 2018), we decided to apply

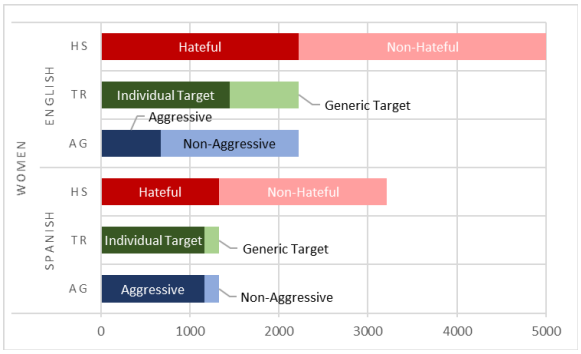

Figure 1: Distribution of the annotated categories in English and Spanish training and development set for the target women.

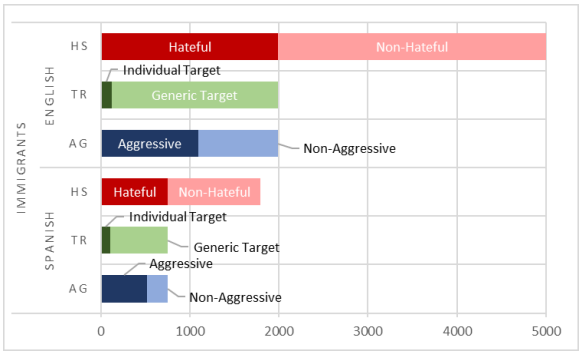

Figure 2: Distribution of the annotated categories in English and Spanish training and development set for the target immigrants.

a similar methodology by adding two more expert annotations to all the crowd-annotated data, provided by native or near-native speakers of British English and Castilian Spanish, having a long experience in annotating data for the specific task's subject. We assigned the final label for this data based on majority voting from *crowd*, *expert1*, and *expert2*. This does not erase the contribution of the crowd, but hopefully maximises consistency with the guidelines in order to provide a solid evaluation benchmark for this task.

For data release and distribution each post has been identified by a newly generated index which substitutes the original Twitter's IDs.

2.2 Training, Development and Test Data

Data for training and development were released according to the distribution described in Figures 1 and 2 across languages (Spanish and English) and targets (women and immigrants). For what concerns Spanish, the training and development set includes 5,000 tweets, (3,209 for the target women and 1,991 for immigrants), while for English it in-

[5] http://www.figure-eight.com/

[6] http://hatespeech.di.unito.it/ihateprejudice.html.

[7] Annotation guidelines provided are accessible here: https://github.com/msang/hateval/blob/master/annotation_guidelines.md.

cludes 10,000 tweets (5,000 for each target). For a cross-language perspective see Figures 1 and 2. It can be also observed that the distribution across categories is pivoting around the main task category, HS, while the other ones more freely vary. Indeed, in order to provide a more balanced distribution of the HS and non-HS categories in the dataset released for Subtask A, we altered the natural distribution: both in the training and test set, hateful tweets are over-represented with respect to the distribution observed in the data we collected from Twitter[8]. Instead, the distribution of the other categories which are relevant for Subtask B is not constrained, and naturally follows from the selection of tweets for representing the classes relevant for the main Subtask A.

As far as the test set is concerned, 3,000 tweets have been annotated for English, half with target women and half immigrants, and 1,600 for Spanish distributed with the same proportion across the targets of hate: 1,260 hateful tweets and 1,740 non-hateful tweets for English, 660 hateful tweets and 940 non-hateful tweets for Spanish.

According to the schema described above, the format of an annotated tweet in the training and development set has the following pattern:

ID, Tweet-text, HS, TR, AG

where ID is a progressive number denoting the tweet within the dataset, Tweet-text is the given text of the tweet, while the other parts of the pattern, given in the training data and to be predicted in the test set, are: Hate Speech [HS] (1 or 0), Target Range [TR] (0 for group or 1 for individual), and Aggressiveness [AG] (0 or 1). Data included in the test instead only include ID and Tweet-text, the annotation of HS, TR and AG to be provided by participants according to the subtask.
An example of annotation is the following:

7, lol, chop her head off and rape the bitch https://t.co/ZB8CosmSD8, 1, 1, 1

which has been considered by the annotators as hateful, against an individual target, and aggressive. The latter category is not necessarily associated to HS, as shown in the following example, where a hateful content is expressed against a generic group of people in terms of disrespect and misogynistic stereotypes rather than using an aggressive language:

11, WOW can't believe all these women riding the subway today? Shouldn't these bitches be making sandwiches LOL #ihatefemales.., 1, 0, 0

3 Task Description

The task is articulated around two related subtasks. The first consists of a basic detection of HS, where participants are asked to mark the presence of hateful content. In the second subtask instead fine-grained features of hateful contents are investigated in order to understand how existing approaches may deal with the identification of especially dangerous forms of hate, i.e., those where the incitement is against an individual rather than against a group of people, and where an aggressive behaviour of the author can be identified as a prominent feature of the expression of hate. The participants will be asked in this latter subtask to identify if the target of hate is a single human or a group of persons, and if the message author intends to be aggressive, harmful, or even to incite, in various forms, to violent acts against the target (see e.g. (Sanguinetti et al., 2018)).

3.1 Subtask A - Hate Speech Detection against immigrants and women

Subtask A is a two-class (or binary) classification task where the system has to predict whether a tweet in English or in Spanish with a given target (women or immigrants) contains HS or not. The following sentences present examples of a hateful and non-hateful tweet where the targets are women.

[hateful]

[id: 32411] Cállate @ y la gran puta madre que te reparió. Que le diste a la poltica...nada. Basura.
Shut up @ you motherfucker. What did you do for politics... nothing. Trash.[9]

[non-hateful]

[id: 33033] @ This is inhumane 😔 Karma is a bitch she ll get around these brainless heartless assholes!

[8]The whole original annotated dataset was very skewed towards the non-HS class (only about 10% of the annotated data contained hate speech).

[9]The target of the misogynistic hate here is Victoria Donda Prez, an Argentinian woman, human rights activist and member of the Argentine National Congress (mentioned in the at-mention of the original tweet).

3.2 Subtask B - Aggressive behaviour and Target Classification

Next, in Subtask B systems are asked to classify hateful tweets (e.g., tweets where HS against our targets has been identified) regarding both aggressive attitude and the target harassed. On one hand, the kind of target must be classified, and the task is binary:

- *Individual*: the text includes hateful messages purposely sent to a specific target.
- *Generic*: it refers to hateful messages posted to many potential receivers.

[**Individual**]:

> [id: 4723] @ Come on box I show you my cock darling

[**Generic**]:

> [id: 5823] Women are equal and deserve respect. Just kidding, they should suck my dick.

On the other hand, the aggressive behaviour has to be identified, then we propose a two-class classification task also for this feature. A tweet must be classified as aggressive or not:

[**Aggressive**]

> [id:1890] Sick barstewards! This is what happens when we put up the refugees welcome signs! They not only rape our wives or girlfriends, our daughters but our ruddy mothers too!! https://t.co/XAYLr6FjNk

[**Non-Aggressive**]

> [id: 945] @EmmanuelMacron Hello?? Stop groping my nation.Schneider: current migrant crisis represents a plan orchestrated and prepared for a long time by international powers to radically alter Christian and national identity of European peoples.http

3.3 Evaluation Measures and Baseline

The evaluation of the results considers different strategies and metrics for Subtasks A and B in order to allow more fine-grained scores.

Subtask A. Systems will be evaluated using standard evaluation metrics, including Accuracy, Precision, Recall and macro-averaged F_1-score.

In order to provide a measure that is independent on the class size, the submissions will be ranked by macro-averaged F_1-score, computed as described in (Özgür et al., 2005). The metrics will be computed as follows:

$$Accuracy = \frac{number\ of\ correctly\ predicted\ instances}{total\ number\ of\ instances} \quad (1)$$

$$Precision = \frac{number\ of\ correctly\ predicted\ instances}{number\ of\ predicted\ labels} \quad (2)$$

$$Recall = \frac{number\ of\ correctly\ predicted\ labels}{number\ labels\ in\ the\ gold\ standard} \quad (3)$$

$$F_1\text{-score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

Subtask B. The evaluation of systems participating to Subtask B will be based on two criteria: (1) partial match and (2) exact match. Regarding the partial match, each dimension to be predicted (HS , TR and AG) will be evaluated independently from the others using standard evaluation metrics, including accuracy, precision, recall and macro-averaged F_1-score. We will report to the participants all the measures and a summary of the performance in terms of macro-averaged F_1-score, computed as follows:

$$F_1\text{-score} = \frac{F_1(HS) + F_1(AG) + F_1(TR)}{3} \quad (5)$$

Concerning the exact match, all the dimensions to be predicted will be jointly considered computing the Exact Match Ratio (Kazawa et al., 2005). Given the multi-label dataset consisting of n multi-label samples (x_i, Y_i), where x_i denotes the i-th instance and Y_i represents the corresponding set of labels to be predicted (HS $\in \{0, 1\}$, TR $\in \{0, 1\}$ and AG $\in \{0, 1\}$), the Exact Match Ratio (EMR) will be computed as follows:

$$EMR = \frac{1}{n}\sum_{i=1}^{n} I(Y_i, Z_i) \quad (6)$$

where Z_i denotes the set of labels predicted for the i-th instance and I is the indicator function. The submissions will be ranked by EMR. This choice is motivated by the willingness to capture the difficulty of modeling the entire phenomenon, and therefore to identify the most dangerous behaviours against the targets.

Baselines. In order to provide a benchmark for the comparison of the submitted systems, we

considered two different baselines. The first one (*MFC baseline*) is a trivial model that assigns the most frequent label, estimated on the training set, to all the instances in the test set. The second one (*SVC baseline*) is a linear Support Vector Machine (SVM) based on a TF-IDF representation, where the hyper-parameters are the default values set by the scikit-learn Python library (Pedregosa et al., 2011).

4 Participant Systems and Results

HatEval has been one of the most popular tasks in SemEval-2019 with a total of 108 submitted runs for Subtask A and 70 runs for Subtask B. We received submission from 74 different teams, of which 22 teams participated to all the subtasks for the two languages[10].

Besides traditional Machine Learning approaches, it has been observed that more than half of the participants investigated Deep Learning models. In particular, most of the systems adopted models known to be particularly suitable for dealing with texts, from Recurrent Neural Networks to recently proposed language models (Sabour et al., 2017; Cer et al., 2018). Consequently, external resources such as pre-trained Word Embeddings on tweets have been widely adopted as input features. Only a few works deepen the linguistic features analysis, probably due to the high expectations on the ability of Deep Learning models to extract high-level features. Most of the submitted systems adopted traditional preprocessing techniques, such as tokenization, lowercase, stopwords, URLs and punctuation removal. Some participants investigated Twitter-driven preprocessing procedures such as hashtag segmentation, slang conversion in correct English and emoji translation into words. It is worth mentioning that the construction of customized hate lexicons derived by the detection of language patterns in the training set has been preferred to the use of external hate lexicons expressing a more universal knowledge about the hate speech phenomenon, additionally demonstrating the need of developing more advanced approaches for detecting hate speech towards women and immigrants.

[10]The evaluation results are published here: https://docs.google.com/spreadsheets/d/1wSFKh1hvwwQIoY8_XBVkhjxacDmwXFpkshYzLx4bw-0/

4.1 Subtask A - Hate Speech Detection against immigrants and women

We received 69 submissions to the English Subtask A, of which 49% and 96% outperformed the SVC and MFC baseline respectively, in terms of macro-averaged F_1-score. Among the five best performing teams, only the team of *Panaetius*, which obtained the second position (0.571), has not provided a description of their system. The higher macro-averaged F_1-score (0.651) has been obtained by the Fermi team. They trained a SVM model with RBF kernel only on the provided data, exploiting sentence embeddings from Google's Universal Sentence Encoder (Cer et al., 2018) as features. Both the third, fourth and fifth ranked teams employ Neural Network models and, more specifically, Convolutional Neural Networks (CNNs) and Long Short Term Memory networks (LSTMs). In particular, the third position has been obtained by the *YNU_DYX* team, which system achieved 0.535 macro-averaged F_1-score by training a stacked Bidirectional Gated Recurrent Units (BiGRUs) (Cho et al., 2014) exploiting fastText word embeddings (Joulin et al., 2017). Then, the output of BiGRU is fed as input to the capsule network (Sabour et al., 2017). The textual preprocessing has been conducted with standard procedures, e.g. punctuation removal, tokenization, contraction normalization, use of tags for hyperlinks, numbers and mentions. The fourth place has been achieved by the team of *alonzorz* (0.535), which used a novel type of CNN called Multiple Choice CNN on the top of contextual embeddings. These embeddings have been created with a model similar to Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) trained using 50 million unique tweets from the Twitter Firehose dataset. The *SINAI-DL* team ranked fifth with a F_1-score of 0.519. They employ a LSTM model based on the pre-trained GloVe Word Embeddings from Stanford-NLP group (Pennington et al., 2014). Since Deep Learning models require a large amount of data for training, they perform data augmentation through the use of paraphrasing tools. For preprocessing the texts in the specific Twitter domain, they convert all the mentions to a common tag and they tokenized hashtags according to the Camel Case procedure, i.e. the practice of writing phrases such that each word or abbreviation in the middle of the phrase begins with a capital letter, with no inter-

vening spaces or punctuation.

For Subtask A in Spanish, we received 39 submissions of which 51% and 100% outperformed the SVC and MFC baseline respectively, in terms of macro-averaged F_1-score. The *Atalaya* and *MineriaUNAM* teams obtained the best macro-averaged F_1-score of 0.73, both taking advantage of Support Vector Machines. The *Atalaya* team studied several sophisticated systems, however the best performances have been obtained by a linear-kernel SVM trained on a text representation composed of bag-of-words, bag-of-characters and tweet embeddings, computed from fastText sentiment-oriented word vectors. The system proposed by the *MineriaUNAM* team is based on a linear-kernel SVM. The study has focused on a combinatorial framework used to search for the best feature configuration among a combination of linguistic patterns features, a lexicon of aggressive words and different types of n-grams (characters, words, POS tags, aggressive words, word jumps, function words and punctuation symbols). The *MITRE* team has achieved the performance of 0.729, presenting a novel method for adapting pre-trained BERT models to Twitter data using a corpus of tweets collected during the same time period of the HatEval training dataset. The *CIC-2* team achieved 0.727 with a word-based representation by combining Logistic Regression, Multinomial Naïve Bayes, Classifiers Chain and Majority Voting. They used TF and TF/IDF after removing HTML tags, punctuation marks and special characters, converting slang and short forms into correct English words and stemming. The participants did not use external resources and trained their systems only with the provided data. Finally, the *GSI-UPM* team obtained the macro-averaged F_1-score of 0.725 with a system where the linear-kernel SVM has been trained on an automated selection of linguistic and semantic features, sentiment indicators, word embeddings, topic modeling features, and word and character TF-IDF n-grams.

Table 3 shows basic statistics computed both for Subtasks A and B, with respect to the relative performance measures. The statistics comprise mean, standard deviation (StdDev), minimum, maximum, median and the first and third quartiles (Q1 and Q3). Concerning Subtask A, we notice that the maximum value in Spanish (0.7300) is higher than the English one (0.6510),

	Subtask A		**Subtask B**	
	English	**Spanish**	**English**	**Spanish**
Min.	0.3500	0.4930	0.1590	0.4280
Q1	0.4050	0.6665	0.2790	0.5820
Mean	0.4484	0.6821	0.3223	0.6013
Median	0.4500	0.7010	0.3120	0.6160
StdDev	0.0569	0.0521	0.0890	0.0662
Q3	0.4880	0.7165	0.3570	0.6365
Max.	0.6510	0.7300	0.5700	0.7050
SVC Baseline	0.451	0.701	0.308	0.588
MFC Baseline	0.367	0.370	0.580	0.605

Table 3: Basic statistics of the results for the participating system and baselines in Subtask A and Subtask B expressed in terms of macro-averaged F_1-score and EMR respectively.

while the difference is even higher (23 points) when considering the mean value, from 0.6821 to 0.4484. On the other hand, the variability is very similar between English (0.0569) and Spanish (0.0521).

4.2 Subtask B - Aggressive behaviour and Target Classification

For Subtask B in English, we received 39 submissions, of which no system has been able to outperform the MFC baseline, which achieved 0.580 of EMR, while 61% outperformed the SVC baseline. Among the five best performing teams, only the team of *scmhl5*, which obtained the third position (0.483), has not provided us with a description of the system. The higher EMR result has been obtained by the *LT3* team with a value of 0.570. They considered a supervised classification-based approach with SVM models which combines a variety of standard lexical and syntactic features with specific features for capturing offensive language exploiting external lexicons. The second position has been obtained by the *CIC-1* team. The team achieved 0.568 in EMR with Logistic Regression and Classifier Chains. They trained their model only with the provided data, with a word-based representation and without external resources. The only preprocessing action was stemming and stop words removal. The fourth position was obtained by the team named The Titans. They achieved 0.471 of EMR with LSTM and TF/IDF-based Multilayer Perceptron. To represent the documents, they used the tweet words after removing links, mentions and spaces. They also tokenized hashtags into word tokens. The MITRE team exploits the same approach used for participating in Subtask A, obtaining 0.399 EMR. It is worth men-

tioning that, despite the fact that the baseline could not be overcome in terms of EMR, the five first performing systems obtained higher F-values. For example, while the baseline obtained 0.421, the *scmhl5* (0.632) and the *MITRE* team (0.614) systems obtained about 20 points over it.

For Subtask B in Spanish, we received 23 submissions of which 52% and 70% outperformed the SVC and MFC baseline respectively, in terms of EMR. The first position has been achieved by the CIC-2 team with 0.705 in terms of EMR, proposing the same approach for Subtask A in Spanish. The *CIC-1* and *MITRE* teams, described previously, achieved the second and third positions with 0.675 and 0.675 in EMR respectively. The fourth position was obtained by the *Atalaya* team that achieved 0.657 EMR by extending the previously presented approach for Subtask A to a 5-way classification problem for all the possible label combinations. Finally, the team of *Oscar-Garibo* achieved the fifth position (0.6444) with Support Vector Machines and statistical embeddings to represent the texts. The proposed method, a variation of LDSE (Rangel et al., 2016), consists of finding thresholds on the frequencies of use of the different terms in the corpora depending on the class they belong to. In this subtask, the correlation between EMR and macro-averaged F_1-score is more homogeneous than in English. However, it is worth mentioning the case of the *CIC-1* team since its macro-averaged F_1-score decreases with respect to the EMR and is 10 points lower than the rest of the best five performing teams.

The comparative results between all the performing teams in the two languages show interesting insights (see Table 3). Firstly, the best result is much higher in the case of Spanish (0.7050) than in English (0.5700) in more than 13 points. In the case of the fifth best results, the difference is much higher (0.2454), from 0.3990 in English to 0.6440 in Spanish. The average value changes from 0.3223 in English to 0.6013 in Spanish, with a difference of 28 points. The variability is also higher in English (0.0890) with respect to the value in Spanish (0.0662).

We can also derive further conclusions by comparing the statistics of the two Subtasks. Looking at the median, it is possible to notice that in both languages, the performances obtained on Subtask B are lower than the performances of Subtask A, with a difference between Subtask A and B of 14 and 8 points for English and Spanish respectively. This suggests that participant systems found much harder to predict the aggressiveness and targets than just the presence of hate speech. The quartile Q1 has highlighted that for the English language 75% of the systems obtained a score higher than 0.41 and 0.28 for Subtasks A and B, in particular 50 out of 69 for Subtask A and 31 out of 41 for Subtask B. While Q3 shows that 25% of the systems achieved a score value higher than 0.49 and 0.36 for Subtasks A and B, in particular 18 out of 69 for Subtask A and 11 out of 41 for Subtask B. For the Spanish language, the value of Q1 indicates that 75% of the systems have a score higher than 0.67 and 0.58 for Subtasks A and B, in particular 30 out of 39 for Subtask A and 17 out of 23 for Subtask B. Observing the quartile Q3, it is possible to observe that 25% of the systems achieved a value higher than 0.72 and 0.64 for Subtasks A and B, in particular 10 out of 39 for Subtask A and 6 out of 23 for Subtask B. Moreover, it is worth mentioning that the smaller the standard deviation the closer are the data to the mean value, highlighting that the Subtask B has shown high variability in terms of results than Subtask A. This statistics remarks again the difficulties of addressing Subtask B compared to Subtask A.

5 Error Analysis

In order to gain deeper insight into the results of the HatEval evaluation, we conducted a first error analysis experiment. For both languages, we selected the three top-ranked systems and checked the instances in the test set that were wrongly labeled by all three of them.

In the English Subtask A, the three top systems (*Fermi*, *Panaetius*, and *YNU_DYX*) predicted the same wrong labels 569 times out of 2,971 (19.1%). In the Spanish Subtask A, the three top systems (*Atalaya*, *mineriaUNAM*, and *MITRE*) predicted the same wrong labels 234 times out of 1,600 (14.6%). The results showing the percentages by wrongly assigned labels are summarized in Table 4.

Subtask	Errors	Predicted 1	Predicted 0
EN A	569	507 (89.1%)	62 (10.9%)
ES A	234	178 (76.1%)	56 (23.9%)

Table 4: Number of instances mislabeled by all the three top-ranked systems, broken down by wrongly assigned label.

The common errors are highly skewed towards the false positives. However, the unbalance is stronger for English (89.1% false positives) than for Spanish (76% false positives).

Two English examples, respectively a false positive and a false negative, are:

> [id: 30249] My mom FaceTimed me to show off new shoes she got and was like "no cabe duda que soy una Bitch" i love her 😂
>
> [id: 30542] @ @ There are NO INNOCENT people in detention centres #SendThemBack

The false positive contains a swear word ("Bitch") used in a humorous, not offensive context, which is a potential source of confusion for a classifier. The false negative is a hateful message towards migrants, but phrased in a slightly convoluted way, in particular due to the use of negation ("no innocent people").

Similarly, a false positive and a false negative in Spanish:

> [id: 33119] Soy un sudaca haciendo sudokus 🙃 https://t.co/vA7nQsfm85
> *I am a sudaca doing sudokus*
>
> [id: 34455] Estoy escuchando una puta canción y la pelotuda de Demi Lovato se pone a hablar en el medio. CANTÁ Y CALLATE LA BOCA.
> *I am listening to a fucking song and that asshole Demi Lovato starts talking in the middle of it. SING AND SHUT YOUR MOUTH.*

Like in the English example, in this false positive a negative word ("sudaca") is used humorously, for the purpose of a wordplay. In the false negative, there a misogynistic message is expressed, although covertly, implying that the target should "shut up and sing".

6 Conclusion

The very high number of participating teams at HatEval 2019 confirms the growing interest of the community around abusive language in social media and hate speech detection in particular. The presence of this task at SemEval 2019 was indeed very timely and the multilingual perspective we applied by developing data in two different widespread languages, English and Spanish, contributed to include and raise interest in a wider community of scholars. 38 teams sent their system reports to describe the approaches and the details of their participation to the task, contributing in shedding light on this difficult task. Some of the HatEval participants also participated to the OffensEval[11], another task related to abusive language identification, but with an accent on the different notion of *offensiveness*, an orthogonal notion that can characterize also expressions that cannot be featured as hate speech[12]. Overall, results confirm that hate speech detection against women and immigrants in micro-blogging texts is challenging, with a large room for improvement. We hope that the dataset made available as part of the shared task will foster further research on this topic, including its multilingual perspective.

Acknowledgments

Valerio Basile, Cristina Bosco, Viviana Patti and Manuela Sanguinetti are partially supported by Progetto di Ateneo/CSP 2016 (*Immigrants, Hate and Prejudice in Social Media*, S1618_L2_BOSC_01).

References

Valerio Basile, Nicole Novielli, Danilo Croce, Francesco Barbieri, Malvina Nissim, and Viviana Patti. 2018. Sentiment polarity classification at evalita: Lessons learned and open challenges. *IEEE Transactions on Affective Computing*.

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 the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*. CEUR-WS.org.

Cristina Bosco, Patti Viviana, Marcello Bogetti, Michelangelo Conoscenti, Giancarlo Ruffo, Rossano Schifanella, and Marco Stranisci. 2017. Tools and Resources for Detecting Hate and Prejudice Against Immigrants in Social Media. In *Proceedings of First Symposium on Social Interactions in Complex Intelligent Systems (SICIS), AISB Convention 2017, AI and Society*.

Miguel Ángel Álvarez Carmona, Estefanía Guzmán-Falcón, Manuel Montes-y-Gómez, Hugo Jair Escalante, Luis Villaseñor Pineda, Verónica Reyes-Meza, and Antonio Rico Sulayes. 2018. Overview

[11] https://competitions.codalab.org/competitions/20011

[12] See (Sanguinetti et al., 2018) for a deeper reflection on hate speech and offensiveness.

of MEX-A3T at IberEval 2018: Authorship and Aggressiveness Analysis in Mexican Spanish Tweets. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*. CEUR-WS.org.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *CoRR*, abs/1803.11175.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *CoRR*, abs/1703.04009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018a. Overview of the EVALITA 2018 Task on Automatic Misogyny Identification (AMI). In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*. CEUR-WS.org.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018b. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*. CEUR-WS.org.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 427–431.

Hideto Kazawa, Tomonori Izumitani, Hirotoshi Taira, and Eisaku Maeda. 2005. Maximal margin labeling for multi-topic text categorization. In *Advances in Neural Information Processing Systems*, pages 649–656.

Ritesh Kumar, Atul Kr. Ojha, Marcos Zampieri, and Shervin Malmasi, editors. 2018. *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. ACL.

Kate Manne. 2017. *Down Girl. The Logic of Misogyny*. Oxford University Press.

John T. Nockleby. 2000. Hate speech. *Encyclopedia of the American Constitution (2nd ed., edited by Leonard W. Levy, Kenneth L. Karst et al., New York: Macmillan, 2000)*, pages 1277–1279.

Arzucan Özgür, Levent Özgür, and Tunga Güngör. 2005. Text categorization with class-based and corpus-based keyword selection. In *International Symposium on Computer and Information Sciences*, pages 606–615. Springer.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Fabio Poletto, Marco Stranisci, Manuela Sanguinetti, Viviana Patti, and Cristina Bosco. 2017. Hate Speech Annotation: Analysis of an Italian Twitter Corpus. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*. CEUR-WS.org.

Francisco Rangel, Paolo Rosso, and Marc Franco-Salvador. 2016. A low dimensionality representation for language variety identification. In *17th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing*. Springer-Verlag, LNCS.

Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. 2017. Dynamic routing between capsules. In *Advances in neural information processing systems*, pages 3856–3866.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter Corpus of Hate Speech against Immigrants. In *Proceedings of the 11th Language Resources and Evaluation Conference 2018*.

Mariona Taulé, Maria 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 Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017)*. CEUR-WS.org.

Zeerak Waseem, Wendy Hui Kyong Chung, Dirk Hovy, and Joel Tetreault, editors. 2017. *Proceedings of the First Workshop on Abusive Language Online*. ACL.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. In *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)*.