Tweetaneuse @ AMI EVALITA2018: Character-based Models for the Automatic Misogyny Identification Task (Short Paper)

Davide Buscaldi
LIPN, Université Paris 13
Sorbonne Paris Cité
99, Avenue Jean-Baptiste Clément
93430 Villetaneuse (France)
buscaldi@lipn.fr

Abstract

English. This paper presents the participation of the RCLN team with the Tweetaneuse system to the AMI task at Evalita 2018. Our participation was focused on the use of language-independent, character-based methods.

Italiano. *Quest'articolo presenta la partecipazione del team RCLN con il sistema Tweetaneuse al challenge AMI di Evalita 2018. La nostra partecipazione era orientata sull'utilizzo di metodi multilingue e basati sui caratteri.*

1 Introduction

The language used on social media and especially Twitter is particularly noisy. The reasons are various; among them, the abuse of abbreviations induced by the limitations on the size of the messages, and the use of different ways to refer to the same event or concept, strengthened by the availability of hashtags (for instance: *World Cup in Russia*, *#WorldCup2018*, *#WC18* all refer to the same event).

Recently, some character-level neural network based models have been developed to take into account these problems for tasks such as sentiment analysis (Zhang et al., 2017) or other classification tasks (Yang et al., 2016). Another advantage of these methods, apart the robustness to the noisy text that can be found in tweets, is that they are completely language independent and they don't need lexical information to carry out the classification task.

The Automatic Misogyny Identification task at Evalita2018 (Fersini et al., 2018) presented an interesting and novel challenge. Misogyny is a type of hate speech that targets specifically women in different ways. The language used in such messages is characterised by the use of profanities, specific hashtags, threats and other intimidating language. This task is an ideal test bed for character-based models, and (Anzovino et al., 2018) already reported that character n-grams play an important role in the misogyny identification task.

We participated to the French Sentiment Analysis challenge DEFT 2018 (Paroubek et al., 2018) earlier this year with language-independent character-based models, based both on neural networks and classic machine learning algorithms. For our participation to AMI@Evalita2018 our objective was to verify whether the same models could be applied to this task while keeping a comparable accuracy.

The rest of the paper is structured as follows: in Section 2 we describe the two methods that were developed for the challenge; in Section 3 we present and discuss the obtained results, and finally in Section 4 we draw some conclusions about our experience and participation to the AMI challenge.

2 Methods

2.1 Locally-weighted Bag-of-Ngrams

This method is based on a Random Forest (RF) classifier (Breiman, 2001) with character n-grams features, scored on the basis of their relative position in the tweet. One of the first parameters to choose was the size of the n-grams to work with. According to our previous experience, we chose to use all the character n-grams (excluding spaces) of size 3 to 6, with a minimum frequency of 5 in the training corpus.

The weight of each n-gram n in tweet t is calculated as:

$$s(n,t) = \begin{cases} 0 & if absent \\ \sum_{i=1}^{occ(n,t)} 1 + \frac{pos(n_i)}{len(t)} & otherwise \end{cases}$$

where $occ(n,t)$ is the number of occurences of n-gram n in t, $pos(n_i)$ indicates the position of the first character of the i-th occurrence of the n-gram n and $len(t)$ is the length of the tweet as number of characters. The hypothesis behind the use of this positional scoring scheme is that the presence of some words (or symbols) at the end or the beginning of a tweet may be more important that the mere presence of the symbol. For instance, in some cases the conclusion is more important than the first part of the sentence, especially when people are evaluating different aspects of an item or they have mixed feelings: *I liked the screen, but the battery duration is horrible*.

2.2 Char and Word-level bi-LSTM

This method was only tested before and after the participation, since we observed that it performed worse than the Random Forest method.

In this method we use a recurrent neural network to implement a LSTM classifier (Hochreiter and Schmidhuber, 1997), which are now widely used in Natural Language Processing. The classification is carried out in three steps:

First, the text is split on spaces. Every resulting text fragment is read as a character sequence, first from left to right, then from right to left, by two recurrent NN at character level. The vectors obtained after the training phase are summed up to provide a character-based representation of the fragment (compositional representation). For a character sequence $s = c_1 \dots c_m$, we compute for each position $h_i = LSTM_o(h_{i-1}, e(c_i))$ et $h'_i = LSTM_{o'}(h'_{i+1}, e(c_i))$, where e is the embedding function, and $LSTM$ indicates a LSTM recurrent node. The fragment compositional representation is then $c(s) = h_m + h'_1$.

Subsequently, the sequence of fragments (i.e., the sentence) is read again from left to right and vice versa by other two recurrent NNs at word level. These RNNs take as input the compositional representation obtained in the previous step for the fragments to which a vectorial representation is concatenated. This vectorial representation is obtained from the training corpus and is considered only if the textual fragment has a frequence ≥ 10. For a sequence of textual fragments $p = s_1 \dots s_n$, we calculate $l_i = LSTM_m(l_{i-1}, c(s_i) + e(s_i))$, $l'_i = LSTM_{m'}(l_{i+1}, c(s_i) + e(s_i))$, where c is the compositional representation introduced above and e the embedding function. The final states obtained after the bi-directional reading are added and they are required to represent the input sentence, $r(p) = l_n + l'_1$.

Finally, these vectors are used as input to a multi-layer perceptron which is responsible for the final classification: $o(p) = \sigma(O \times \max(0, (W \times r(p) + b)))$, where σ is the *softmax* operator, W, O are matrices and b a vector. The output is interpreted as a probability distribution on the tweets' categories.

The size of character embeddings is 16, those of the text fragments 32, the input layer for the perceptron is of size 64 and the hidden layer 32. The output layer is size 2 for subtask A and 6 for subtask B. We used the DYNET[1] library.

3 Results

We report in Table 1 the results on the development set for the two methods.

Italian		
Subtask	**lw RF**	**bi-LSTM**
Misogyny identification	0.891	0.872
Behaviour classification	0.692	0.770
English		
Misogyny identification	0.821	0.757
Behaviour classification	0.303	0.575

Table 1: Results on the dev test (macro F1). In both cases, the results were obtained on a random 90%-10% split of the dev dataset.

From these results we could see that the locally weighted n-grams model using Random Forest was better in the identification tasks, while the bi-LSTM was more accurate for the misogynistic behaviour classification sub-task. However, a closer look to these results showed us that the bi-LSTM was classifying all instances but two in the majority class. Finally, due to these problems, we decided to participate to the task with just the locally weighted n-grams model.

The official results obtained by this model are detailed in Table 2. We do not consider the *derailing* category for which the system obtained 0 accuracy.

We also conducted a "post-mortem" test with the bi-LSTM model for which we obtained the following results:

[1] https://github.com/clab/dynet

Overall		
Subtask	**Italian**	**English**
Misogyny identification	0.824	0.586
Behaviour classification	0.473	0.165
Per-class accuracy (sub-task B)		
discredit	**0.694**	0.432
dominance	0.250	0.184
sexual_harassment	0.722	0.169
stereotype	0.699	0.040
active	0.816	0.541
passive	0.028	0.248

Table 2: Official results on the test set (macro F1) obtained by the locally-weighted bag of n-grams model.

Subtask	**Italian**	**English**
Misogyny identification	0.821	0.626
Behaviour classification	0.355	0.141

Table 3: Unofficial results on the test set (macro F1) obtained by the bi-LSTM character model.

As it can be observed, the results confirmed those obtained in the dev test for the misogyny identification sub-task, and in any case we observed that the "deep" model performed overall worse than its more "classical" counterpart.

The results obtained by our system were in general underwhelming and below the expectations, except for the *discredit* category, for which our system was ranked 1st and 3rd in Italian and English respectively. An analysis of the most relevant features according to information gain (Lee and Lee, 2006) showed that the 5 most informative n-grams are *tta*, *utt*, *che*, *tan*, *utta* for Italian and *you*, *the*, *tch*, *itc*, *itch* for English. They are clearly part of some swear words that can appear in different forms, or conjunctions like *che* that may indicate some linguistic phenomena such as emphasization (for instance, as in *"che brutta!"* - "what a ugly girl!"). On the other hand, another category for which some keywords seemed particularly important is the *dominance* one, but in that case the information gain obtained by sequences like *stfu* in English or *zitt* in Italian (related to the "shut up" meaning) was marginal. We suspect that the main problem may be related to the unbalanced training corpus in which the discredit category is dominant, but without knowing whether the other participants adopted some balancing technique it is difficult to analyze our results.

4 Conclusions

Our participation to the AMI task at EVALITA 2018 was not as successful as we hoped it to be; our systems in particular were not able to repeat the excellent results that they obtained at the DEFT 2018 challenge, although for a different task, the detection of messages related to public transportation in tweets. In particular, the bi-LSTM model underperformed and was outclassed by a simpler Random Forest model that uses locally weighted n-grams as features. At the time of writing, we are not able to assess if this was due to a misconfiguration of the neural network, or to the nature of the data, or the dataset. We hope that this participation and the comparison to the other systems will allow us to better understand where we have failed and why in view of future participations. The most positive point of our contribution is that the systems that we proposed are completely language-independent and we did not make any adjustment to adapt the systems that participated in a French task to the Italian or English language that were targeted in the AMI task.

Acknowledgments

We would like to thank the program "Investissements d'Avenir" overseen by the French National Research Agency, ANR-10-LABX-0083 (Labex EFL) for the support given to this work.

References

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *Natural Language Processing and Information Systems - 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings*, pages 57–64.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the evalita 2018 task on automatic misogyny identification (ami). In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18)*, Turin, Italy. CEUR.org.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Changki Lee and Gary Geunbae Lee. 2006. Information gain and divergence-based feature selection for machine learning-based text categorization. *Information processing & management*, 42(1):155–165.

Patrick Paroubek, Cyril Grouin, Patrice Bellot, Vincent Claveau, Iris Eshkol-Taravella, Amel Fraisse, Agata Jackiewicz, Jihen Karoui, Laura Monceaux, and Torres-Moreno Juan-Manuel. 2018. Deft2018: recherche d'information et analyse de sentiments dans des tweets concernant les transports en île de france. In *14ème atelier Défi Fouille de Texte 2018*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. 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*, pages 1480–1489.

Shiwei Zhang, Xiuzhen Zhang, and Jeffrey Chan. 2017. A word-character convolutional neural network for language-agnostic twitter sentiment analysis. In *Proceedings of the 22Nd Australasian Document Computing Symposium*, ADCS 2017, pages 12:1–12:4, New York, NY, USA. ACM.