UTFPR at IEST 2018:
Exploring Character-to-Word Composition for Emotion Analysis

Gustavo H. Paetzold
Federal University of Technology - Paraná / Brazil
ghpaetzold@utfpr.edu.br

Abstract

We introduce the UTFPR system for the Implicit Emotions Shared Task of 2018: A compositional character-to-word recurrent neural network that does not exploit heavy and/or hard-to-obtain resources. We find that our approach can outperform multiple baselines, and offers an elegant and effective solution to the problem of orthographic variance in tweets.

1 Introduction

Emotion analysis has become one of the most prominent tasks in Natural Language Processing (NLP) in recent years. It can be framed as either a regression task, where one wants to gauge the degree of some emotion, such as how optimistic a certain opinion is with respect to a given matter, or a classification task, where one wants to decide on which type of emotion is being conveyed, such as happiness, fear, anger, etc. The task is particularly interesting for industry applications, since it can potentially allow for institutions to automatically assess the public opinion on things like initiatives, products, etc. The task can also be applied in many interesting natural language domains. For instance, one can try to determine whether a certain restaurant review posted in a major news outlet favors the establishment, or whether a certain tweet about a celebrity or institution conveys support or disdain.

The type of target domain can greatly influence how an emotion analysis system is structured. If the targets are formally written and well revised articles from newspapers and magazines, then one can expect to find only orthographically correct words and appropriately structured sentences in the input. The authors of tweets, on the other hand, often use many distinct orthographic variants of the same word (ex: you, u, youu), tend to have less regard for form, and use non-textual symbols to express meaning, such as emojis. Also, articles tend to be much longer than tweets, which have a size limit of just a few hundred characters. Systems for the later type of domain must address a lot of challenges that systems for the former do not have to, which can compel them to be much more complex.

The SeerNet system (Duppada et al., 2018), one of the best performing systems of the SemEval 2018 shared task on affect in tweets (Mohammad et al., 2018), is a great example of that. In this shared task, participants were asked to create both regression (for emotion intensity) and classification (for emotion decision) systems for emotion analysis in English, Arabic, and Spanish. In order to overcome the challenge of analyzing tweets, the SeerNet system resorts to a wide range of specialized resources, such as special tokenizers and embedding models for tweets, emoji analyzers, and even off-the-shelf systems trained on large amounts of curated data. Though undoubtedly effective, the SeerNet has a very complex architecture that would be difficult to replicate, specially for under-resourced languages.

In an effort to offer a simpler solution to emotion analysis in tweets, we present the UTFPR system submitted to Implicit Emotions Shared Task (IEST) of 2018 (Klinger et al., 2018). Ours is a character-to-word recurrent neural network architecture that offers an elegant solution to orthographic variance within tweets, and does not rely

Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 176–181
Brussels, Belgium, October 31, 2018. ©2018 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

on any resources other than the input provided by the shared task organizers. We describe our approach in what follows.

2 Task Description

The UTFPR system is a contribution to the IEST 2018 shared task, hosted at the 9th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA 2018). In this shared task, participants were tasked with classifying tweets with respect to the emotion they convey.

The task organizers provided participants a training set composed of $153,383$ instances, a trial set with $9,591$ instances, and an unlabeled test set with $28,757$ instances. Each instance is composed of a tweet with a target emotion word replaced with a [#TRIGGERWORD#] marker, and an emotion label. There are six possible emotions in this setup: joy, sad, disgust, anger, surprise, and fear.

The organizers also provided a wide array of external resources that could complement the systems created, such as emotion dictionaries, lexicons, datasets from previous tasks, etc.

3 Preliminary Experiment

Before conceiving the final version of the UTFPR system, we conducted a preliminary experiment with baseline classification models in order to test some model design options, and hence guide the creation of the UTFPR approach. More specifically, we assessed two design options with respect to input:

- **Structure:** Since each instance contains an omitted target emotion word, we tested whether it is more productive to address the entire tweet as a bag of words, or to individually model the words to the left and right of the target.
- **Enhancement:** We also tested whether or not it is helpful to complement the training set with data gathered in unsupervised fashion.

3.1 Experimental Setup

In this section, we delve into the details of how our preliminary experiment was structured.

Data: We used 90% of the training data from the shared task for training, and 10% for testing. Notice that we did not use the trial data for testing because the labels had not been made available at the time this experiment was conducted.

Models: We tested three types of machine learning models; logistic regression, decision trees, and random forests. All these models were implemented with the help of scikit-learn[1].

Input Features: We tried two types of features; TF-IDF weights from a bag-of-words model trained over our input training data (TF-IDF), and the average word embedding values of the words in the tweet (Embeddings). We used the 300-dimension word embeddings model of Paetzold and Specia (2016), which was trained using the CBOW model (Mikolov et al., 2013) over a corpus of $\tilde{7}$ billion words from assorted sources, such as news articles, subtitles, tweets, etc.

Input Structure: We tested two types of inputs to the models; one in which we calculate and concatenate two separate feature representations of the words to the left and right of the target (Separate), and another in which we calculate only one feature representation of all words in the tweet aside from the target (Joint).

Input Enhancement: We tested two variants of each model; one trained only on our training data (TR), and another trained on the training data plus a set of $1,774,423$ automatically extracted complementary instances (TR+E). To produce the complementary instances we first extracted all morphological variants and synonyms of the words "joy", "sad", "disgust", "anger", "surprise", and "fear", then looked for sentences containing these words in the same 7 billion word corpus used to train our embeddings. Finally, we replaced the emotion word in each sentence with [#TRIGGERWORD#], and assigned the appropriate emotion label to the instance.

3.2 Preliminary Results

The macro F-score obtained by each variant tested is featured in Table 1. The results reveal that, across almost all scenarios, modeling the words to the left and right of the target (Separate) without data enhancement (TR) yields the most promising results.

[1] http://scikit-learn.org

	TF-IDF				Embeddings			
	Joint		Separate		Joint		Separate	
	TR	TR+E	TR	TR+E	TR	TR+E	TR	TR+E
Logistic Regression	0.270	0.406	0.277	0.220	0.139	0.095	0.160	0.058
Decision Trees	0.201	0.165	0.213	0.191	0.112	0.101	0.155	0.145
Random Forests	0.228	0.186	0.243	0.219	0.122	0.096	0.177	0.166

Table 1: Preliminary experiment results. Each cell represents the macro F-score obtained by a given model.

4 The UTFPR System

The UTFPR system is a compositional character-to-word recurrent neural network model that attempts to address the challenges of working with tweets in an elegant way. Guided by the findings from our preliminary experiment, we structured our as illustrated in Figure 1.

First, the words to the left and right of the target word, which we henceforth refer to as left and right contexts, are passed through a character embedding layer. The embeddings are then passed onto a bidirectional GRU network that produce a numerical representation for each word in the sentence. The representations of the words in the left and right contexts are then passed on to two separate bidirectional GRU networks, which in turn produce a final representation of each context. Finally, these representations are concatenated and passed on to a final perceptron layer, which predicts the most likely emotion based on softmax.

The model uses no external resources other than the sentence itself as input, and because it is a neural model, it can be configured with respect to the size of embeddings, type of recurrent layers used, number of layers, activation function, etc. We describe our experiments and configuration of the UTFPR system in what follows.

5 Experimental Setup

For our experiments, we configure the UTFPR system as follows:

- **Character embedding size:** 25
- **RNN layer type:** Gated Recurrent Units (GRU)
- **RNN layer depth:** 2
- **RNN layer size:** 50
- **Dropout proportion:** 50%
- **Loss function:** Cross-entropy
- **Framework used:** PyTorch[2]

As mentioned in section 2, we submited the UTFPR system to the IEST 2018 shared task. We trained the UTFPR system over the entire training set provided by the organizers, and validated it over the trial set. Our final submission was the model resulting from iteration with the lowest cross-entropy error on the trial set.

In order to offer some points of comparison and highlight the importance of some design decisions made when creating UTFPR, we trained two other variants of UTFPR:

- **UTFPR-C**: A version of UTFPR without the character-to-word layers. Instead, it uses as input word embeddings extracted from the word embeddings model described in section 3.
- **UTFPR-CD**: A version of UTFPR-C trained without dropout.

Due to the limited amount of computing resources available to us, we were not able to train any more variants of UTFPR. We also include in our performance comparison all the baseline models described in section 3, the baselines provided by the IEST 2018 organizers, and the 5 systems with the highest macro F-scores in the shared task.

6 Results

Table 2 showcases the micro and macro Precision, Recall, and F-scores of our UTFPR variants, as well the IEST 2018 baselines and top 5 systems. Although our approach did not manage to reach the top of the leaderboards, the results do highlight the impact of some design decisions made when creating the final UTFPR system. As it can be noticed, incorporating dropout and adding a character-to-word encoder to our model slightly increases its performance. While the complete UTFPR system

[2] https://pytorch.org

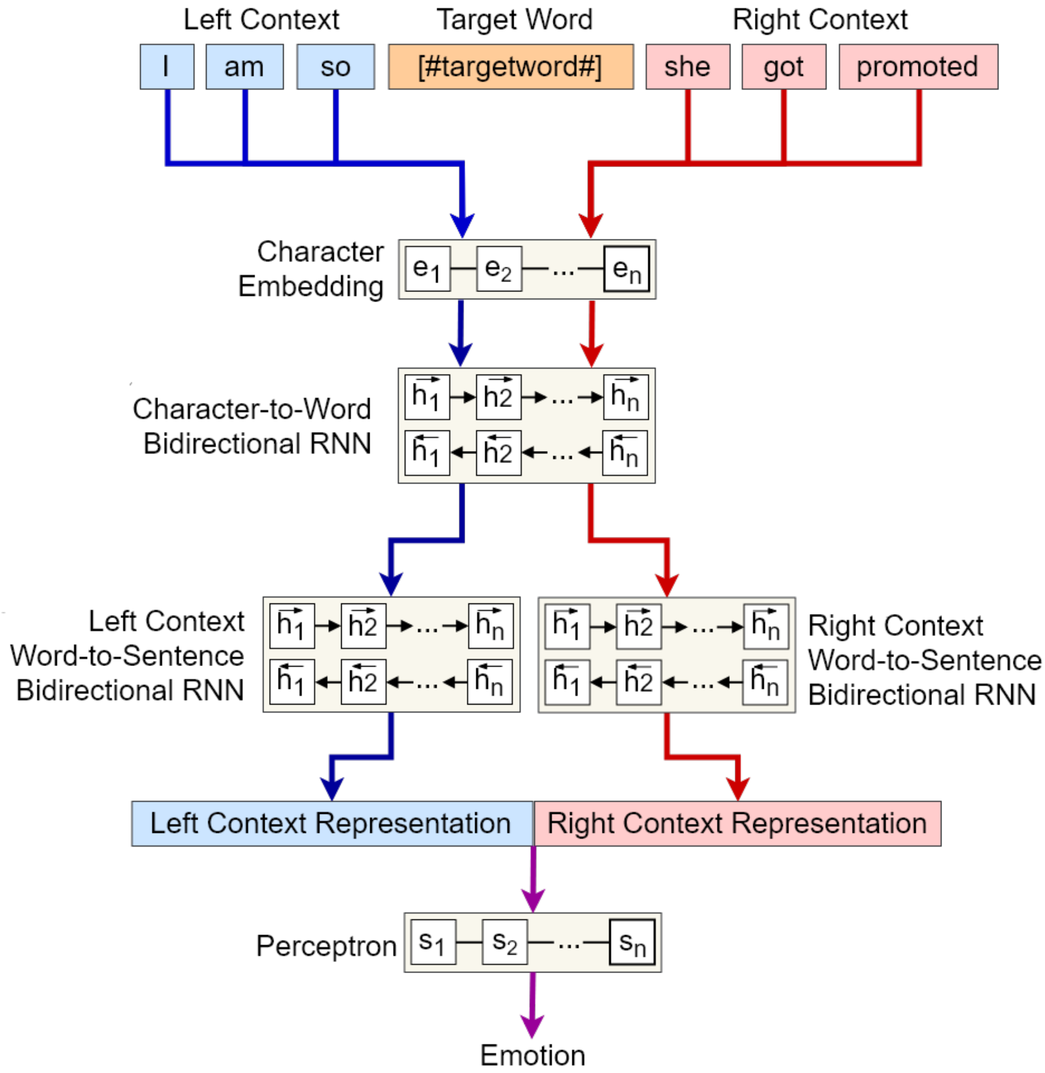
Left Context
Target Word
Right Context
I
am
so
[#targetword#]
she
got
promoted
Character Embedding
e_1 — e_2 — ... — e_n
Character-to-Word Bidirectional RNN
$\overrightarrow{h}_1$ → $\overrightarrow{h}2$ → ... → $\overrightarrow{h}_n$
$\overleftarrow{h}_1$ ← $\overleftarrow{h}2$ ← ... ← $\overleftarrow{h}_n$
Left Context Word-to-Sentence Bidirectional RNN
Right Context Word-to-Sentence Bidirectional RNN
Left Context Representation
Right Context Representation
Perceptron
s_1 — s_2 — ... — s_n
Emotion

Figure 1: Architecture of the UTFPR system

	Micro			Macro		
	P	R	F	P	R	F
Amobee	-	-	-	-	-	0.714
IIIDYT	-	-	-	-	-	0.710
NTUA-SLP	-	-	-	-	-	0.703
UBC-NLP	-	-	-	-	-	0.693
Sentylic	-	-	-	-	-	0.692
BOW MaxEnt	-	-	-	-	-	0.599
“Joy” Always	-	-	-	-	-	0.051
UTFPR-CD	0.541	0.541	0.541	0.550	0.544	0.541
UTFPR-C	0.545	0.545	0.545	0.551	0.546	0.545
UTFPR	**0.568**	**0.568**	**0.568**	**0.575**	**0.569**	**0.569**

Table 2: Official micro and macro scores obtained by the UTFPR systems. Bold-case scores showcase the highest scores obtained by the UTFPR systems. The first five lines showcase the scores for the top 5 IEST 2018 systems, and the following two the ones for the IEST 2018 baselines.

	TF-IDF				Embeddings			
	Joint		Separate		Joint		Separate	
	TR	TR+E	TR	TR+E	TR	TR+E	TR	TR+E
Logistic Regression	0.496	0.403	0.500	0.406	0.255	0.093	0.293	0.104
Decision Trees	0.369	0.304	0.387	0.351	0.211	0.188	0.284	0.267
Random Forests	0.418	0.346	0.449	0.405	0.232	0.182	0.332	0.310

Table 3: Macro F-scores obtained by our baseline models on the official IEST 2018 test set.

managed to place 23rd in the competition, either of its two variants would place 26th.

Table 3 features the macro F-score results obtained by the baseline systems described in section 3 on the IEST 2018 test set. The results are consistent with the ones in Table 1, which brings some reassurance with respect to the usefulness of our preliminary experiment. Nonetheless, none of the baseline systems managed to outperform the more elaborate UTFPR systems.

7 Analysis

In addition to our performance comparison, we also conducted complementary analyses that allowed us to delve into the merits and limitations of the UTFPR system. First we produced its confusion matrix, which is illustrated in Table 4. Although the mistakes made by the UTFPR system are well spread out throughout the matrix, it can be noticed that, despite the labels in the reference set being present in even proportions, the UTFPR model is slightly biased towards the anger and surprise labels.

Finally, we conducted an experiment comparing the robustness of the three UTFPR variants described in section 5 (UTFPR, UTFPR-C, UTFPR-CD). For this experiment, we first created an orthographically “jammed” version of the IEST 2018 test set. For each tweet in the test set, we randomly selected 75% of its words, and then either duplicated (50% chance) or removed (50% chance) a randomly selected letter. We then trained the UTFPR variants on the regular IEST 2018 training and trial set, and tested them over our jammed test set.

The results in Table 5 show that adding the compositional character-to-word encoder to our model greatly increases its robustness with respect to orthographic variance. While jamming the words cost the UTFPR-C and UTFPR-CD variants upwards of $14,3\%$ in macro F-score, the performance of our complete UTFPR system only dropped by $2,2\%$.

8 Conclusions

In this contribution, we introduced the UTFPR emotion analysis system for the IEST 2018 shared task. Unlike current state-of-the-art approaches, our model does not rely on external resources, and employs instead a single compositional recurrent neural network that learns representations of sentences based on its words, and of words based on its characters.

Through our experiments we found that, although the UTFPR system cannot compete with more elaborate, resource-heavy approaches, it does offer a promising solution to the task that is very robust to orthographic variance. In the future, we aim to create more sophisticated variants of the UTFPR approach that incorporate other cost-effective sources of information to better inform the model and hence increase its performance.

9 Acknowledgments

We would like to thank the Federal University of Technology - Paraná for supporting this contribution.

References

Duppada, Venkatesh, Royal Jain, and Sushant Hiray. 2018. Seernet at semeval-2018 task 1: Domain adaptation for affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 18–23. Association for Computational Linguistics.

Klinger, Roman, Orphée de Clercq, Saif M. Mohammad, and Alexandra Balahur. 2018. Iest: Wassa-2018 implicit emotions shared task. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

	joy	**sad**	**disgust**	**anger**	**surprise**	**fear**	**% Ref.**
joy	2790	595	295	646	563	357	18.24%
sad	343	2582	451	467	316	181	15.09%
disgust	203	524	2784	429	677	177	16.67%
anger	399	417	435	2530	686	327	16.67%
surprise	290	296	478	442	2971	315	16.66%
fear	323	326	275	523	660	2684	16.66%
% Pred.	15.12%	16.48%	16.41%	17.52%	20.42%	14.05%	-

Table 4: Confusion matrix of the UTFPR system. Lines represent reference labels and columns represent predictions. The last column and line feature the occurrence proportion of each emotion in the reference and predicted label set, respectively.

	Micro			Macro		
	P	R	F	P	R	F
UTFPR-CD	0.400	0.400	0.400	0.414	0.403	0.398
UTFPR-C	0.408	0.408	0.408	0.414	0.407	0.404
UTFPR	**0.546**	**0.546**	**0.546**	**0.552**	**0.547**	**0.547**

Table 5: Official micro and macro scores obtained by the UTFPR systems on the jammed test set. Bold-case scores showcase the highest scores obtained by the UTFPR systems.

Mohammad, Saif, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17. Association for Computational Linguistics.

Paetzold, Gustavo H. and Lucia Specia. 2016. Unsupervised lexical simplification for non-native speakers. In *Proceedings of the 13th AAAI Conference on Artificial Intelligence*, pages 3761–3767.