UTFPR at SemEval-2019 Task 5:
Hate Speech Identification with Recurrent Neural Networks

Gustavo Henrique Paetzold[1], Shervin Malmasi[2], Marcos Zampieri[3]
[1]Universidade Tecnológica Federal do Paraná, Toledo-PR, Brazil
[2]Harvard Medical School, Boston, United States
[3]University of Wolverhampton, Wolverhampton, United Kingdom
`ghpaetzold@utfpr.edu.br`

Abstract

In this paper we revisit the problem of automatically identifying hate speech in posts from social media. We approach the task using a system based on minimalistic compositional Recurrent Neural Networks (RNN). We tested our approach on the SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter (HatEval) shared task dataset. The dataset made available by the HatEval organizers contained English and Spanish posts retrieved from Twitter annotated with respect to the presence of hateful content and its target. In this paper we present the results obtained by our system in comparison to the other entries in the shared task. Our system achieved competitive performance ranking 7th in sub-task A out of 62 systems in the English track.

1 Introduction

Abusive and offensive content such as aggression, cyberbulling, and hate speech have become pervasive in social media. The widespread of offensive content in social media is a reason of concern for governments worldwide and technology companies, which have been heavily investing in ways to cope with such content using human moderation of posts, triage of content, deletion of offensive posts, and banning abusive users.

One of the most common and effective strategies to tackle the problem of offensive language online is to train systems capable of recognizing such content. Several studies have been published in the last few years on identifying abusive language (Nobata et al., 2016), cyber aggression (Kumar et al., 2018), cyber bullying (Dadvar et al., 2013), and hate speech (Burnap and Williams, 2015; Davidson et al., 2017). As evidenced in two recent surveys (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018) and in a number of other

studies (Malmasi and Zampieri, 2017; Gambäck and Sikdar, 2017; ElSherief et al., 2018; Zhang et al., 2018), the identification of hate speech is the most popular of what Waseem et al. (2017) refers to as "abusive language detection sub-tasks".

This paper deals with the hate speech identification in English and Spanish posts from social media. We present our submissions to the SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter (HatEval) shared task. We participated in sub-task A which is a binary classification task in which systems are trained to discriminate between posts containing hate speech and posts which do not contain any form of hate speech. Our approach, presented in detail in Section 4, combines compositional Recurrent Neural Networks (RNN) and transfer learning and achieved competitive performance in the shared task.

2 Related Work

As evidenced in the introduction of this paper, there have been a number of studies on automatic hate speech identification published in the last few years. One of the most influential recent papers on hate speech identification is the one by Davidson et al. (2017). In this paper, the authors presented the Hate Speech Detection dataset which contains posts retrieved from social media labeled with three categories: OK (posts not containing profanity or hate speech), Offensive (posts containing swear words and general profanity), and Hate (posts containing hate speech). It has been noted in Davidson et al. (2017), and in other works (Malmasi and Zampieri, 2018), that training models to discriminate between general profanity and hate speech is far from trivial due to, for example, the fact that a significant percentage of hate speech posts contain swear words. It has been ar-

gued that annotating texts with respect to the presence of hate speech has an intrinsic degree of subjectivity (Malmasi and Zampieri, 2018).

Along with the recent studies published, there have been a few related shared tasks organized on the topic. These include GermEval (Wiegand et al., 2018) for German, TRAC (Kumar et al., 2018) for English and Hindi, and OffensEval[1] (Zampieri et al., 2019b) for English. The latter is also organized within the scope of SemEval-2019. OffensEval considers offensive language in general whereas HatEval focuses on hate speech.

Waseem et al. (2017) proposes a typology of abusive language detection sub-tasks taking two factors into account: the target of the message and whether the content is explicit or implicit. Considering that hate speech is commonly understood as speech attacking a group based on ethnicity, religion, etc, and that cyber bulling, for example, is understood as an attack towards an individual, the target factor plays an important role in the identification and the definition of hate speech when compared to other forms of abusive content.

The two SemEval-2019 shared tasks, HatEval and OffensEval, both include a sub-task on target identification as discussed in Waseem et al. (2017). HatEval includes the target annotation in its sub-task B with two classes (individual or group) whereas OffensEval includes it in its sub-task C with three classes (individual, group or others). Another important similarity between these two tasks is that both include a more basic binary classification task in sub-task A. In HatEval, posts are labeled as as to whether they contain hate speech or not and in OffensEval, posts are labeled as being offensive or not. As OffensEval considers multiple types of offensive contents, the hierarchical annotation model used to annotate OLID (Zampieri et al., 2019a), the dataset used in OffensEval, includes an annotation level distinguishing between the type of offensive content that posts include with two classes: insults and threats, and general profanity. This type annotation is used in OffensEval's sub-task B.

3 Task Description

HatEval (Basile et al., 2019) provides participants with annotated datasets to create systems capable of properly identifying hate speech in tweets writ-

ten in both English and Spanish.

The training, development, trial, and test sets provided for English are composed of 9,000, 1,000, 100 and 3,000 instances, respectively. The training, development, trial and test sets provided for Spanish are composed of 4,500, 500, 100 and 1,600 instances, respectively. Each instance is composed of a tweet and three binary labels: One that indicates whether or not hate speech is featured in the tweet, one indicating whether the hate speech targets a group or an individual, and another indicating whether or not the author of the tweet is aggressive. HatEval has 2 sub-tasks:

- **Sub-task A:** Judging whether or not a tweet is hateful.

- **Sub-task B:** Correctly predicting all three of the aforementioned labels.

In this paper, we focus on Task A exclusively, for both English and Spanish. We participated in the competition using the team name UTFPR.

4 The UTFPR Models

The UTFPR models are minimalistic Recurrent Neural Networks (RNNs) that learn compositional numerical representations of words based on the sequence of characters that compose them, then use them to learn a final representation for the sentence being analyzed. These models, of which the architecture is illustrated in Figure 1, are somewhat similar to those of Ling et al. (2015) and Paetzold (2018), who use RNNs to create compositional neural models for different tasks.

As illustrated, the UTFPR models take as input a sentence, split it into words, then split the words into a sequence of characters in order to pass them through a character embedding layer. The character embeddings are passed onto a set of bidirectional RNN layers that produces word representations, then a second set of layers produces a final representation of the sentence. Finally, this representation is passed through a softmax dense layer that produces a final classification label.

For each language, we created two variants of UTFPR: one trained exclusively over the training data provided by the organizers (UTFPR/O), and another that uses a pre-trained set of character-to-word RNN layers extracted from the models introduced by Paetzold (2018) (UTFPR/W). The pre-trained model was trained for the English multi-class classification Emotion Analysis shared task
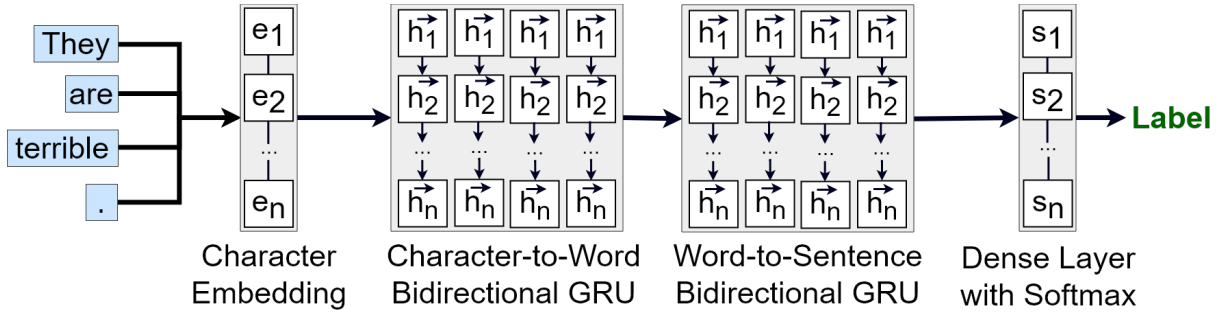

Figure 1: Architecture of the UTFPR models.

of WASSA 2018, which featured a training set of $153,383$ instances composed of a tweet and an emotion label. This pre-trained model for English was used for the UTFPR/W variant of both languages, since we wanted to test the hypothesis that pre-training a character-to-word RNN on a large dataset for English can improve the performance of compositional models for both English and Spanish.

We use 25 dimensions for the size of our character embeddings, and two layers of Gated Recurrent Units for our bidirectional RNNs with 60 hidden nodes each and 50% dropout. We saved a model after each training iteration and picked the one with the lowest error on the development set. The UTFPR/W model went through the same training process as UTFPR/O, with the pre-trained character-to-word RNN layers being fine-tuned for the task at hand.

Table 1 showcases the F-scores obtained by the UTFPR systems on the trial set of Task A. Because of its superior performance, we chose to submit the UTFPR/W variants as our official entry.

| System | F-scores | |
	English	Spanish
UTFPR/O	0.509	0.601
UTFPR/W	0.570	0.665

Table 1: F-scores obtained for the trial set at HatEval Task A for both languages.

5 Results and Discussion

5.1 Shared Task Performance

Tables 2 and 3 feature the F-scores obtained by the UTFPR systems and the 3 best and worst performing systems at HatEval Task A for English and Spanish, respectively. Ultimately, the UTFPR/W systems submitted ranked 7th out of 62 valid sub-

missions for English, and 31st out of 35 valid submissions for Spanish.

System	F-scores
FERMI	0.650
Panaetius	0.570
YNU_DYX	0.550
UTFPR/O	0.524
UTFPR/W	0.513
MELODI	0.350
INGEOTEC	0.350
INAOE-CIMAT	0.340

Table 2: F-scores obtained at HatEval Task A for the English language. At the top and bottom of the table are featured the top and bottom 3 systems submitted to the shared task, respectively.

System	F-scores
mineriaUNAM	0.730
Atalaya	0.730
MITRE	0.730
UTFPR/O	0.664
UTFPR/W	0.636
jhouston	.630
LU team	0.620
TuEval	0.620

Table 3: F-scores obtained at HatEval Task A for the Spanish language. At the top and bottom of the table are featured the top and bottom 3 systems submitted to the shared task, respectively.

One of the aspects we wanted to test with our participation in this shared task was the extent to which pre-training a character-to-word RNN over a larger dataset for an analogous task helped the models. Our results show that, even though using a pre-trained RNN considerably improved the performance of our models in the trial experiments, it actually compromised their performance for the
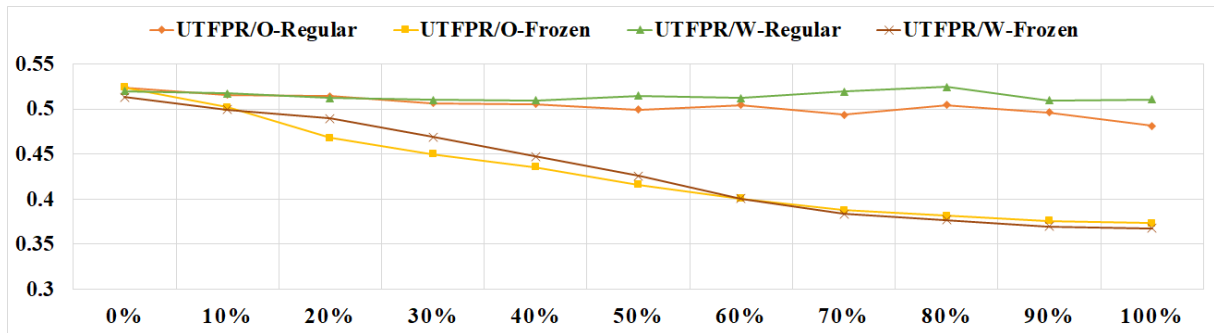
Figure 2: F-scores of our robustness experiments for English. The horizontal axis represents the proportion of noisy words in the input sentence, and the vertical axis the F-scores.
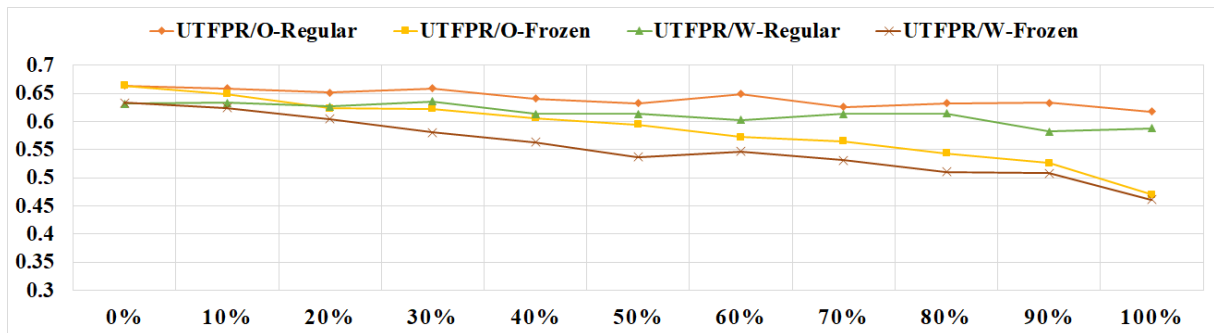

Figure 3: F-scores of our robustness experiments for Spanish. The horizontal axis represents the proportion of noisy words in the input sentence, and the vertical axis the F-scores.

test set a little. We believe that this was caused because the development set was more representative of the trial than the test set. Overall, submitting UTFPR/W instead of UTFPR/O cost us 2 ranks for English and 3 for Spanish.

5.2 Robustness Assessment

In order to test the robustness of the UTFPR systems, we had to generate different noisy versions of the test set with increasing volumes of noise artificially added to them.

To do so, we introduced a modification to $N\%$ of randomly selected words in each sentence in the datasets. The modifications could be either the deletion of a randomly selected character (50% chance) or its duplication (50% chance). We used $0 \leq N \leq 100$ in intervals of 10, resulting in a total of 11 increasingly noisy versions. The next step was to create "frozen" versions of the UTFPR models that act as if any word out of the training set's vocabulary is unknown. If a word of the test set is not present in the vocabulary of the training set, it produces a numerical vector full of 1's that represents an out-of-vocabulary word.

Figures 2 and 3 show the results obtained for English and Spanish, respectively. As it can be

noticed, our compositional models are much more robust than the frozen alternatives, suffering very faint losses in F-score even when 100% of the words in the input sentence are noisy.

6 Conclusions

In this contribution, we presented the UTFPR systems submitted to the HatEval 2019 shared task. The systems are based on compositional RNN models trained exclusively over the training data provided by the organizers. We introduced two variants of our models: one trained entirely on the shared task's data (UTFPR/O), and another with a set of pre-trained character-to-word RNN layers fine-tuned to the task at hand (UTFPR/W). Our results show that, despite its simplicity, the UTFPR/O model attained competitive results for English, placing it 7th out of 62 submissions. Furthermore, the results of this shared task indicate that our models are very robust, being able to handle even substantially noisy inputs. In the future, we intend to test more reliable ways of re-using pre-trained compositional models.

Acknowledgements

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.

Pete Burnap and Matthew L Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242.

Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving cyberbullying detection with user context. In *Advances in Information Retrieval*, pages 693–696. Springer.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of ICWSM*.

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate Lingo: A Target-based Linguistic Analysis of Hate Speech in Social Media. *arXiv preprint arXiv:1804.04257*.

Paula Fortuna and Sérgio Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys (CSUR)*, 51(4):85.

Björn Gambäck and Utpal Kumar Sikdar. 2017. Using Convolutional Neural Networks to Classify Hate-speech. In *Proceedings of the First Workshop on Abusive Language Online*, pages 85–90.

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 Cyberbulling (TRAC)*, Santa Fe, USA.

Wang Ling, Chris Dyer, Alan W Black, Isabel Trancoso, Ramon Fermandez, Silvio Amir, Luis Marujo, and Tiago Luis. 2015. Finding function in form: Compositional character models for open vocabulary word representation. In *Proceedings of the 2015 EMNLP*, pages 1520–1530. Association for Computational Linguistics.

Shervin Malmasi and Marcos Zampieri. 2017. Detecting Hate Speech in Social Media. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP)*, pages 467–472.

Shervin Malmasi and Marcos Zampieri. 2018. Challenges in Discriminating Profanity from Hate Speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30:1–16.

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*, pages 145–153. International World Wide Web Conferences Steering Committee.

Gustavo Paetzold. 2018. Utfpr at iest 2018: Exploring character-to-word composition for emotion analysis. In *Proceedings of the 9th EMNLP*, pages 176–181. Association for Computational Linguistics.

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*, pages 1–10, Valencia, Spain.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Langauge Online*.

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*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of NAACL*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval)*.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network. In *Lecture Notes in Computer Science*. Springer Verlag.