ISSN 2007-9737

Data-Driven and Psycholinguistics-Motivated Approaches to Hate Speech Detection

Samuel Caetano da Silva[1], Thiago Castro Ferreira[2], Ricelli Moreira Silva Ramos[1], Ivandré Paraboni[1]

[1] University of São Paulo, School of Arts, Sciences and Humanities,
Brazil

[2] Federal University of Minas Gerais, Arts Faculty,
Brazil

{samuel.caetano.silva,ricelliramos,ivandre}@usp.br,
thiago.castro.ferreira@gmail.com

Abstract. Computational models of hate speech detection and related tasks (e.g., detecting misogyny, racism, xenophobia, homophobia etc.) have emerged as major Natural Language Processing (NLP) research topics in recent years. In the present work, we investigate a range of alternative implementations of three of these tasks - namely, hate speech, aggressive behavior and target group recognition - by presenting a number of experiments involving different learning methods, including regularized logistic regression, convolutional neural networks (CNN) and deep bidirectional transformers (BERT), and using word embeddings, word n-grams, character n-grams and psycholinguistics-motivated (LIWC) features alike. Results suggest that a purely data-driven BERT model, and to some extent also a hybrid psycholinguisticly informed CNN model, generally outperform the alternatives under consideration for all tasks in both English and Spanish languages.

Keywords. Natural language processing, hate speech, aggressive language detection.

1 Introduction

Aggressiveness, threats and other forms of abuse that may harm individuals and disrupt social relations are all ubiquitous in the language employed in on-line communication.

As a response to these challenges, 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 the Natural Language Processing (NLP) field. Existing methods are usually based on supervised machine learning, often making use of Twitter data [2, 5, 26] and, to a lesser extent, Facebook data [13, 24], or both [3].

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 [2, 3, 26], and aggressive language detection [5, 13], among others.

In the present work, we will focus on three forms of abusive language towards two target groups - namely, women and immigrants - as proposed in [2]. More specifically, given a message conveying potentially abusive language, we address the issues of hate speech, aggressiveness (i.e., whether abusive language is aggressive or not) and target group classification (i.e., whether the abuse is directed towards a specific individual or towards a general group). Examples of each of these tasks - taken from the Twitter corpus presented in [2] - are illustrated as follows.

Hate speech:

— *#Refugees go home* (hateful),

— *iMMIGRANT SONG !* (non-hateful).

doi: 10.13053/CyS-24-3-3478

Aggressiveness:

— *Bad girls get spankings* (hateful, aggressive) ,

— *Good girls send nudes* (hateful, non-aggressive).

Target group:

— *THIS.#IllegalAliens* (hateful towards a general group),

— *Stupid Skank !* (hateful towards an individual).

Based on messages of this kind, the present work will investigate a number of data-driven and psycholinguistics-motivated models that make use of word embeddings, word n-grams, character n-grams and psycholinguistics-motivated (LIWC) features alike, and which are based on regularized logistic regression, convolutional neural networks (CNN), and pre-trained BERT deep bidirectional transformers [7]. In doing so, we would like to determine the models that are best suited to each task in both Spanish and English languages.

The rest of this paper is organized as follows. Section 2 reviews related work in the field. Section 3 discusses the dataset that supports the three tasks under consideration. Section 4 introduces a number of models for these tasks. Section 5 describes the evaluation procedure and results. Section 6 presents our final remarks and points to ongoing and future work.

2 Related Work

Hate speech detection and related tasks are now mainstream NLP research topics. A gentle introduction to this subject is provided in [22], in which multiple methods, features, and research questions are discussed. The study also addresses the use of word- and character-based features by discussing their predictive effectiveness in 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 models has so far shown somewhat mixed results. On the one hand, studies as in [13] (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 provided that careful feature selection is performed. Moreover, we notice that all of the top-performing systems at HatEval subtasks made use of either SVM or logistic regression models, and presented results that are superior to those obtained by alternatives based on CNNs, LSTMs and other possibly more sophisticated approaches [2].

On the other hand, deep learning methods have of course been highly effective in many NLP tasks, including some of the tasks presently under discussion. Studies as in [12, 14] described in the next section provide some evidence in favour of these methods, and particularly so in cases where larger datasets happen to be available. In what follows, a number of recent works in the field are briefly reviewed.

The work in [4] addresses the issue of hate speech detection from Twitter text focused on a number of pre-defined target topics (e.g., racial issues, disabilities, sexual orientation etc.) A number of simple strategies are evaluated, including both bag-of-words and typed dependency models, and using both support vector machine (SVM) and Random Forest classifiers.

The work in [17] makes use of a regression method to detect hate speech from on-line 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 character unigram and bigram counts, punctuation, word length etc.) The study also suggests that character n-gram models are particularly suitable for handling noisy data of this kind.

The use of character n-grams also plays a central role in [25], in which character-level information is found to be generally superior to using word-level features and others. The study makes use of logistic regression to classify racist and sexist tweets in English.

The work in [12] presents a neural-network based approach to classify hateful, racist and sexist language use. The proposal makes use of word embeddings and max/mean pooling from fully connected embeddings transformations, and outperforms a number of existing models – including the aforementioned work in [25] 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 [9]. The corpus has been labeled with multiple categories (namely, offensive, abusive, hateful speech, aggressive, 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 [9], the work in [14] presents a comparative study of learning models of hate speech detection. Among a large number of strategies - including Naive Bayes, SVM, Logistic regression, CNN and recurrent neural networks (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.

Neural models are also at the centre of the experiments described in [28].

The study makes use of word embeddings and a CNN network with max pooling to provide input vectors to a GRU neural network.

The work in [15] addresses the issue of abusive language detection (racism and sexism) based on the Twitter corpus in [9]. In this approach, graph convolutional networks (GCNs) are designed so as to capture both the structure of on-line communities and their linguistic behavior. Results of this approach combined with logistic regression generally outperform a number of baseline systems based on logistic regression and others, including the current best-performing system for this corpus [16].

The work in [27] introduces a dataset of 14,100 English tweets - called Offensive Language Identification Dataset (OLID) - annotated with three levels of information: offensive language detection (offensive / not offensive), offensive language categorization (untargeted / targeted insult) and target identification (individual, group, other). Similarities with the HatEval corpus [2] are discussed, and reference results based on SVM, Bi-LSTM and CNN models are reported.

We notice also 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 [3, 5, 13, 26] and, in particular, to the HatEval shared task results [2], whose training and test datasets were taken as the basis of the present work as well, as discussed in the next sections.

3 Data

The present work makes use of the HatEval shared task corpus data discussed in [2]. The corpus contains 19,600 tweets (13,000 in English and 6,600 in Spanish) potentially conveying offensive language towards women and migrants, and targeting either single individuals or general groups.

Corpus texts are provided with labels representing three kinds of binary information: hateful versus non-hateful, aggressive versus non-aggressive, and individual versus general target. This organization gives rise to the three computational tasks under discussion, namely, hate speech, aggressive behavior and target group classification.

Being originally part of a shared task competition, the HatEval corpus is provided in two fixed, pre-defined development and test datasets. The development dataset contains 5,000 tweets in Spanish (3,209 of which targeted women, and 1,991 targeted migrants) and 10,000 tweets in English (with 5,000 instances for each target.)

The test dataset contains further 1,600 tweets in Spanish and 3,000 tweets in English, in both cases keeping the same class distribution of the development dataset for each language.

4 Methods

In what follows we introduce a number of models to address the issues of hate speech, aggressiveness and target group classification as supported by the corpus HatEval [2] described in the previous section. In doing so, each problem is modelled as an independent binary classification task.

For each task, we ran several pilot experiments based on a range of learning methods, and using word embeddings, word n-grams, character n-grams and psycholinguistic features alike. However, since discussing every possible combination would be beyond the scope of the present work, in what follows we shall focus on the two best-performing alternatives, hereby called *BERT-En/Sp* and *CNN.glove+liwc*. These two models, alongside four baseline systems, are summarized in Table 1 and further discussed in the next sections.

4.1 Pre-Trained Deep Bidirectional Transformer Model

The use of pre-trained BERT deep bidirectional transformer models [7] is now a staple in the NLP field, with state-of-the-art results for a wide range of tasks, including general language understanding evaluation, question answering, and many others. In the present work, we will use English and Spanish BERT models - hereby called *BERT-En/Sp* - as examples of a purely data-drive approaches to hate speech, aggressiveness and target group classification.

We presently discuss results obtained by the bert-base-cased (for English) and bert-base-multilingual-cased (Spanish) versions of the model presented in [7], which have been fine-tuned to each individual task and language. This consists of 12 layers with hidden-layers of size 768, 12 self-attention heads and feed-forward layers of size (4 * hidden size) 3072, comprising 110 million parameters in total.

The *BERT-En/Sp* approach works as follows. First, text words are embedded and summed up with positional embeddings. Next, the normalized embedded words are fed to the BERT encoder, where each layer self-attends the input word vectors (BERTSelfAttention) and transforms the outcome through three dense layers (called BertSelfOutput, BertIntermediate and BertOutput.) Once the encoding process is completed, the output representation is transformed by a dense layer with a tanh activation function (BertPooler). Finally, a classifier layer is fed with the current representation to make a binary decision based on the underlying task. The *BERT-En/Sp* architecture is illustrated in Figure 1.

4.2 Hybrid Psycholinguistics-Motivated CNN Model

Before the recent popularity of pre-trained BERT models, convolutional neural networks have long been a popular choice in a large number of NLP tasks [11]. At least in the case of the HatEval hate speech detection shared task [2], however, models of this kind have been outperformed by much simpler alternatives (namely, based on logistic regression.)

A possible explanation for this outcome (and which was indeed confirmed by our own pilot experiments, presently not reported) is the relatively small size of the dataset and, perhaps to a lesser extent, the issue of data imbalance. Thus, as a means to overcome some of these difficulties, we developed a hybrid CNN model that takes into account not only the input text itself (i.e., in a data-driven fashion) but also a second input representation conveying psycholinguistics-motivated features provided by the *Linguistic Inquiry and Word Count* (LIWC) dictionary [19]. LIWC includes word categories such as ‘negative emotions’, ‘sexual’ and others, which may potentially correlate with affective language use, and may arguably enhance results of a purely data-driven model. For further details regarding LIWC word categories, we refer to [19].

For text representations, we use Glove word embeddings [20]. For LIWC vectors, we use two different language-dependent knowledge sources as follows. In the case of the English language tasks, LIWC features are computed from the 93-feature LIWC set in [18]. For the Spanish language tasks, since we do not have access to the appropriate (Spanish) LIWC

Table 1. Models under consideration

Model name	Method	Features
BERT-En/Sp	BERT-base	cased words
CNN.glove+liwc	CNN	word embeddings + LIWC
LR.LIWC	log.regression	LIWC word counts
LR.char	log.regression	TF-IDF character n-grams
LR.word	log.regression	TF-IDF word n-grams
Majority	majority class	na

dictionary, LIWC features were computed from a Portuguese machine-translated version of the original (Spanish) corpus obtained from Google Translate with no further revision, and using the 64-feature set for this language instead [1]. 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.

All LIWC vectors were built by counting the number of words in the text that belong to each LIWC dictionary category. When a word belongs to more than one categories (e.g. 'he' is both a personal pronoun and a masculine word, among other possibilities), all related counts are updated. Finally, word counts are normalized by the number of words in each text, hence obtaining feature values within the 0..1 range.

The resulting model, hereby called *CNN.glove+liwc*, takes as an input parallel word embeddings and LIWC vectors. The model consists of a word-based CNN with two convolution channels ($conv1$ and $conv2$) with filters of size 2 and 3 with a mapping of size $= 64$ followed by a Batch Normalization and a MaxPooling layer with a 50% dropout. LIWC features are fed into a convolution with filter size = 1 and mapping size = 64. This is followed by a Batch Normalization and a MaxPooling layers, and then concatenated with $conv1$ and $conv2$. The LIWC layer ($liwc\_layer$) is concatenated with the two convolutions ($conv1 \oplus conv2 \oplus liwc\_layer$) with a 50% dropout layer, and finally fed into a softmax output layer. Training was performed using k-fold cross-validation with a 16 batch size and using Adam as a solver with a 5-fold optimisation function. To this end, the training portion of the data is divided using a 80:20 split. The *CNN.glove+liwc* architecture is illustrated in Figure 2.

4.3 Baseline Systems

Given the prominence of shallow (e.g., logistic regression and SVM) methods for hate speech detection in [2], our two main approaches discussed above are to be compared against a number of baseline systems of this kind. More specifically, we will consider three alternatives that make use of regularized logistic regression based on TF-IDF word unigrams and bigrams hereby called *LR.word*, variable length TF-IDF character n-grams *LR.char*, and LIWC counts (as discussed in the previous section) hereby called *LR.LIWC*.

Character n-grams in *LR.char* are in the 2..4 range for the English models, and in the 2..5 range for the Spanish models. These parameters were set by performing grid search over development data.

The three LR baseline systems make use of k-best univariate feature selection using ANOVA F1 as a score function. Optimal k values for each task, language and model were obtained by performing grid search over development data.

In the case of *LR.word*, k values were sampled in the X to 1000 range at -500 intervals, where X is the number of features of each model. In the case of *LR.char*, k values were sampled in the 25000 to 5000 range at -1000 intervals. In the case of *LR.LIWC* all possible k values were attempted from X to 5. Optimal k values for each model are summarized in Table 2.

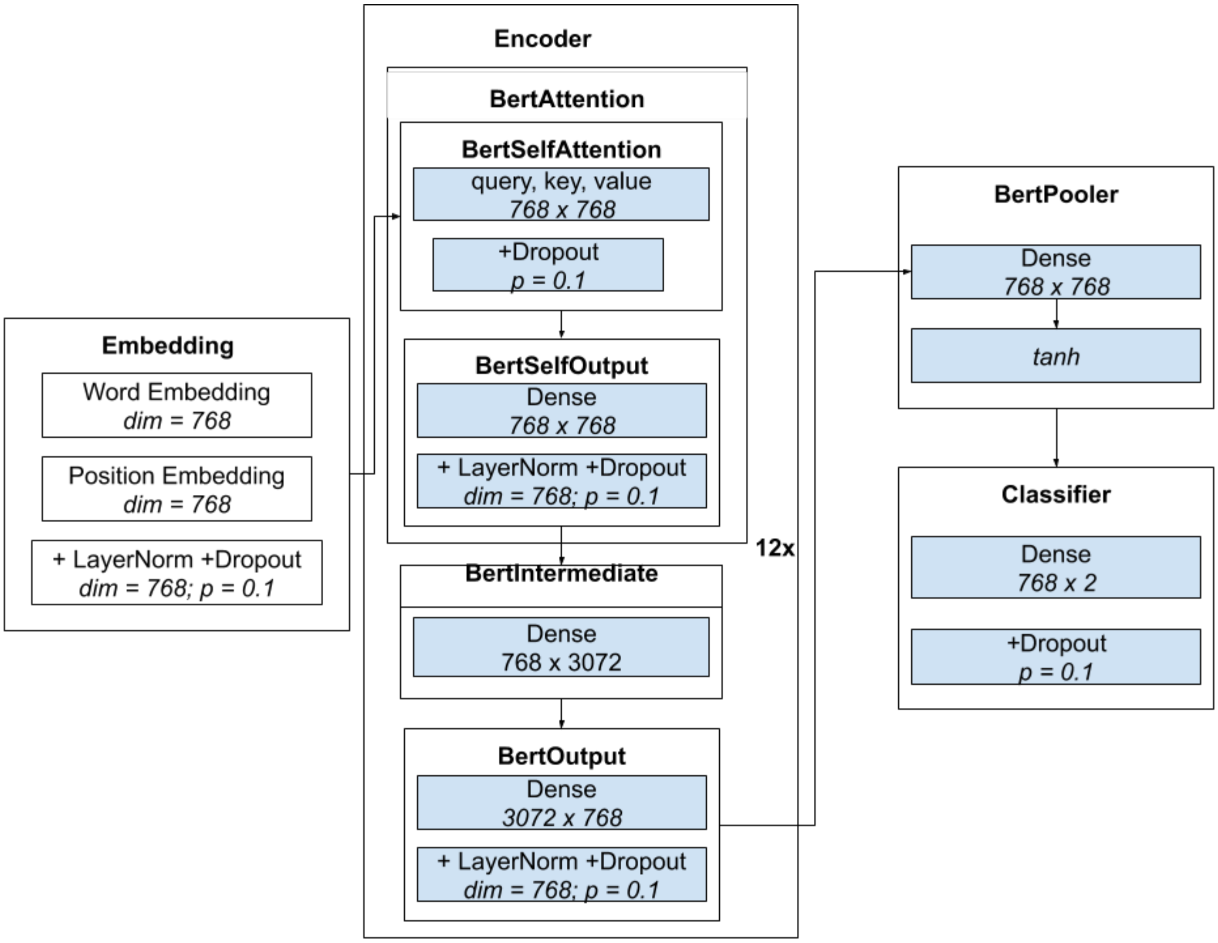

Fig. 1. BERT-En/Sp architecture

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 [2].

5 Evaluation

In what follows, we report results provided by our two main models - *BERT.En/Sp* and *CNN.glove+liwc* - and by the four baseline systems *LR.char*, *LR.word*, *LR.LIWC* and *Majority*, all of which discussed in the previous sections.

These are to be compared against the best-performing systems reported in [2].

5.1 Procedure

Evaluation was carried out by using the previously unseen test portion of the data and by measuring F1 scores.

Comparison with the participant systems in the HatEval shared task is complicated by the fact that different systems in [2] address different tasks and languages, and that multiple evaluation criteria were considered. Thus, the present analysis includes the best results taken from [2] regarding HatEval sub task B (which combines the three tasks presently under discussion) according to their averaged F1 scores. Although details about these participant systems remain scarce, the shared task summary report points out that the

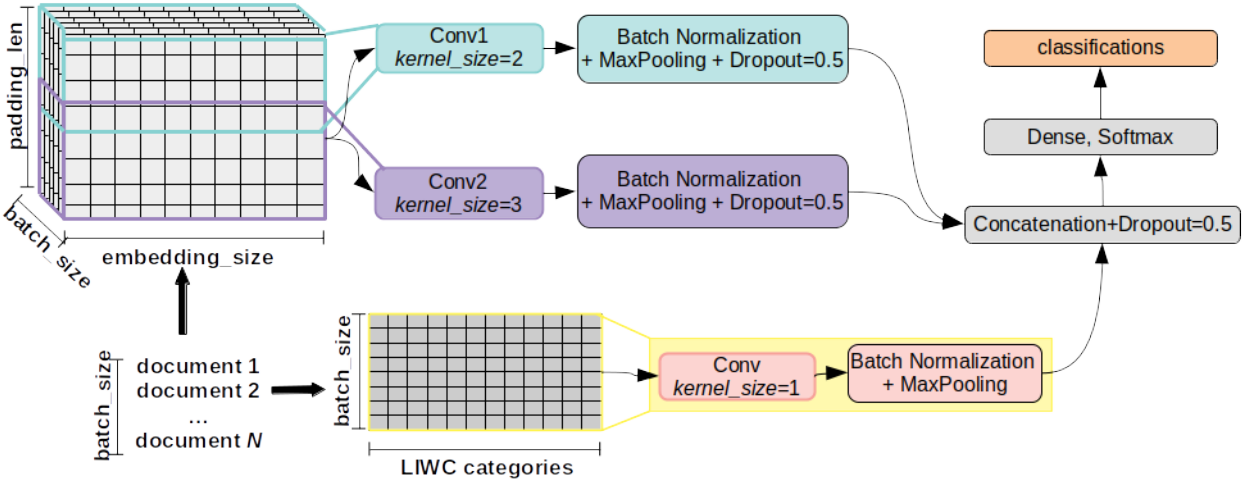

Fig. 2. CNN.glove+liwc architecture

Table 2. Baseline univariate feature selection optimal k-values for each task, language and feature set (words, characters or LIWC features

Language	Task	word	char	LIWC
English	Hate speech	3000	19000	78
	Aggressiveness	5731	19000	60
	Target Group	5731	17000	7
Spanish	Hate speech	2762	17000	64
	Aggressiveness	2681	13000	52
	Target Group	2681	9000	40

best-performing systems were mostly based on logistic regression or SVM models, and that deep learning approaches were not ranked at the first positions in any of the three tasks at hand.

The systems taken from [2] as a basis for comparison with our current work are summarized in Table 3.

5.2 Results

Results for each classification task in are shown in Table 4 for the English dataset, and in Table 5 for Spanish. In both cases, best F1 scores are highlighted, and the Mean column presents the average of the three tasks, corresponding to the 'partial match' metrics adopted in [2].

5.3 Discussion

From the current results, a number of observations are warranted. First, we notice that complexity varies considerable across tasks and languages. Generally speaking, the English tasks were more difficult than their Spanish counterparts. This may be due to the richer morphology of the Spanish language, which may favour the use of word and subword-based models. Moreover, we notice that hate speech classification is generally more difficult than the other tasks in both languages. Target group classification, by contrast, obtained the overall highest accuracy among the three tasks.

Regarding the comparison between models, we notice that *BERT-En/Sp* generally outperforms the alternatives, and particularly so in the case of target group classification. Differences from the

Table 3. HatEval subtask B best-performing systems according to averaged F1 scores for each task and language

Language	Task	System
English	Hate speech	scmhl5
	Aggressiveness	alonzorz
	Target Group	YNU_NLP
Spanish	Hate speech	gertner
	Aggressiveness	gertner
	Target Group	Saagie

Table 4. F1 scores results for the English dataset. Best results for each task and language are highlighted

Model	Hate speech	Aggressiveness	Target group	Mean
Majority	0.37	0.44	0.45	0.42
LR.char	0.37	0.39	0.63	0.46
LR.word	0.39	0.39	0.64	0.47
LR.liwc	0.49	0.48	0.64	0.54
CNN.glove+liwc	**0.60**	0.59	0.82	0.67
BERT-En	0.53	0.68	**0.85**	**0.69**
HatEval best	**0.60**	**0.75**	0.62	0.66

HatEval participant systems are in some cases small, but it is worth remembering that results on the 'HatEval best result' row actually represent the best out of five different systems, and that none of these would provide results that are as consistently high as those obtained from *BERT-En/Sp* across tasks and languages.

Leaving the comparison with the previous work aside, we notice that both *BERT-En/Sp* and *CNN.glove+liwc* outperform the baseline alternatives by a considerable margin. Despite some positive results obtained by shallow methods in [2], this outcome was to some extent to be expected as both models are considerably more complex than the baseline systems. Moreover, we notice that purely data-driven *BERT-En/Sp* generally outperforms the more informed *CNN.glove+liwc* model.

This, in our view, may be seen as further evidence of the positive results obtained by pre-trained deep bidirectional transformer models in other NLP tasks as reported in [7].

Finally, regarding the issue of using translated text instead of the original Spanish data, we notice that the present results do not seem to have been harmed by this as the proposed methods are still, on average, highly comparable to the previous work (presumably developed using the original Spanish data.) Thus, text translation may represent a useful strategy - at least for lexically-motivated tasks of this kind - when certain NLP resources are unavailable, as it was the case of LIWC data for the Spanish language.

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 - in Spanish and English as proposed in [2]. Among these, a purely data-driven BERT model, and to some extent also a hybrid psycholinguisticly informed CNN model, were found to generally outperform previous works in the field for certain task and language combinations.

The present work leaves a number of opportunities are open to further investigation. In particular, we intend to revisit the use of deep learning methods, and further investigate the dictionary-based LIWC approach as an alternative

Table 5. F1 scores results for the Spanish dataset. Best results for each task and language are highlighted

Model	Hate speech	Aggressiveness	Target group	Mean
Majority	0.37	0.41	0.42	0.40
LR.char	0.65	0.60	0.71	0.65
LR.word	0.67	0.66	0.74	0.69
LR.liwc	0.60	0.65	0.69	0.65
CNN.glove+liwc	0.50	**0.83**	0.81	0.71
BERT-Sp	0.73	**0.83**	**0.93**	**0.83**
HatEval best	**0.76**	0.82	0.76	0.78

to using large training datasets. Moreover, we also intend to investigate the possible relation between hate speech and moral stance classification [8], as well as in author profiling [10, 21, 23] and authorship attribution [6].

Acknowledgements

The first and second authors received support from the Brazilian Foundation CAPES - Coordination for the Improvement of Higher Education Personnel, under grants 88882.378103/2019-01 and 88887.367980/2019-00, respectively. The last author received support from the University of São Paulo PRP grant 668/2018. The authors are also grateful to Valerio Basile and to the HatEval team for making the dataset publicly available.

References

1. **Balage Filho, P. P., Aluísio, S. M., & Pardo, T. (2013).** An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. *9th Brazilian Symposium in Information and Human Language Technology - STIL*, Fortaleza, Brazil, pp. 215–219.

2. **Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel, F., Rosso, P., & Sanguinetti, M. (2019).** SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)*, Association for Computational Linguistics.

3. **Bosco, C., Dell'Orletta, F., Poletto, F., Sanguinetti, M., & Tesconi, M. (2018).** Overview of the EVALITA 2018 hate speech detection task. *Proceedings of EVALITA 2018*, CEUR-WS.org.

4. **Burnap, P. & Williams, M. L. (2016).** Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data Science*, Vol. 5, No. 1, pp. 11.

5. **Carmona, M. Á. Á., Guzmán-Falcón, E., y Gómez, M. M., Escalante, H. J., Pineda, L. V., Reyes-Meza, V., & Sulayes, A. R. (2018).** Overview of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in mexican spanish tweets. *IberEval@SEPLN*.

6. **Custódio, J. E. & Paraboni, I. (2018).** EACH-USP ensemble cross-domain authorship attribution. *Working Notes Papers of the Conference and Labs of the Evaluation Forum (CLEF-2018) vol.2125*, Avignon, France.

7. **Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019).** BERT: Pre-training of deep bidirectional transformers for language understanding. *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Minneapolis, USA, pp. 4171–4186.

8. **dos Santos, W. R. & Paraboni, I. (2019).** Moral Stance Recognition and Polarity Classification from Twitter and Elicited Text. *Recents Advances in Natural Language Processing (RANLP-2019)*, Varna, Bulgaria, pp. 1069–1075.

9. **Founta, A. M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018).** Large scale crowdsourcing and characterization of twitter abusive behavior. *Twelfth International AAAI Conference on Web and Social Media*, AAAI Publications.

10. **Hsieh, F. C., Dias, R. F. S., & Paraboni, I. (2018).** Author profiling from facebook corpora. *11th*

International Conference on Language Resources and Evaluation (LREC-2018), ELRA, Miyazaki, Japan, pp. 2566–2570.

11. **Kim, Y. (2014).** Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Doha, Qatar, pp. 1746–1751.

12. **Kshirsagar, R., Cukuvac, T., McKeown, K., & McGregor, S. (2018).** Predictive embeddings for hate speech detection on twitter. *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, Association for Computational Linguistics, Brussels, Belgium, pp. 26–32.

13. **Kumar, R., Ojha, A. K., Malmasi, S., & Zampieri, M. (2018).** Benchmarking aggression identification in social media. *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, Association for Computational Linguistics, Santa Fe, USA, pp. 1–11.

14. **Lee, Y., Yoon, S., & Jung, K. (2018).** Comparative studies of detecting abusive language on twitter. *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, Association for Computational Linguistics, Brussels, Belgium.

15. **Mishra, P., Tredici, M. D., Yannakoudakis, H., & Shutova, E. (2019).** Abusive language detection with graph convolutional networks. *Proceedings of NAACL-HLT 2019*, Association for Computational Linguistics, Minneapolis, USA, pp. 2145–2150.

16. **Mishra, P., Yannakoudakis, H., & Shutova, E. (2018).** Neural character-based composition models for abuse detection. *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, Association for Computational Linguistics, pp. 1–10.

17. **Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016).** Abusive language detection in online user content. *Proceedings of the 25th International Conference on World Wide Web*, International World Wide Web Conferences Steering Committee, Geneva, Switzerland, pp. 145–153.

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

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

20. **Pennington, J., Socher, R., & Manning, C. D. (2014).** GloVe: Global Vectors for Word Representation. *Proceedings of EMNLP-2014*, pp. 1532–1543.

21. **Ramos, R. M. S., Neto, G. B. S., Silva, B. B. C., Monteiro, D. S., Paraboni, I., & Dias, R. F. S. (2018).** Building a corpus for personality-dependent natural language understanding and generation. *11th International Conference on Language Resources and Evaluation (LREC-2018)*, ELRA, Miyazaki, Japan, pp. 1138–1145.

22. **Schmidt, A. & Wiegand, M. (2017).** A survey on hate speech detection using natural language processing. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, Association for Computational Linguistics, Valencia, Spain, pp. 1–10.

23. **Silva, B. B. C. & Paraboni, I. (2018).** Learning personality traits from Facebook text. *IEEE Latin America Transactions*, Vol. 16, No. 4, pp. 1256–1262.

24. **Vigna, F. D., Cimino, A., Dell'Orletta, F., Petrocchi, M., & Tesconi, M. (2017).** Hate me, hate me not: Hate speech detection on facebook. *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, Venice, Italy, pp. 86–95.

25. **Waseem, Z. & Hovy, D. (2016).** Hateful symbols or hateful people? predictive features for hate speech detection on twitter. *Proceedings of the NAACL Student Research Workshop*, Association for Computational Linguistics, San Diego, USA, pp. 88–93.

26. **Wiegand, M., Siegel, M., & Ruppenhofer, J. (2018).** Overview of the GermEval 2018 shared task on the identification of offensive language. *14th Conference on Natural Language Processing (KONVENS 2018).*

27. **Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019).** Predicting the type and target of offensive posts in social media. *Proceedings of NAACL-HLT 2019*, Association for Computational Linguistics, Minneapolis, USA, pp. 1415–1420.

28. **Zhang, Z., Robinson, D., & Tepper, J. (2018).** Detecting hate speech on twitter using a convolution GRU based deep neural network. *The Semantic Web*, Springer International Publishing, Cham, pp. 745–760.

Article received on 14/96/2020; accepted on 21/07/2020.
Corresponding author is Ivandré Paraboni.