UTFPR at SemEval-2019 Task 6: Relying on Compositionality to Find Offense

Gustavo Henrique Paetzold
Universidade Tecnológica Federal do Paraná / Toledo-PR, Brazil
ghpaetzold@utfpr.edu.br

Abstract

We present the UTFPR system for the OffensEval shared task of SemEval 2019: A character-to-word-to-sentence compositional RNN model trained exclusively over the training data provided by the organizers. We find that, although not very competitive for the task at hand, it offers a robust solution to the orthographic irregularity inherent to tweets.

1 Introduction

Text classification tasks can take a wide variety of forms, and some of them, such as sentiment and emotion analysis, have managed to grab a lot of attention from researchers in recent years. More recently, however, the public's growing engagement in debates on the topics of free speech and politics has led the Natural Language Processing (NLP) and Machine Learning (ML) communities to take an interest in classification tasks related to identifying and categorizing patterns of profanity, hate speech and offense. The prominence of shared tasks held on these topics, such as those of Fersini et al. (2018) and Wiegand et al. (2018), are great examples of the research community coming together to attempt to create more reliable solutions for these challenges.

While hate speech is commonly characterized by specific slurs and other offensive expressions that convey prejudice against a certain group or individual, offensive speech is often more challenging to identify because it encompasses a more broad spectrum of language, featuring expressions that do not necessarily convey prejudice (Malmasi and Zampieri, 2018). Technologies that identify these types of patterns could, for instance, help a social media platform on profiling users and take appropriate action whenever someone breaks user agreements and/or terms of use.

Identifying offensive language within the content of social media platforms is particularly challenging, since this type of content is usually littered with irregular orthography, meta-characters, slang and others. Since a lot of the effort from the research community focuses on identifying offensive language in social media platforms, an NLP approach for such a task must be able to overcome those hurdles in some way. Some of the preferred methods for handling the orthographic irregularity of social media content are using word embeddings trained over tweets (Rozental et al., 2018) or regularizing unusual spellings (Bertaglia and das Graças Volpe Nunes, 2017), but neither of them ensure that every possible orthographically irregular word will be understood by the NLP model in question. Recently, however, there have been a lot of contributions that present compositional neural models that learn numerical representations of words based on the sequence of characters that compose them (Kim et al., 2016; Ling et al., 2015; Balazs et al., 2018; Paetzold, 2018). These models have been demonstrated to be both effective in text classification, and robust when faced with orthographic irregularity.

In this paper, we present the UTFPR system submitted to the OffensEval shared task of SemEval 2019, which employs compositional neural models to identify offensive language in tweets. In the following sections we describe the task (section 2), our model (section 3) and experiments (sections 4 to 5).

2 Task Summary

The UTFPR systems described herein are a contribution to the OffensEval shared task held at the SemEval 2019 workshop (Zampieri et al., 2019b). In this shared task, participants were tasked with creating innovative classifiers capable of identify-

Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pages 801–805
Minneapolis, Minnesota, USA, June 6–7, 2019.

ing and categorizing offensive tweets. This shared tasks has 3 sub-tasks:

- **Task A:** Binary classification task that consists in judging whether a tweet is offensive or not.
- **Task B:** Binary classification task that consists in identifying whether or not an offensive tweet was targeted towards a specific person or group.
- **Task C:** Consists in identifying whether an offensive tweet was targeted at a person, group or something else (3-class classification).

We decided to focus our efforts on Task A exclusively. The organizers provided participants a training set with $13,240$ instances, a trial set with 320, and a test set with 860. Each instance is composed of a tweet and its respective labels for tasks A, B and C. The datasets were annotated by humans of undisclosed background (Zampieri et al., 2019a).

3 The UTFPR Model

As we have previously mentioned, ours is a compositional Recurrent Neural Network (RNN) inspired by the ones introduced by Ling et al. (2015) and Paetzold (2018). Our RNN learns word representations based on the sequence of characters that compose them, then learns sentence representations based on the word representations previously learned. Figure 1 illustrates the architecture of our model in detail.

The model takes as input a potentially offensive tweet. It first produces character embeddings for the characters of each word in the sentence, then passes them through a sequence of bidirectional RNN layers in order to produce character-to-word numerical representations for them. These word representations are then passed onto another sequence of bidirectional RNN layers, which in turn produce a single word-to-sentence numerical representation for the sentence. A dense layer connected to a softmax layer produces the final binary class, which can be OFF (for offensive) and NOT (for not offensive).

Because the dataset provided for training is rather small, we suspected that the character-to-word representations produced through this training data would not be reliable enough for the task. Because of that, we decided to train two different model variants:

- **UTFPR-Scratch:** The model depicted in Figure 1 trained from scratch over the shared task's training set exclusively.
- **UTFPR-Reuse:** The same model depicted in Figure 1, except instead of training its character-to-word RNN layers from scratch along with the rest of the model, they are taken from a similar compositional model pre-trained by Paetzold (2018) over a much larger dataset for the Emotion Analysis shared task held at WASSA 2018 (Klinger et al., 2018). The training set of the WASSA 2018 shared task has $153,383$ instances, each composed of a tweet with a target emotion word replaced with a [#TRIGGERWORD#] marker, and an emotion label that could be either joy, sad, disgust, anger, surprise, or fear.

The architecture of our models is identical, and their specifications are:

- **Size of character embeddings:** 15
- **RNN layer type:** Gated Recurrent Units
- **RNN layer depth:** 2 (for all layers sets)
- **RNN layer size:** 60 (for all layers)
- **Dropout proportion:** 25%
- **Loss function:** Cross-entropy
- **Framework used:** PyTorch[1]

We chose to use the PyTorch framework due to the fact that it employs dynamic computational graphs, and hence they do not require us to set a fixed maximum size for the words in the dataset. This feature of PyTorch only allows us to create a much more flexible model that can handle any word size, but also disregards the needs for padding.

To train our models, we split the training set into a training portion ($10,000$ instances) and a development portion ($3,240$ instances). The models were left training for hundreds of iterations, and after each iteration a version of each model was saved. The final selected models were the ones with the lowest attained error on the development

[1] https://pytorch.org

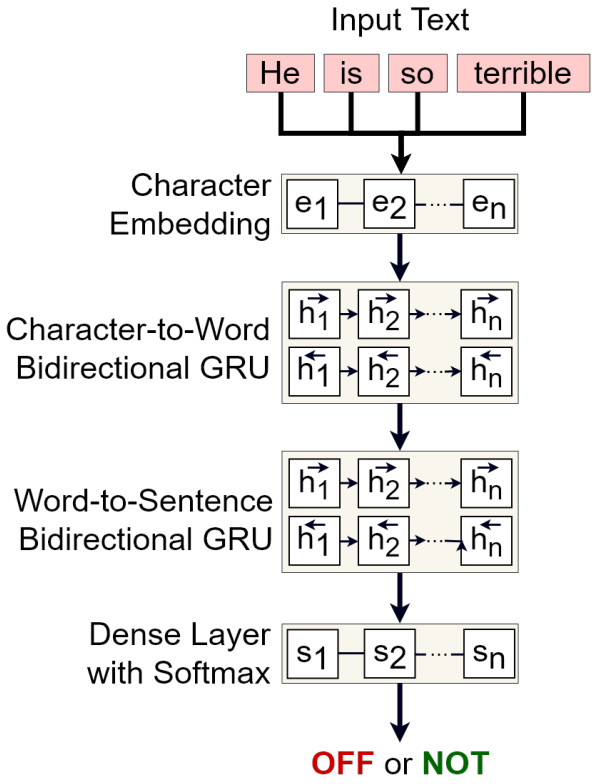

Figure 1: Architecture of the UTFPR system.

portion of the data. We conducted a preliminary evaluation over the trial set to determine which of the variants to submit. The macro-averaged F-scores, which are illustrated in Table 2, show that using pre-trained character-to-word RNN layers actually compromised the performance of our model in this instance, hence we opted to submit the UTFPR-Scratch variant.

System	F-score
UTFPR-Scratch	0.770
UTFPR-Reuse	0.599

Table 1: Macro-averaged F-scores for the trial set

4 Performance on Shared Task

The systems submitted to the shared task were evaluated through their macro-averaged F1-score. The results on Table 2 showcase the results obtained by UTFPR-Scratch, as well as the top 3 and bottom 3 systems submitted to Task A. As it can be noticed, our system did not perform very well, placing 93rd out of 103 teams. The confusion matrix of UTFPR-Reuse in Figure 3 shows that the main reason behind this poor showing was the large amount of false negatives predicted. Upon inspecting the labels predicted, we found that the UTFPR-Scratch system would predict offense mostly for tweets with a lot of profanity and with hashtags associated with the Donald Trump administration, such as "#BuildTheWall".

Rank	System	F-score
1	pliu19	0.829
2	anikolov	0.815
3	lukez	0.814
93	UTFPR	0.528
101	hamadanayel	0.458
102	magnito60	0.422
103	AyushS	0.171

Table 2: Macro-averaged F-scores for the trial set

5 Robustness Assessment

As we've already mentioned, one of the main advantages of compositional RNN models that learn word representations from character sequences is the fact that they handle low-frequency and out-of-vocabulary words in an elegant way, since they are able to produce a numerical representation for any word. Because of that, these models tend to be much more resilient when presented with noisy

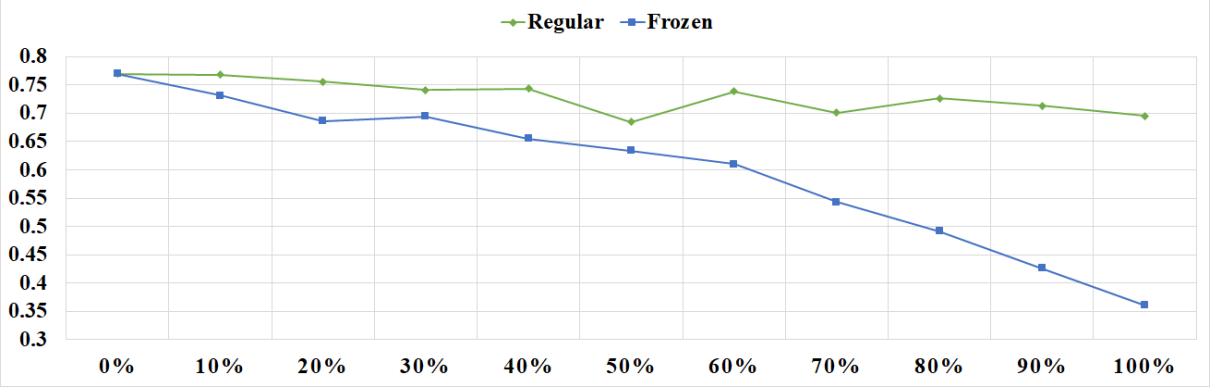

Figure 2: Results of our robustness experiments. The vertical and horizontal axes presents macro-averaged F-scores and the percentage of words with noise introduced to them, respectively. The dots represent the scores obtained by the regular UTFPR-Scratch model and a frozen version that treats all words outside of the training set as out-of-vocabulary words.

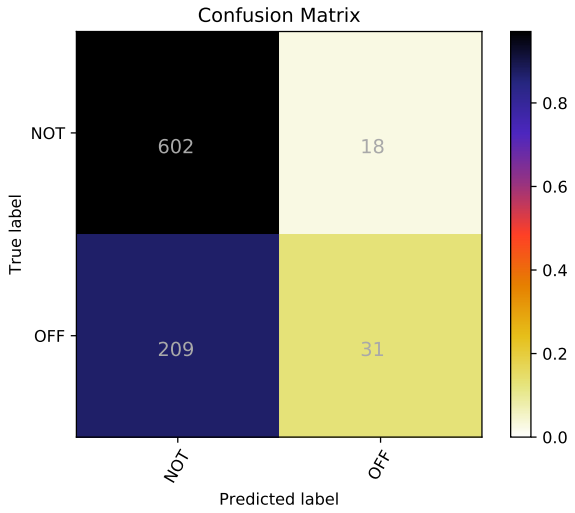

Figure 3: Confusion matrix of the UTFPR-Scratch model on the test set.

input that differs considerably from the input presented during training.

In this experiment, we assess the robustness of our UTFPR models. First, we generated "jammed" versions of the shared task's trial set (since the test set was not made available) with increasing volumes of noise introduced to them. To create a jammed test set, we simply added a noise-inducing modification to $N\%$ randomly selected words of each sentence. The modifications were randomly selected between either removing a randomly selected character form the word (50% chance) or duplicating it (50% chance). We created 11 jammed versions by using $0 \leq N \leq 100$ in intervals of 10. Words with a single character that were subjected to removal were simply discarded from the sentence.

We compared the regular UTFPR-Scratch model (Regular) with a modified version with frozen character-to-word RNN layers (Frozen). The frozen version only produces a numerical representation of a word if it is present in the training set, otherwise, it produces a vector full of 1's signaling an out-of-vocabulary word. The results in Figure 2 show that using the frozen version is much less robust than the regular model, specially when the input sentence has 70% or more of its words out of the training set vocabulary.

6 Conclusions

In this paper we presented the UTFPR system for the OffensEval shared task held at SemEval 2019, which is a compositional RNN model that learns numerical representations of words based on its characters. Our experiments reveal that, although our model is not very competitive for this task specifically (placing 93rd out of 103 participants), it offers a very robust solution to the problem of out-of-vocabulary words. Inspecting the model's output we found that the main cause for its poor performance was the fact that it learned a bias towards the "not offensive" label, which caused it to predict a lot of false negatives. Also, we found that our model was actually better at identifying profanity and controversial topics rather than offense itself. In the future, we intend to explore combining our numerical word representations with richer semantic features in order to train more reliable compositional models for this task.

7 Acknowledgments

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

References

Jorge Balazs, Edison Marrese-Taylor, and Yutaka Matsuo. 2018. Iiidyt at iest 2018: Implicit emotion classification with deep contextualized word representations. In *Proceedings of the 9th WASSA*, pages 50–56.

Thales Felipe Costa Bertaglia and Maria das Graças Volpe Nunes. 2017. Exploring word embeddings for unsupervised textual user-generated content normalization. *CoRR*, abs/1704.02963.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the evalita 2018 task on automatic misogyny identification (ami). In *Proceedings of the 6th EVALITA*.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2016. Character-aware neural language models. In *Proceedings of the 2016 AAAI*, pages 2741–2749.

Roman Klinger, Orphee De Clercq, Saif Mohammad, and Alexandra Balahur. 2018. Iest: Wassa-2018 implicit emotions shared task. In *Proceedings of the 9th WASSA*, pages 31–42.

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.

Shervin Malmasi and Marcos Zampieri. 2018. Challenges in discriminating profanity from hate speech. *Journal of Experimental & Theoretical Artificial Intelligence*, 30(2):187–202.

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

Alon Rozental, Daniel Fleischer, and Zohar Kelrich. 2018. Amobee at iest 2018: Transfer learning from language models. In *Proceedings of the 9th WASSA*, pages 43–49.

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*.

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)*.