Experiments in Hate Speech Detection

Anonymous ACL submission

Abstract

Hate speech detection and related tasks have emerged as major NLP research topics. In this paper we benefit from the resources and reference results provided by the recent HatEval shared task to investigate a range of alternative implementations of these tasks, based on both shallow (SVM and logistic regression) and deep (CNN and LSTM) learning methods, and using word n-grams, char n-grams and psycholinguistic features alike. From these experiments, a simple model combining logistic regression and n-gram feature selection emerges as a potentially suitable approach to the hate speech detection subtask, with results that are comparable to those obtained by the top-performing systems at the competition.

1 Introduction

Hate speech detection and related tasks (e.g., the recognition of offensive or abusive language use, aggressiveness, misogyny, racism, xenophobia, homophobia etc.) have emerged as major research topics in Natural Language Processing (NLP). Existing methods are usually based on supervised machine learning, often making use of Twitter (Carmona et al., 2018; Wiegand et al., 2018; Basile et al., 2019) and, to a lesser extent, Facebook (Vigna et al., 2017; Kumar et al., 2018) data, or both (Bosco et al., 2018).

As evidence of their popularity, tasks of this kind have been the focus of several recent events (or shared tasks), including the case of hate or otherwise abusive speech detection (Wiegand et al., 2018; Bosco et al., 2018; Basile et al., 2019), and aggressive language detection (Carmona et al., 2018; Kumar et al., 2018), among others.

Of particular interest for the present work, the HatEval shared task (Basile et al., 2019) has distributed a labelled corpus of English and Spanish tweets conveying statements about two target groups, namely, women and immigrants. The corpus supports the study of at least three forms of language abuse: hate speech and aggressiveness detection, and target group classification (i.e., towards a particular individual or general aggression) and, accordingly, has been the centre piece of the shared task in (Basile et al., 2019).

In the current work we benefit from the resources and reference results in (Basile et al., 2019) to investigate a range of alternative implementations of these tasks, based on both shallow (SVM and logistic regression) and deep (CNN and LSTM) learning methods, and using word n-grams, char n-grams and psycholinguistic features alike. From these experiments, a simple model combining logistic regression and n-gram feature selection emerges as a potentially suitable approach to the hate speech detection subtask, with results that are comparable to those obtained by the top-performing systems at the competition.

The rest of this paper is organised as follows. Section 2 discusses the three subtasks under consideration, and reviews related work in the field. Section 3 described our main models and compares their results based on the HatEval development dataset. Section 5 focuses on the hate speech detection subtask, and applies our overall best-performing model to the test data provided. Finally, Section 6 presents additional remarks and points to future work.

2 Background

2.1 Overview

Hate speech detection and related tasks are now mainstream NLP research topics. A gentle intro-

duction to this subject is provided in (Schmidt and Wiegand, 2017), in which methods, features and research questions are discussed. The study also addresses the use of word- and character-based features and their predictive effectiveness for the task, and points out to a number of possible improvements such as the use of word clusters and others.

Tasks of this kind are usually implemented by making use of supervised machine learning methods. Interestingly, however, the use of deep learning has so far shown some mixed results. On the one hand, studies as in (Kumar et al., 2018) (in the context of aggressive language detection) suggest that neural network models may show little improvement over more traditional methods, and that standard approaches such as SVM and logistic regression may produce similar results with careful feature selection. Moreover, we notice that all of the top-performing systems in both of HatEval subtasks make use of either SVM or logistic regression models, and present results that are superior to those obtained by alternatives based on CNNs, LSTMs and other deep architectures (Basile et al., 2019).

On the other hand, deep learning methods are of course highly effective in many NLP tasks, including some of the tasks presently under discussion. Studies as in (Kshirsagar et al., 2018; Lee et al., 2018) (see below) provide evidence of this for certain tasks and domains, and particularly so when large datasets are available. In what follows, a number of recent works in the field is briefly reviewed.

2.2 Related work

The work in (Burnap and Williams, 2016) addresses the issue of hate speech detection from Twitter text based on multiple targets (e.g., race, disabilities, sexual orientation etc.) A number of simple strategies are evaluated, including both bag-of-words and typed dependency models, and using both SVM and Random Forest classifiers.

The work in (Nobata et al., 2016) makes use of a regression method to detect hate speech on online user comments posted on Yahoo! Finance and News. A wide range of features are investigated, including syntactic features and word embedding variations, which are found to be particularly effective when combined with standard text features (e.g., word and char unigrams and bigrams, punctuation, word length etc.) The study also suggests that char n-gram models are particularly suitable for handling noisy data of this kind.

The use of char n-grams also play a central role in (Waseem and Hovy, 2016). The study makes use of logistic regression to classify racist and sexist tweets.

The work in (Kshirsagar et al., 2018) presents a neural-network based approach to three classification tasks: general hate speech, racist and sexist language use. The proposal makes us of word embeddings and max/mean pooling from fully connected embeddings transformations, and outperforms a number of existing models, including the work in (Waseem and Hovy, 2016) and others - while using a significantly lower number of features.

A number of recent studies in hate speech detection from Twitter have been based on the 80k tweet dataset described in (Founta et al., 2018). The corpus has been labelled with multiple categories (namely, offensive, abusive, hateful speech, aggressive, cyber bullying, spam, and normal) through crowd sourcing, and supports a potentially wide range of studies focused on the English language.

Using the Twitter dataset provided in (Founta et al., 2018), the work in (Lee et al., 2018) presents a comparative study of learning models of hate speech detection. Among a wide range of models - including Naive Bayes, SVM, Logistic regression, CNN and RNN classifiers - a Bidirectional Gate recurrent Unit (GRU) network model trained on word-based features and using Latent Topic Clustering is shown to outperform the alternatives.

Deep learning methods are also at the centre of the experiments described in (Zhang et al., 2018). The study makes use of word embeddings and a CNN network with max pooling to provide input vectors to a GRU neural network.

Finally, we notice that a large number of recent events and shared tasks have addressed the issues of hate speech recognition and aggressive language use. For further details, we report to (Kumar et al., 2018; Bosco et al., 2018; Carmona et al., 2018; Wiegand et al., 2018) and, in particular, results from the recent HatEval shared task (Basile et al., 2019), whose training and test datasets were taken as the basis of the present work as well.

3 Current work

3.1 Motivation

The present work is based on the corpus provided by the recent HatEval shared task (Basile et al., 2019). The corpus contains 19,600 tweets (13,000 in English and 6,600 in Spanish) potentially conveying offensive language towards women and migrants.

Being part of a shared task competition, the corpus was released in two portions (development and test data.) In this section we will first investigate a range of models and algorithms for these tasks, and present results based on the development dataset only. In Section 5, our best-performing model will be validated using the HatEval test dataset.

As discussed in (Basile et al., 2019), the HatEval corpus is labelled with three kinds of information: hateful x non-hateful, aggressive x non-aggressive, and individual x general target. In the original shared task, two subtasks were proposed: subtask A concerns hate speech detection, which exploits the hateful x non-hateful corpus labels, and subtask B concerns aggressive behaviour and target classification, which exploits the combination of both aggressive x non-aggressive and individual x general target corpus labels. Hate speech detection results are measured by macro-averaged F1 scores, and aggressiveness/target classification results are measured by a combined Exact Match Ratio (EMR) metrics.

In the present work, we address the hate speech detection task exactly as proposed by subtask A in (Basile et al., 2019). On the other hand, we will address the issues of aggressiveness and target group classification as being independent from each other. Thus, our results for hate speech detection will be directly comparable to those reported in (Basile et al., 2019), whereas our results for aggressiveness and target group classification will not.

3.2 Models

The tasks at hand - hate speech detection, aggressiveness and target group classification based on the HatEval corpus - are presently modelled as three independent binary classification problems. For each task, we ran several pilot experiments based on both shallow (SVM and logistic regression) and deep (CNNs and LSTMs) learning methods, and using word n-grams, char n-grams and psycholinguistic features. However, since discussing every possible combination would be beyond the scope of the present work, only four of these models were selected for the present discussion. These models, and a simple baseline system, are summarised in Table 1 and further discussed below.

Table 1: Models under consideration

Model name	Method	Features
CNN.word	CNN	word n-grams
LSTM.char	LSTM	char n-grams
LIWC	log.reg.	LIWC word counts
LR.Tfidf	log.reg.	k-best TF-IDF counts
Majority	baseline	na

Our first two models are standard implementation of deep learning methods that are popular in hate speech detection and many other NLP tasks, namely, convolutional (CNN) and long short-term memory (LSTM) networks. Both models resort to SMOTE oversampling (Chawla et al., 2002) to minimize class imbalance.

CNN.word consists of a word-based CNN with two convolution channels with filters of size 3 and 4 with a mapping of size$=$ 64, and using one-hot encoding. The model uses ReLu as an activation function with L2 regularisation of 0.003, and max pooling size $=$ 2. The model uses a fully-connected layer with 1024 neurons using drop-out regularisation $=$ 0.3, and a softmax output layer. Training is performed in mini batches of size $=$ 32 using RMSProp optimisation, and using cross-entropy loss as a cost function. Validation is performed over a 20% portion of the training data with Early Stopping.

LSTM.char consists of a character-based LSTM with attention mechanism and using one-hot encoding as well. The model uses memory units of size $=$ 64 and drop-out regularisation $=$ 0.12. This is followed by a hidden layer containing 1024 neurons using ReLu as an activation function, and a softmax output layer. Training is performed using AdaDelta optimisation and using cross-entropy loss as a cost function. Once again, validation is performed over a 20% portion of the training data with Early Stopping.

In addition to these two deep learning methods, our third model, hereby called *LIWC*, makes us of logistic regression over a set of psycholinguistics-

Table 2: Optimal k-values for the *LR.Tfidf* model

Task	English	Spanish
Hate speech	4500	1000
Aggressiveness	3000	10000
Target Group	2000	2000

motivated word category counts provided by the Linguistic Inquiry and Word Count (LIWC) dictionary (Pennebaker et al., 2001). This (perhaps unusual) choice for the tasks at hand is motivated by the observation that LIWC contains word categories such as 'negative emotion', 'sexual' and others, which potentially open up the opportunity for detecting hate speech without training examples, that is, in a dictionary-based fashion. No oversampling was performed in this case.

For the English language tasks, our current *LIWC* model makes use of the 93-feature set in (Pennebaker et al., 2015). For the Spanish language tasks, since we did not have access of the appropriate LIWC version, the corpus was machine-translated into Portuguese, and our models made use of the 64-feature set for this language instead (Filho et al., 2013). Although in this case some accuracy loss is to be expected, we assume that the Spanish and Portuguese languages are still sufficiently close - at least for the purpose of building lexically-motivated models of this kind - and that the use of machine-translated text will not have a major impact on the results if compared to the corresponding tasks for the English language.

Our fourth model, hereby called *LR.Tfidf*, makes use of logistic regression over unigram and bigram TF-IDF counts subject to univariate feature selection using ANOVA F1 as a score function. As in the case of *LIWC*, no oversampling was performed. Optimal k values for each task and language were obtained by performing grid search over the training portion of the data in the 1000 to 15000 range at 500 intervals. These are summarised in Table 2.

Finally, a *Majority* class baseline model is also included in the present evaluation for illustration purposes. This is similar to the MFC baseline employed at the HatEval shared task (Basile et al., 2019).

4 Evaluation on development data

In what follows we report results provided by our four models - *CNN.word*, *LSTM.char*, *LIWC* and *LR.Tfidf* - and by the *Majority* class baseline. All results are solely based on the development portion of the HatEval corpus, which was subject to a stratified random 80:20 split.

4.1 Hate speech detection

Table 3 shows results for the HatEval hate speech detection task using only the development dataset provided by the HatEval organisers. Best F1 scores for each language are highlighted.

From these results, we notice that the *LR.Tfidf* model largely outperforms all alternatives in both languages. This outcome is in principle consistent with the overall results of the HatEval shared task, as will be discussed in Section 5.

4.2 Aggressiveness detection

Table 4 shows results for aggressive language detection, once again using only the HatEval development data. Best F1 scores for each language are highlighted.

Onge again, we notice that the *LR.Tfidf model* outperforms the alternatives for the Spanish dataset. In the case of the English dataset, however, *LR-Tfidf* was outperformed by *LSTM.char*. These results are in principle consistent with (Kumar et al., 2018), in which the use of simple learning algorithms with feature selection was found to produce competitive results if compared to deep learning approaches to aggressive language detection.

4.3 Target group classification

Finally, Table 5 shows results for target group detection based on HatEval development data. Best F1 scores for each language are highlighted.

The *LR.Tfidf* model outperforms all alternatives in both languages as well. We notice also that this task was considerable easier than the previous two, with particularly high accuracy for the Spanish dataset.

5 Evaluation on test data (hate speech detection)

From the series of experiments described in the previous section, we found little motivation to further apply deep learning methods to the HateEval subtasks. Moreover, given that our experiments are only fully comparable with HatEval subtask A (hate speech recognition), in this section we will focus on this subtask only, assessing our overall

Table 3: Hate speech detection results using HatEval development data. Best F1 scores for each language are highlighted.

	English			Spanish		
Model	P	R	F	P	R	F
Majority	0.34	0.58	0.43	0.34	0.59	0.43
CNN.word	0.51	0.51	0.51	0.51	0.50	0.41
LSTM.char	0.52	0.52	0.51	0.50	0.50	0.50
LIWC	0.68	0.68	0.68	0.66	0.66	0.66
LR.Tfidf	0.77	0.77	**0.77**	0.81	0.81	**0.81**

Table 4: Aggressive language detection results using HatEval development data. Best F1 scores for each language are highlighted.

	English			Spanish		
Model	P	R	F	P	R	F
Majority	0.35	0.59	0.44	0.65	0.81	0.72
CNN.word	0.52	0.51	0.49	0.58	0.54	0.47
LSTM.char	0.44	0.45	0.42	0.90	0.88	**0.88**
LIWC	0.63	0.62	0.62	0.81	0.65	0.68
LR.Tfidf	0.68	0.67	**0.68**	0.82	0.83	0.83

Table 5: Target group detection results using HatEval development data. Best F1 scores for each language are highlighted.

	English			Spanish		
Model	P	R	F	P	R	F
Majority	0.42	0.65	0.51	0.37	0.61	0.46
CNN.word	0.52	0.51	0.48	0.51	0.51	0.51
LSTM.char	0.67	0.66	0.65	0.48	0.48	0.45
LIWC	0.81	0.80	0.81	0.82	0.81	0.81
LR.Tfidf	0.85	0.85	**0.85**	0.92	0.92	**0.92**

best-performing *LR.Tfidf* model - based on logistic regression over feature-selected TF-IDF counts - in the hate speech recognition task.

In what follows, the *LR.Tfidf model* is evaluated against the (unseen) HatEval test data, and results are compared to those obtained by the shared task best-performing participants as reported in (Basile et al., 2019). These results are summarised in Table 6.

In the case of the English language dataset, we notice that *LR.Tfidf* approach outperforms both HatEval baseline systems (called SVC and MFC) which, according to (Basile et al., 2019), obtained 0.45 and 0.37 F1 scores, respectively. However, this is still well below the 0.65 F1 score obtained by the top-performing participant system - Fermi - as reported in (Basile et al., 2019).

In the case of the Spanish dataset, our approach (with a 0.73 macro F1 score) also outperforms both HatEval baseline systems. The gain over SVC (which obtained 0.70) is relatively small, but the gain over MFC (0.30) is large. This result is the same obtained by the two top-performing systems - Atalaya and MineriaUNAM - reported in (Basile et al., 2019), which might suggest an upper limit for this particular task.

6 Final remarks

In this paper we have investigated a range of learning methods and features for three tasks - hate speech detection, aggressiveness and target group classification - as proposed in the HatEval shared task (Basile et al., 2019), and taking advantage of the labelled corpus and reference results provided by the task organisers. In a first round of experiments, the three tasks were investigated by making use of the development data, and then our overall best-performing model for hate speech detection was validated on the shared task test dataset.

The present approach to hate speech detection consisted of a logistic regression model based on word n-grams and univariate feature selection. All results were superior to those obtained by the HatEval baseline systems in (Basile et al., 2019) and, in the case of the Spanish dataset, were similar to those obtained by the top-performing participants in the shared task. Only in the case of the English dataset the current results remained below those obtained by the HatEval winner.

As in the case of the top-performing systems discussed in (Basile et al., 2019), our current results seem to confirm that, at least for relatively small datasets of this kind, the use of standard learning methods and feature selection techniques may still produce results that are generally comparable to those obtained by deep architectures. This seems to be suggested also by the results obtained by our simple *LIWC* lexical model, which is based on psycholinguistic word category counts. The *LIWC* model was rated as the second best strategy in most of all experiments, including even those developed for the Spanish language, which made use of machine-translated text instead of the actual corpus data.

As future work, we intend to revisit the current deep learning models, and further investigate the dictionary-based approach as an alternative to using large training datasets. The assessment of the current models as a joint task as proposed in HatEval subtask B (aggressiveness and target groupd classification) also remains to be implemented.

Acknowledgements

This work received support by (anon.) The authors are also grateful to Valerio Basile and all the HatEval team for providing us with the dataset and organising the shared task.

References

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Rangel, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)*. Association for Computational Linguistics.

Cristina Bosco, Felice Dell'Orletta, Fabio Poletto, Manuela Sanguinetti, and Maurizio Tesconi. 2018. Overview of the EVALITA 2018 hate speech detection task.

Pete Burnap and Matthew L. Williams. 2016. Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data Science* 5(1):11. https://doi.org/10.1140/epjds/s13688-016-0072-6.

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 of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in mexican spanish tweets. In *IberEval@SEPLN*.

Table 6: LR.Tfidf Hate speech recognition final results using HatEval test data.

	English			Spanish		
Class	P	R	F	P	R	F
0	0.78	0.22	0.34	0.75	0.83	0.79
1	0.46	0.92	0.61	0.72	0.61	0.66
macro avg	0.62	0.57	0.48	0.74	0.72	0.73

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Int. Res.* 16(1):321–357.

Pedro P. Balage Filho, Sandra M. Aluísio, and T.A.S. Pardo. 2013. An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In *9th Brazilian Symposium in Information and Human Language Technology - STIL*. Fortaleza, Brazil, pages 215–219.

Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*. AAAI Publications.

Rohan Kshirsagar, Tyrus Cukuvac, Kathy McKeown, and Susan McGregor. 2018. Predictive embeddings for hate speech detection on twitter. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Association for Computational Linguistics, Brussels, Belgium, pages 26–32.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. Association for Computational Linguistics, Santa Fe, USA, pages 1–11.

Younghun Lee, Seunghyun Yoon, and Kyomin Jung. 2018. Comparative studies of detecting abusive language on twitter. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Association for Computational Linguistics, Brussels, Belgium.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, Geneva, Switzerland, pages 145–153. https://doi.org/10.1145/2872427.2883062.

J. W. Pennebaker, R. L. Boyd, and K. Jordanand K. Blackburn. 2015. The development and psychometric properties of LIWC2015. Technical report, University of Texas at Austin.

J. W. Pennebaker, M. E. Francis, and R. J. Booth. 2001. *Inquiry and Word Count: LIWC*. Lawrence Erlbaum, Mahwah, NJ.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. Association for Computational Linguistics, Valencia, Spain, pages 1–10. https://doi.org/10.18653/v1/W17-1101.

Fabio Del Vigna, Andrea Cimino, Felice Dell?Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*. Venice, Italy, pages 86–95.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*. San Diego, USA.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 shared task on the identification of offensive language. In *14th Conference on Natural Language Processing (KONVENS 2018)*.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution GRU based deep neural network. In *The Semantic Web*. Springer International Publishing, Cham, pages 745–760.