Attention mechanism for aggressive detection

Carlos Enrique Muñiz Cuza[1], Gretel Liz De la Peña Sarracén[1], Paolo Rosso[2]

[1]Center for Pattern Recognition and Data Mining, Cuba
{carlos,gretel}@cerpamid.co.cu
[2]PRHLT, Universitat Politècnica de València, Spain
prosso@dsic.upv.es

Abstract. This paper describes the system we developed for IberEval 2018 on Aggressive detection track of Authorship and Aggressiveness Analysis on Twitter (MEX-A3T)[1]. The task focuses on the detection of aggressive comments in tweets that come from Mexican users. Systems must be able to determine whether a tweet is aggressive or not. Our approach is an Attention-based Long Short-Term Memory Network Recurrent Neural Network where the attention layer helps to calculate the contribution of each word towards targeted aggressive classes. In particular, we build a Bidireccional LSTM to extract information from the word embeddings over the sentence, then apply attention over the hidden states to estimate the importance of each word and finally feed this context vector to another LSTM model to estimate whether the tweet is aggressive or not. The experimental results show that our model achieves outstanding results.

Keywords: Deep Learning, Attention-based Neural Network, LSTM Model, Aggressive Detection Track, Twitter

1 Introduction

Recurrent Neural Networks (RNNs) is a type of deep neural network designed for sequence modeling. These kinds of models are greatly studied due to their flexibility in capturing nonlinear relationships. However, traditional RNNs suffer from problems known as exploding or vanishing gradients and, therefore, have difficulty in capturing long-term dependencies. Long Short-Term Memory networks (LSTM) [1] are one of the most used in Natural Language Processing (NLP) to overcome this limitation. They are able to learn the dependencies in lengths of considerably large chains.

Moreover, attention models have become an effective mechanism to obtain better results [2–6]. In [7], the authors use a hierarchical attention network for document classification. The model has two levels of attention mechanisms applied at the word and sentence-level, enabling it to attend differentially to more and less important content when constructing the document representation. The

[1] https://mexa3t.wixsite.com/home

experiments show that the architecture outperforms previous methods by a substantial margin.

In this paper, we propose a similar Attention-based LSTM for the IberEval 2018 track on Authorship and aggressiveness analysis in Twitter (MEX-A3T) [8]. The attention layer is applied on the top of a Bidirectional LSTM to generate a context vector for each word embedding which is then fed to another LSTM network to detect whether the tweet is aggressive or not. To the best of our knowledge, there has been no other work exploring the use of attention-based architectures for the task.

The task focuses on the detection of aggressive comments. This is a topic that has not been widely studied in the community. The aim is to determine whether a tweet, which comes from Mexican users, is aggressive or not.

The paper is organized as follows. Section 2 describes our system. Experimental results are then discussed in Section 3. Finally, we present our conclusions with a summary of our findings in Section 4.

2 System

2.1 Preprocessing

In the preprocessing step, the tweets are cleaned. Firstly, the emoticons are recognized and replaced by corresponding words that express the sentiment they convey. Also, we remove all links and urls. Afterwards, tweets are morphologically analyzed by FreeLing [9]. In this way, for each resulting token, its lemma is assigned. Then, the tweets are represented as vectors with a word embedding model. This model was generated by using the word2vec algorithm [10] from the Wikipedia collection in Spanish.

2.2 Method

We propose a model that consists of a Bidirectional LSTM neural network (Bi-LSTM) at the word level. At each time step t the Bi-LSTM gets as input a word vector w_t with syntactic and semantic information, known as word embedding [10]. Afterward, an attention layer is applied over each hidden state h_t. The attention weights are learned using the concatenation of the current hidden state h_t of the Bi-LSTM and the past hidden state s_{t-1} of a Post-Attention LSTM (Pos-Att-LSTM). Finally, the target aggressiveness of the tweet is predicted by this final Pos-Att-LSTM network.

2.3 Bidireccional LSTM Recurrent Neural Network

In NLP problems, standard LSTM receives sequentially (left to right order) at each time step a word embedding w_t and produces a hidden state h_t. Each hidden state h_t is calculated as follow:

$$
\begin{aligned}
i_t &= \sigma(W^{(i)}x_t + U^{(i)}h_{t-1} + b^{(i)}) && \text{(input gate)}\\
f_t &= \sigma(W^{(f)}x_t + U^{(f)}h_{t-1} + b^{(f)}) && \text{(forget gate)}\\
o_t &= \sigma(W^{(o)}x_t + U^{(i)}h_{t-1} + b^{(o)}) && \text{(output gate)}\\
u_t &= \sigma(W^{(u)}x_t + U^{(u)}h_{t-1} + b^{(u)}) && \text{(new memory cell)}\\
c_t &= i_t \otimes u_t + f_t \otimes c_{t-1} && \text{(final memory cell)}\\
h_t &= o_t \otimes tanh(c_t)
\end{aligned}
$$

Where all W_*, U_* and b_* are parameters to be learned during training. The function σ is the sigmoid function and $\otimes$ stands for element-wise multiplication.

The bidirectional LSTM, on the other hand, makes the same operations as standard LSTM but, processes the incoming text in a left-to-right and a right-to-left order in parallel. Thus, the output is a two hidden state at each time step $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$.

The proposed method uses a Bidirectional LSTM network which considers each new hidden state as the concatenation of these two $\hat{h_t} = [\overrightarrow{h_t}, \overleftarrow{h_t}]$. The idea of this Bi-LSTM is to capture long-range and backwards dependencies.

2.4 Attention Layer

With an attention mechanism we allow the Bi-LSTM to decide which part of the sentence should "attend". Importantly, we let the model learn what to attend on the basis of the input sentence and what it has produced so far.

Let $H \in R^{2*N_h \times T_x}$ the matrix of hidden states $[\hat{h_1}, \hat{h_2}, ..., \hat{h_{T_x}}]$ produced by the Bi-LSTM, where N_h is the size of the hidden state and T_x is the length of the given sentence. The goal is then to derive a context vector c_t that captures relevant information and feed it as input to the next level (Pos-Att-LSTM). Each c_t is calculate as follow:

$$
c_t = \sum_{t'=1}^{T_x} \alpha_{t,t'} \hat{h_{t'}} \qquad \alpha_{t,i} = \frac{\beta_{t,i}}{\sum_{j=1}^{T_x} \beta_{t,j}} \qquad \beta = tanh(W_a * [\hat{h_t}, s_{t-1}] + b_a)
$$

Where W_a and b_a are the trainable attention parameters, s_{t-1} is the past hidden state of the Pos-Att-LSTM and $\hat{h_t}$ is the current hidden state. The idea of the concatenation layer is to take into account not only the input sentence but also the past hidden state to produce the attention weights.

2.5 Post-Attention LSTM

The goal of the Post-Att-LSTM is to predict whether the tweet is aggressive or not. This network at each time step receives the context vector c_t which is propagated until the final hidden state s_{T_x}. This vector is a high level representation of the tweet and is used in the final softmax layer as follow:

$$\hat{y} = softmax(W_g * s_{T_x} + b_g)$$

Where W_q and b_q are the parameters for the softmax layer. Finally, cross entropy is used as the loss function, which is defined as:

$$L = -\sum_i y_i * log(\hat{y}_i) \quad (1)$$

Where y_i is the true classification of the tweet.

3 Results

Table 1 shows the results obtained by the proposed method on the aggressive class for two different runs (run1 and run2). A variation in the model was realized for run2, where a linguistic characteristic is added to tweets. This characteristic is based on the study carried out in the work [11], where the authors propose a methodology for the detection of obscene and vulgar phrases in Mexican tweets. In this way, the characteristic defines the presence or not of obscene or vulgar words in the tweets according to the resource generated by the cited work. These results reveal that the linguistic characteristic incorporated in the second run marked performance improvement based on F-measure aggressive class, reaching the second position of the ranking among all the participants in the task.

Table 1. Performance on the testing set

Run	F-measure	Precision	Recall
run 1	0.3091	0.5724	0.2117
run 2	0.45	0.3815	0.5485

4 Conclusion

We propose an Attention-based Long Short-Term Memory Network Recurrent Neural Network for the IberEval 2018 track on aggressive detection in Twitter. The model consists of a bidirectional LSTM neural network with an attention mechanism that allows to estimate the importance of each word and then, this context vector is used with another LSTM model to estimate whether the tweet is aggressive or not. The results showed that the use of a linguistic characteristic based on the occurrence of obscene or vulgar phrases in the tweets allows to improve the F.measure of the aggressive class.

Acknowledgments. The work of the third author was partially supported by the SomEMBED TIN2015-71147-C2-1-P MINECO research project.

References

1. Hochreiter, Sepp, and Jürgen Schmidhuber. Long Short-Term Memory. Neural Computation 9(8), pp. 1735–1780. (1997).
2. Yang, Min and Tu, Wenting and Wang, Jingxuan and Xu, Fei and Chen, Xiaojun. Attention Based LSTM for Target Dependent Sentiment Classification. In AAAI Conference on Artificial Intelligence. pp. 5013–5014. (2017).
3. Zhang, Yu and Zhang, Pengyuan and Yan, Yonghong. Attention-based LSTM with Multi-task Learning for Distant Speech Recognition. Proc. Interspeech 2017. pp. 3857–3861. (2017).
4. Wang, Yequan and Huang, Minlie and Zhao, Li and Zhu, Xiaoyan. Attention-based LSTM for Aspect-level Sentiment Classification. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 606–615. (2016).
5. Lin, Kai and Lin, Dazhen and Cao, Donglin. Sentiment Analysis Model Based on Structure Attention Mechanism. In UK Workshop on Computational Intelligence. pp. 17–27. Springer, Cham (2017).
6. Rush, Alexander M and Chopra, Sumit and Weston, Jason. A Neural Attention Model for Abstractive Sentence Summarization. arXiv preprint arXiv:1509.00685. (2015).
7. Yang, Zichao and Yang, Diyi and Dyer, Chris and He, Xiaodong and Smola, Alex and Hovy, Eduard. Hierarchical Attention Networks for Document Classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 1480–1489. (2016).
8. Álvarez-Carmona, Miguel Á and Guzmán-Falcón, Estefanía and Montes-y-Gómez, Manuel and Escalante, Hugo Jair and Villaseñor-Pineda, Luis and Reyes-Meza, Verónica and Rico-Sulayes, Antonio. Overview of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in Mexican Spanish tweets. Notebook Papers of 3rd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL), Seville, Spain, September. (2018).
9. Padró, Lluís and Stanilovsky, Evgeny. Freeling 3.0: Towards Wider Multilinguality. Proceedings of the Language Resources and Evaluation Conference (LREC 2012). (2013).
10. Mikolov, Tomas and Sutskever, Ilya and Chen, Kai and Corrado, Greg S and Dean, Jeff. Distributed Representations of Words and Phrases and their Compositionality. In Advances in Neural Information Processing Systems. pp. 3111–3119. (2013).
11. Guzmán, Estefania and Beltrán, Beatriz and Tovar, Mireya and Vázquez, Andrés and Martínez, Rodolfo. Clasificación de Frases Obscenas o Vulgares dentro de Tweets. Research in Computing Science. 85, pp. 65–74. (2014).