Automatic Misogyny Identification Using Neural Networks

I. Goenaga, A. Atutxa, K. Gojenola, A. Casillas, A. Díaz de Ilarraza, N. Ezeiza, M. Oronoz, A. Pérez and O. Perez de Viñaspre

IXA Group - University of the Basque Country (EHU/UPV)
Ixa Taldea EHU/UPV Informatika Fakultatea M. Lardizabal 1 20008 Donostia
http://ixa.si.ehu.es/
iakesg@gmail.com

Abstract. In this paper we present our approach to automatically identify misogyny in Twitter tweets. That task is one of the two sub-tasks organized by AMI-IberEval 2018 organization. In order to carry out the task, we present a neural network approach. Neural network models have been demonstrated to be capable of achieving remarkable performance in sentence and document modeling. Convolutional neural network (CNN) and recurrent neural network (RNN) are two mainstream architectures for such modeling tasks, which adopt totally different ways of understanding natural languages. In this work we focus on RNN approach using a Bidirectional Long Short Term Memory (Bi-LSTM) with Conditional Random Fields (CRF) and we evaluate the proposed architecture on misogyny identification task (text classification). The experimental results show that the system can achieve good performance on this task obtaining 78.9 accuracy on English tweets and 76.8 accuracy on Spanish tweets.

Keywords: Shared task · Misogyny · Neural Networks.

1 Introduction

In the last couple of years we have started to see deep learning making significant inroads into areas where computers have previously seen limited success. Rather than requiring a set of fixed rules that are defined by the programmer, deep learning uses neural networks that learn rich non-linear relationships directly from data. Deep learning has also seen some success in NLP, for example in text classification. Text classification is an essential component in many applications, such as web searching, information filtering, and sentiment analysis [3].

A key problem in text classification is feature representation, which is commonly based on the bag-of-words (BoW) model, where unigrams, bigrams, n-grams or some specific patterns are extracted as features. Moreover, several feature selection methods, such as pLSA [4] or LDA [5] are applied to select more discriminative features. Nevertheless, traditional feature representation methods often have problems when they try to capture the semantics of the words because they ignore contextual information. This is a problem in text classification

because contextual information is the key in order to correctly classify a text. Although high-order n-grams and more complex features are designed to capture more contextual information and word orders, the data sparsity problem remains, which heavily affects the classification accuracy.

In a recurrent neural network approach, the models analyze a text word by word and store the semantics of all the previous text in a fixed-sized hidden layer [6]. They receive as input a sequence of vectors and return another sequence that represents some information about the sequence at every step in the input. Although RNNs can learn long dependencies, they often fail to do so and tend to be biased towards their most recent inputs in the sequence [8]. Likewise, Long Short-term Memory Networks (LSTMs) incorporate a memory-cell and have been shown effective capturing long-range dependencies.

Classic LSTMs create the representation of each word of the sentence using only the left context. Is interesting to use also the right context if we want to create a more complete representation of the words, though. This can be done with a second LSTM that reads the same sequence in reverse. This type of LSTMs are named bidirectional LSTMs (BI-LSTMs) [9] and they create the representation of the words concatenating the left representation and the right representation. These representations effectively include a representation of a word in context, which is useful for numerous tasks.

On the other hand, Conditional Random Fields (CRF) are a probabilistic framework for labeling and segmenting structured data, such as sequences and trees. The underlying idea is that of defining a conditional probability distribution over label sequences given a particular observation sequence, rather than a joint distribution over both label and observation sequences. The primary advantage of CRFs is the relaxation of the independence assumption. Independence assumption states that the variables do not depend on each other and they do not affect each other in any way and this is not always the case and, consequently, it can lead to serious inaccuracies. Likewise, CRFs have been shown really effective in different tasks such as POS tagging [7], text processing [10] or computer vision [11].

Taking that into account, in this work we have employed a BI-LSTM with Conditional Random Fields (CRF) [7] in order to prove its effectiveness in misogynous tweet identification. In this area, one of the last works is [1] where the authors address the problem of automatic detection and categorization of misogynous language in online social media, and they set the bases to organize AMI-IberEval 2018 shared task [2].

In the rest of this paper we will first present the experimental setup we have used to carry out our experiments in section 2, followed by the results obtained in the shared task test set in section 3 and the main conclusions of the work in section 4.

2 System Description

We have divided this section into three subsections. In subsection 2.1 we explain the preprocessors we have used to tokenize and normalize the tweets, in subsection 2.2 the data resources we have employed in addition to the data shared by the organization, while in subsection 2.3 we focus on the used system.

2.1 Preprocessors

We have made use of one of python's packages designed for preprocessing tweets [12]. The tool performs tokenization, word normalization, word segmentation (for splitting hash-tags) and spell correction, using word statistics from 2 big corpora (English Wikipedia, twitter - 330mil English tweets). In addition, for Spanish we have used a set of simple rules proposed in [13] for spell correction.

2.2 External Data

Our AMI System needs word-embeddings in order to create a better word representation for each word we find in the corpus. Thus, we have used word-embeddings extracted from the Spanish Billion Word Corpus [14] and from Wikipedia 2014 and Gigaword 5 [15].

2.3 AMI System

In order to classify the tweets we have employed a neural network based architecture, more precisely a specific Bi-LSTM (an RNN subclass) with a CRF on top of it as proposed in [7]. This kind of neural network is widely used to pursue sequence to sequence tagging. One of the advantages of using Bi-LSTM in contrast to other machine learning techniques such as SVM, Perceptron or CRFs is that the size of the context is automatically learned by the LSTM and there is no need to perform any complicated text preprocessing to obtain features to feed the tool. As we mentioned previously, our system is a tagger and marks the beginning and the next words of the sequences (IOB) we want to label. In this case we want to predict whether a tweet contains misogynous content or not. Thus, we introduce the tweets and the word-embeddings at the beginning of the process as in [7]. When a word is missing in the word-embeddings, the system replaces the word with unknown (UNK) label.

In all cases the system returns every word of each tweet tagged with *Yes* label when the tweet contains misogyny and with *No* label otherwise. If the opposite happened, we would consider a tweet as misogynous if at least has one *Yes* label. The examples below are the output of the system for two tweets written in English and represent the aforementioned:

[**B-Yes**]< *user* > [**I-Yes**]bitch [**I-Yes**]is [**I-Yes**]a [**I-Yes**]psyco [**I-Yes**], [**I-Yes**]* [**I-Yes**]dry [**I-Yes**]pussy [**I-Yes**]detected [**I-Yes**]*

[**B-No**]you [**I-No**]give [**I-No**]me [**I-No**]life [**I-No**]! [**I-No**]< *repeated* > [**I-No**]< *url* >

3 Results and Discussion

In AMI-IberEval 2018 shared task the participants can try their misogynous content identification systems in two languages: English and Spanish. We have participated in both languages and we have included the results for English track in table 1 and the results for Spanish track in table 2.

English								
R	**Team**	**Accu**	**R**	**Team**	**Accu**	**R**	**Team**	**Accu**
1	14-exlab-r1	91.3	**11**	resham-r1	78.5	**21**	JoseSebastian-r1	74.9
2	14-exlab-r2	90.2	**12**	AMI-Baseline	78.3	**22**	Amrita_CEN-r3	73.8
3	14-exlab-r4	89.8	**13**	_vic_-r2	78.0	**23**	_vic_-r1	70.9
4	14-exlab-r3	87.8	**14**	_vic_-r3	78.0	**24**	ITT-r1	70.6
5	SB-r4	87.0	**15**	_vic_-r4	78.0	**25**	_vic_-r5	64.6
6	SB-r5	85.1	**16**	maybelraul-r3	77.9	**26**	Amrita_CEN-r2	56.3
7	14-exlab-r5	82.3	**17**	maybelraul-r1	77.1	**27**	GrCML2016-r3	52.7
8	AnotherOne-r1	79.3	**18**	maybelraul-r4	76.9	**28**	GrCML2016-r2	52.4
9	maybelraul-r2	79.3	**19**	maybelraul-r5	76.0	**29**	Amrita_CEN-r1	51.9
10	**ixaTeam-r1**	**78.9**	**20**	ITT-r2	75.8			

Table 1. Results obtained by participants for English track using only provided training data (constrained).

If we analyze the results for English track, we observe that our position within all participants of the shared task is tenth with 78.9 of accuracy. Although we are far from winning the shared task (- 12.4), we are in the first third of the classification and the two previous systems in the classification are not far (+ 0.4) from us which demonstrates a good performance of our system identifying misogynous tweets in English.

Spanish								
R	**Team**	**Accu**	**R**	**Team**	**Accu**	**R**	**Team**	**Accu**
1	14-exlab-r3	81.4	**10**	SB-r3	80.5	**19**	maybelraul-r1	76.7
2	JoseSebastian-r1	81.4	**11**	SB-r1	80.3	**20**	_vic_-r2	76.6
3	SB-r4	81.3	**12**	AnotherOne-r1	80.2	**21**	Amrita_CEN-r3	74.4
4	14-exlab-r1	81.2	**13**	maybelraul-r5	79.6	**22**	_vic_-r3	65.9
5	14-exlab-r2	81.2	**14**	maybelraul-r2	78.8	**23**	Amrita_CEN-r1	54.2
6	14-exlab-r4	80.9	**15**	maybelraul-r3	78.7	**24**	14-exlab-r5	53.6
7	SB-r2	80.8	**16**	maybelraul-r4	78.2	**25**	Amrita_CEN-r2	52.9
8	SB-r5	80.6	**17**	**ixaTeam-r1**	**76.8**			
9	_vic_-r1	80.5	**18**	AMI-BASELINE	76.7			

Table 2. Results obtained by participants for Spanish track using only provided training data (constrained).

On the other hand, our system's accuracy identifying Spanish written tweets is 76.8. This time our position within the all participants is seventeenth just above the shared task's baseline. However, almost all the participants have obtained accuracies between 81.4 and 76.6 which indicates that the vast majority of the systems are close to each other. Likewise, identifying misogynous content in Spanish written tweets is more difficult mostly because the lack of top quality resources (corpus, word-embeddings, preprocessors ...) we can find relatively easy for English.

Once we analyzed our system's results for both languages, taking into account our system was designed for sequential tagging or sequence labeling we consider the experimental setup has performed well in a task it was not thought for. We realize the best option to do text classification would have been a convolutional neural network (CNN) specially because the best systems of the state of the art employ this type of neural networks. Nevertheless, our main purpose has been to test a BI-LSTM with CRF on text classification task and bearing in mind its constraints the system has achieved reasonable results.

4 Conclusions

This paper presents our approach to automatically identify misogynous content in Twitter tweets. In order to carry out the task, we have chosen a neural network approach due to their ability to achieving remarkable performance in sentence and document modeling. In this work we focus on RNN approach using a Bidirectional Long Short Term Memory (Bi-LSTM) with Conditional Random Fields (CRF) and the experimental results show that the system can achieve good performance identifying misogynous tweets obtaining 78.9 accuracy on English tweets and 76.8 accuracy on Spanish tweets.

Acknowledgments

This work has been partially funded by:

- The Spanish ministry (projects TADEEP: TIN2015-70214-P, PROSA-MED: TIN2016-77820-C3-1-R).
- The Basque Government (projects DETEAMI: 2014111003, ELKAROLA:KK-2015/00098).

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

References

1. M. Anzovino, E. Fersini, P. Rosso. 2018. Automatic Identification and Classification of Misogynistic Language on Twitter. In: M. Silberztein, F. Atigui, E. Kornyshova, E. Métais, F. Meziane. (eds) Natural Language Processing and Information Systems. NLDB 2018. Lecture Notes in Computer Science, vol 10859.

2. E. Fersini, M. Anzovino, P. Rosso. Overview of the Task on Automatic Misogyny Identification at IberEval. In: Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018). CEUR Workshop Proceedings. CEUR-WS.org, Seville, Spain, September 18, 2018.
3. S. Kiritchenko, S. Mohammad, and M. Salameh. 2016. SemEval-2016 Task 7: Determining sentiment intensity of English and Arabic phrases. In Proceedings of the 10th International Workshop on Semantic Evaluation. San Diego, California, USA, SemEval '16, pages 42–51.
4. L. Cai, and T. Hofmann. 2003. Text categorization by boosting automatically extracted concepts. In SIGIR, 182–189.
5. S. Hingmire, S. Chougule, G. K. Palshikar and S. Chakraborti. 2013. Document classification by topic labeling. In SIGIR, 877–880.
6. J. L. Elman. Finding structure in time. 1990. Cognitive science 14(2):179–211.
7. G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami and C. Dyer. 2016. Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360.
8. Y. Bengio, P. Simard, and P. Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. Neural Networks, IEEE Transactions on, 5(2):157–166.
9. A. Graves and J. Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM networks.In Proc. IJCNN.
10. B. Settles. 2005. Abner: an open source tool for automatically tagging genes, proteins, and other entity names in text. Bioinformatics, 21(14):3191–3192.
11. X. He, R. S. Zemel and M. A. Carreira-Perpinian. 2004. Multiscale conditional random fields for image labelling. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition.
12. C. Baziotis, N. Pelekis and C. Doulkeridis. 2017. DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis. Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pages 747-754.
13. I. Alegria, I. Etxeberria and G. Labaka. 2013. Una Cascada de Transductores Simples para Normalizar Tweets. Tweet-Norm@ SEPLN.
14. C. Cardellino. 2016. Spanish Billion Words Corpus and Embeddings. http://crscardellino.me/SBWCE/
15. J. Pennington, R. Socher and C. D. Manning. 2014. GloVe: Global Vectors for Word Representation.