USI-IR at IEST 2018: Sequence Modeling and Pseudo-Relevance Feedback for Implicit Emotion Detection

Esteban A. Ríssola, Anastasia Giachanou and Fabio Crestani
Faculty of Informatics
Università della Svizzera italiana (USI)
{esteban.andres.rissola, anastasia.giachanou, fabio.crestani}@usi.ch

Abstract

This paper describes the participation of USI-IR in WASSA 2018 Implicit Emotion Shared Task. We propose a relevance feedback approach employing a sequential model (biLSTM) and word embeddings derived from a large collection of tweets. To this end, we assume that the top-*k* predictions produce at a first classification step are correct (based on the model accuracy) and use them as new examples to re-train the network.

1 Introduction

Recent years have seen the rapid growth of social media platforms (*e.g.*, Facebook, Twitter, several blogs) that has changed the way that people communicate. Many people express their opinion and emotions on blogs, forums or microblogs. Detecting the emotions that are expressed in social media is a very important problem for a wide variety of applications. For example, enterprises can detect complains of customers about their products or services and act promptly.

Emotion detection aims at identifying various emotions from text. According to (Plutchik, 1980) there are eight basic emotions: anger, joy, sadness, fear, trust, surprise, disgust and anticipation. Considering the abundance of opinions and emotions expressed in microblogs, emotion and sentiment analysis in Twitter has attracted the interest of the research community (Giachanou and Crestani, 2016). In particular, *Implicit Emotion Shared Task (IEST)* is a shared task by WASSA 2018 that focuses on emotion analysis. In this task, participants are asked to develop tools that can predict the emotions in tweets from which a certain emotion word is removed. This is a very challenging problem since the emotion analysis needs to be done without access to an explicit mention of an emotion word and consequently taking advantage of context that surrounds the target word.

In this paper, we describe our submitted system to the *IEST: WASSA-2018 Implicit Emotion Shared Task*. Our system is based on a bidirectional Long Short-Term Memory (biLSTM) network on top of word embeddings which is later inserted in a pseudo-relevance feedback schema. Our results show that even though the model still need more refinement it offers interesting capabilities to address the task at hand.

2 Dataset

To train our model, we employ the dataset provided within the shared task. It is worth mentioning that no other external datasets are used during the training and development phases. There are roughly 153K tweets in the training set, 10K in the development set and 28K in the test set. Each data instance includes the tweet and the emotion class of the word which has been extracted from the text. The test set's golden labels were provided only after the evaluation period. The complete description of the dataset can be found in (Klinger et al., 2018)

3 Proposed Approach

In recent years, Recurrent Neural Networks (RNN) have risen in popularity among different NLP tasks (Mikolov et al., 2010; Graves et al., 2013; Filippova et al., 2015). This success can be attributed to their inherent ability to capture temporal information and learn features directly from the data. In other words, the time based sequentially connected structure of these networks is intuitive to use for sequential inputs, such as sentences or words. For this reason, we decide to model the tweets employing bidirectional Long Short-Term Memory (biLSTM) networks (Graves and Schmidhuber, 2005), which are an alternative RNN architecture that incorporates additional

Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 231–234

https://doi.org/10.18653/v1/P17

structures, called gates, to better control the information across sequential inputs and deal with issues that may arise during training, like the vanishing gradient problem. Moreover, each training sequence is presented forwards and backwards and the output combined at each timestep allowing to improve the overall performance of the network.

In the context of Information Retrieval (IR), relevance feedback refers to a technique designed to refine a query, either automatically or through user interaction. The goal of this process is to construct a query that is a better representation of the information need, and therefore to retrieve better documents (Manning et al., 2008). In particular, pseudo-relevance feedback automates the manual part of relevance feedback as the system simply assumes that the initial top-ranked documents are relevant and uses them to produce a new result set. In order to increase the accuracy of the biLSTM network we develop a pseudo-relevance feedback schema where we assume that the top-k predictions produce at a first classification step are correct (based on the model accuracy) and use them as new examples to re-train the network.

4 Experimental Setup

Preprocessing and tokenization are crucial steps of the pipeline involved in the development of a model: the output produced has an immediate effect in the features learned by the model. This task could turn to be particularly challenging in Twitter since the vocabulary results quite unstable over time and the way that users expressed does not follow traditional patterns. In order to preprocess and tokenize the collection of tweets we employ a python library[1] developed for that purpose which applies different regular expressions to extract particular units, such as hashtags, and separates them from the rest of the tokens. We only conserve words, hashtags, mentions, emojis and smileys. The remaining tokens outside these categories are discarded given that their inclusion did not prove to be useful for the task. Some examples of such tokens are URLs (they were originally replaced with `http://url.removed`), numbers and the unit `[NEWLINE]`. Furthermore, we remove the hash symbols from the hashtags and split the words when possible using the Viterbi algorithm. The prior probabilities are obtained from word statistics from Google Ngram corpus. In particular, we observed a positive impact on the training accuracy of the network. One possible reason could be that the terms contained in the hashtags were probably present in the word embeddings but not as a singular unit (*e.g., #classyCouple*). All the tokens are transformed to lowercase and words which were completely in capitals, emulating a yell in the social media language, were doubled. It is important to remark that stopwords are not removed.

The model is comprised of an embedding layer, a biLSTM layer and a softmax layer. It receives as input a tokenized twitter message treated as a sequence of words. Since the length of different tweets can vary, we set the length of each message to 99 (the maximum message length across training and development data according to the operations performed in the preprocessing step). Tweets that are shorter than this length are zero-padded. It should be noted that the network will ignore everything that goes beyond the last word in the text, *i.e.*, the padding. The weights of the embedding layer are initialized using word2vec (Mikolov et al., 2013) embeddings trained on 400 million tweets (Godin et al., 2015) from the ACL W-NUT share task (Baldwin et al., 2015). We also tried to use the 300-dimensional pre-trained vector trained on Google News dataset[2] combined with emojis pre-trained embeddings (Eisner et al., 2016). However, the performance was slightly worse and for that reason we decided not to employ them. Words out of the embeddings are conserved, albeit their weights are randomly initialize and learn from scratch. A single biLSTM layer with a hidden layer size of 128 neurons follows in the architecture and feeds a softmax layer in order to obtain the final prediction. The network parameters are learned by minimizing Cross-Entropy and by backpropagating the error through layers over 5 epochs, with a batch size of 128, using RMSprop optimization algorithm. Moreover, a dropout rate (Srivastava et al., 2014) of 0.5 is used to address overfitting issues. The aforementioned model was implemented in Python using Tensorflow library (Abadi et al., 2015).

Lastly, as introduced in Section 3, we propose a pseudo-relevance feedback scheme as follows: (a) A first instance of the network is trained using the training and development sets; (b) Sub-

[1] See: https://github.com/s/preprocessor

[2] See: https://code.google.com/archive/p/word2vec/

Emotion	I	you	shehe	adverb	posemo	insight	cause	focuspresent	focusfuture	swear
Anger	7.43	2.21	**2.24**	8.53	2.65	1.9	**3.55**	14.78	1.08	0.85
Disgust	6.63	2.33	1.58	8.98	2.7	**2.54**	2.8	15.15	0.73	**1.08**
Fear	**9.36**	2.64	1.82	7.1	2.69	2.03	2.68	15.95	**2.95**	0.49
Joy	9.11	**3.4**	1.69	**9.54**	**4.58**	1.78	2.94	15.59	1.43	0.47
Sadness	7.35	2.56	1.38	8.85	3.35	2.18	2.95	**16.37**	1.49	0.54
Surprise	7.13	1.95	1.96	8.56	3.1	2.09	2.94	13.18	0.89	0.58

Table 1: LIWC selected categories for the six emotions. The values represent percentages over total words.

sequently, the k percentage of the tweets with the highest class probability is extracted with the corresponding labels to create a new set examples; (c) Finally, the training and development sets along with the new examples are used to re-train the model from scratch and obtain the final predictions. It should be noted that the same hyperparameters are employed at both training and retraining steps.

5 Results and Discussion

Overall, we observe that the effectiveness obtained on the test set by the proposed approach is not as satisfactory as expected (see Table 2). As can be noticed the performance diminishes as the relevance set used to re-trained the network increases in size. One of the reason could be that even though the majority of the new examples are accurately classified the remaining ones correspond to misclassification errors. Consequently, this might introduce certain noise and affect the performance of the model. One possible way to overcome this issue could be to define a threshold and select the cases whose class probability exceeds this limit, instead of just taking the k percentage with the highest chance of being correctly classified. Another reason, could reside in the fact that different training parameters, like the number of epochs, should be again optimized given that the size and content of the new training data has changed.

In addition to the previously mentioned pseudo-relevance feedback schema, we also explore the use of the information provided by a tool known as Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al.). This software is equipped with a series of dictionaries which allows to obtain different psychometric properties that may arise from language use. More specifically, it was developed by psychologists with focus on studying the various emotional, cognitive, and structural components present in individuals' verbal and written speech samples. Interesting findings arose after analyzing the training set with this tool. In particular, the selected categories showed noticeable differences between the tweets expressing different emotions (See Table 1). For instance, the use of the word *I* in the tweets expressing fear, which could correlate with the fact that individuals tend to refer more to themselves when they expressed the perceived danger, threat or even concerns, or the high use of she/he pronouns and causation words (because, hence, thus) when conveying anger. We attempted to incorporate this information to our model by repeating the words in the tweets that were included in a set of selected dictionaries for each emotion. It is clear that the chosen dictionaries should not overlap so as to emphasize the differences among the emotions. Nonetheless, only in the training (and development) set the labels are known in advance, and consequently only these instances can be expanded in this way. The results obtained by following with this approach were not promising and for that reason, we decided to expand the tweets using the dictionaries disregarding of the emotion. Given that the improvement was almost marginal we decided not to include the results in the paper.

k%	%Correct	MacAvg
10	63.72	0.60
20	64.08	0.59
30	64.22	0.58
40	64.35	0.57
50	64.13	0.58

Table 2: Pseudo-relevance feedback schema results on the test set

6 Conclusions

In this work, we introduced a relevance feedback schema employing a sequential model (biLSTM) in order to predict the class of a certain emotion

that has been removed from a tweet. Despite the fact that the performance did not improve as expected, we consider that the method still needs further improvement. For instance, by employing a probability threshold to create a more accurate expansion set. Furthermore, we would like to continue exploring different ways to incorporate LIWC's output to the network. Promising features can be extracted from the presented analysis which might allow to emphasize the differences between the emotions conveyed in the tweets.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Timothy Baldwin, Marie Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. ACL Association for Computational Linguistics.

Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bosnjak, and Sebastian Riedel. 2016. emoji2vec: Learning emoji representations from their description. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 48–54. Association for Computational Linguistics.

Katja Filippova, Enrique Alfonseca, Carlos A. Colmenares, Lukasz Kaiser, and Oriol Vinyals. 2015. Sentence compression by deletion with lstms. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 360–368. Association for Computational Linguistics.

Anastasia Giachanou and Fabio Crestani. 2016. Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys*, 49(2):28.

Fréderic Godin, Baptist Vandersmissen, Wesley De Neve, and Rik Van de Walle. 2015. Multimedia lab @ acl w-nut ner sharedtask: named entity recognition for twitter microposts using distributed word representations. In *ACL 2015 Workshop on Noisy User-generated Text, Proceedings*. Association for Computational Linguistics.

A. Graves, N. Jaitly, and A. r. Mohamed. 2013. Hybrid speech recognition with deep bidirectional lstm. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*.

A. Graves and J. Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2047–2052 vol. 4.

Roman Klinger, 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*, Brussels, Belgium. Association for Computational Linguistics.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *ICLR Worshop*, Scottsdale, AZ, USA.

Tomas Mikolov, Martin Karafit, Luks Burget, Jan Cernock, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH*, pages 1045–1048. ISCA.

James W. Pennebaker, Cindy K. Chung, Molly Ireland, Amy Gonzales, and Roger J. Booth. The development and psychometric properties of liwc2015. Technical report, The University of Texas at Austin.

Robert Plutchik. 1980. Emotion: Theory, research, and experience: Vol. 1. theories of emotion. In R. Plutchik and H. Kellerman, editors, *Approaches to Emotion*, pages 3–33. Academic press.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.