Aggression Identification in Posts - two machine learning approaches

Faneva Ramiandrisoa

Aggression Identification in Posts - two machine learning approaches.

Faneva Ramiandrisoa[1,2] [0000−0001−9386−3531]

[1] IRIT & Université de Toulouse, France
faneva.ramiandrisoa@irit.fr
[2] Université d'Antananarivo

Abstract. Social media have changed the way people communicate. One of the aspects is cyber-aggression and interpersonal aggression that can be catalyzed by perceived anonymity. Automatically monitoring user-generated content in order to help moderating it is thus a hot topic. In this paper, we present and evaluate two supervised machine learning models to identify aggressive content and the level of aggressiveness. The first model uses random forest and linear regression while the second model uses deep learning techniques.

Keywords: Social media; Social media analysis; Cyber-agression; TRAC Trolling, Aggression and Cyberbulling; Machine learning based model

1 Introduction

Social media have changed the way people communicate [3,13,14,5]. One of these aspects is cyber-aggression and interpersonal aggression that can be catalyzed by perceived anonymity [16]. Automatically monitoring user-generated content in order to help moderating social media is thus an important although difficult topic [4,17].

In 2018, the Shared Task on Aggression Identification was organised as part of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC - 1) at COLING 2018 [9]. The objective of this task is to detect aggressive content and the level of aggressiveness. Thirty teams submitted their test runs. The best system obtained a weighted F-score of 0.64 on a data set composed of annotated Facebook comments.

In this paper, we report two models we developed in order to answer the aggression identification task. The first model uses random forest and linear regression which can be considered as relatively mature approaches while the second model combines CNN and LSTM recent deep learning techniques. No strong conclusion could be made on the superiority of one or the other model since it depends on the collection.

This paper is organized as follows: Section 2 reports related works, Section 3 describes our two approaches, Section 4 describes the dataset used in this work, reports the results and discuss them while Section 5 concludes this paper and presents future works.

2 Related work

Approaches based on features and supervised classifiers such as Support Vector Machines (SVM) are often used in order to learn to detect whether a text contains aggressiveness [24]; in recent years, deep learning has been also employed for this task [19,2].

Deep learning has also been used by TRAC challenge participants. TRAC [9] challenge is the first that focuses on detecting aggressive text. The task training set is composed of Facebook posts/comments; there is also two kinds of test sets: one from Facebook and another from Twitter.

Among the thirty participants, Saroyehun [2] obtained the best results. The authors investigated the efficacy of deep neural network by experimenting different models : CNN, LSTM, BiLSTM, and combinations thereof. In their experiments they used translation technique to enlarge the training set and added an external dataset on hate speech[3]. The LSTM model which was trained on the augmented training set only, achieved the best weighted F1 score of 0.6425 on Facebook test set ; it is the first ranked system on TRAC challenge ; the same system does not performed as well on the Twitter data set. The other system of the same team which implements a combination of CNN and LSTM and which was trained on the augmented training set and the additional dataset, achieved a weighted F1 score of 0.5920 and the third rank on the twitter test set.

Raiyani *et. al.* [20], meanwhile, tested different models for text classification in TRAC, from classic machine learning model to deep learning models. At the end, they kept three models: FastText model, Dense neural networks, and Voting of the two. The Dense neural networks gives better performance than the two others and achieved a weighted F1 score of 0.5813 on Facebook test set; it is the fourteenth rank on TRAC challenge. While it achieved the best weighted F1 score of 0.6009 and the first rank on the twitter test set, although it was trained on a Facebook dataset.

3 Machine learning based models

We developed two supervised machine learning based models that we evaluated in this paper. The first method combines random forest and logistic regression while the second approach is deep learning based. We also developed a model based on CNN only for which results can be found in [21]; it performs in between the two models reported in this paper.

3.1 Trac-RF_LR: combination of two classifiers

In this model we combined random forest (RF) based on surface features and linguistic features with logistic regression (LR) based on document vectorization. We chose this combination because a combination of multiple machine learning models placed first in many prestigious machine learning competitions [18], such as Netflix Competition,

[3] `https://github.com/ZeerakW/hatespeech`, accessed on January 10, 2020

Kaggle,... Moreover, when using non-combined models on the training dataset, the results were lower in the case of TRAC as well and this was confirmed on the test set (see section 4.3).

RF Classifier. The random forest model uses different features extracted from the comments as presented in Table 1. Some are adapted from [1,22] where the authors tried to detect depression from texts; another source of inspiration is [7] where the authors suggested an information nutritional label for describing text qualities.

Name	Hypothesis or tool/resource used
Part-of-speech frequency	Normalized frequencies of each tag: adjectives, verbs, nouns and adverbs (four features).
Negation	Normalized frequencies of negative words like: *no, not, didn't, can't*, ... The idea behind is to detect non direct aggressiveness.
Capitalized	The idea behind is that aggressive texts tend to put emphasis on the target they mention. It can indicate feelings or speaking volume.
Punctuation marks	! or ? or any combination of both can emphasize offensiveness of texts.
Emoticons	Another way to express sentiment or feeling.
Sentiment	Use of NRC-Sentiment-Emotion-Lexicons[4] to trace the polarity in text.
Emotions	Frequency of emotions from specific categories: anger, fear, surprise, sadness and disgust. The idea behind is to check the categories related to aggressiveness.
Gunning Fog Index	Estimate of the years of education that a person needs to understand the text at first reading.
Flesch Reading Ease	Measure how difficult to understand a text is.
Linsear Write Formula	Developed for the U.S. Air Force to calculate the readability of their technical manuals[5].
New Dale-Chall Readability	Measure the difficulty of comprehension that persons encounter when reading a text. It is inspired from Flesch Reading Ease measure.
Swear words	The intuition behind is that the texts containing insults are often aggressive.
Lexical analysis with python library *empath*	Empath is a tool for analyzing text across lexical categories. By default, it has 194 lexical categories and each category is considered as feature.

Table 1: List of features used in RF to represent texts (Facebook comments or tweets).

[4] `http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm`, accessed on 2017-02-23

Some of these features are used for abusive language detection, hate speech, cyberbullying and the others are used for sentiment or personality analysis that we judged useful for aggression detection.

A RF classifier was trained on train and validation sets by representing each text (Facebook comment or tweet) with a vector composed by the features we mentioned in Table 1.

The following parameters were used during the training: class_weight="balanced", max_features="sqrt", n_estimators=60, min_weight_fraction_leaf=0.0, criterion='entropy', random_state=2.

At prediction time, a text from the test set is represented with features and then run the trained model. The output is the estimated probabilities for the three classes (overtly aggressive, covertly aggressive and non-aggressive).

LR Classifier. This model is based on document vectorization using *Doc2vec* [12]. *Doc2vec* is used to represent sentences, paragraphs, or whole documents as vectors and it can be trained on small corpora, which is case of the task datasets.

Before building the LR Classifier, we first trained two separate *Doc2vec* models: a Distributed Bag of Words and a Distributed Memory model [12]. For the training, we used the same configuration as in [25] for representing user's text. The two Doc2vec models were trained on the train and validation sets. We used the Python package *gensim*[6][23]. We also concatenated the output vectors of these two models, as done in [25], resulting in a representation by a 200-dimension vector per text.

Then a logistic regression classifier was trained on the vectors for both the train and validation sets with the following parameters : class_weight="balanced", random_state=1, max_iter=100, solver="liblinear".

At prediction time, the texts from the test set were vectorized by using the two *Doc2vec* models and the 200-dimension vectors were given as input of trained classifier. The output is also a set of class probabilities.

Combination of two classifiers. The class probabilities obtained from RF classifier and LR Classifier were averaged and finally the class with the highest probability was considered as the class the text belongs to. We also tested different ways to combine the output probabilities obtained from the two classifiers RF and LR, such as maximum, minimum, etc., but the average method gave the best results.

3.2 Trac-CNN_LSTM: Combination of CNN and LSTM

This model combines two deep learning techniques: CNN and LSTM. The main idea is to pass input representation (sentence matrix in Figure 1) to the CNN and pass the local features learnt by the CNN (concatenated vectors in Figure 1) to the LSTM. Indeed, CNN and LSTM are complementary due to the fact that each of them captures information at different scales [2].

[5] `http://www.streetdirectory.com/travel_guide/15675/writing/how_to_choose_the_best_readability_formula_for_your_document.html`, accessed on 2018-02-25

[6] `https://radimrehurek.com/gensim/index.html`

The architecture of our combined model is illustrated in Figure 1. It is as follows: first, we convert sentences/texts into *sentences matrix*[7] where each row is a vector representation[8] of each word in the sentences/texts. Then, convolutions are applied on the sentences matrix where we used three filter region sizes: bigrams (height = 2), trigrams (height = 3) and fourgrams (height = 4). Each region has 100 filters; thus, in total there are 300 filters. The result of convolutions is called feature maps; vectors with variable-length according to the region filter and each filter region has 100 feature maps. Afterwards, a 1-max pooling is performed over feature maps. More precisely, for each region the largest number from each feature map is kept and then concatenated to form a vector. As a result, we obtain one vector of size 100[9] per region filter. Then, these three vectors are concatenated to form a feature vector and a dropout is applied on this feature vector. The concatenated feature vector is passed to the LSTM layer. Then, we added one fully connected hidden layer to reduce the dimension of the concatenated vector, followed by a dropout. Finally, an output layer, which is also a fully connected layer with three possible output states, is added. On the output layer, the activation function used is the softmax function.

The architecture of our model is inspired from the CNN architecture Zhang *et. al.* [26] proposed and which is used for sentences classification. In that task, their CNN architecture outperforms baseline methods which use SVM as well as the one that used CNN in [8].

4 Evaluation

4.1 Data set

The evaluation is based on the TRAC 2018 shared task [9]. The task dataset is a subset of Kumar et al' [10] and consists in English and Hindi randomly sampled Facebook comments. In this study, we focused on the English part of the dataset which is detailed in Table 2. It is composed of (a) 11,999 Facebook comments for training and 3,001 comments for validation. It is annotated with 3 levels of aggression - Overtly Aggressive (OAG), Covertly Aggressive (CAG) and Non- Aggressive (NAG), (b) 916 English comments for test. Additionally, 1,257 English tweets were given as a second test set.

4.2 Evaluation measure

The evaluation metric used in this paper is the weighted F1 which was also used in the TRAC shared task. The weighted F1 is equal to the average, weighted by the number of instances for each label, of the F1 (given by equation 1) of each class label.

$$F1 = 2\frac{R*P}{R+P} \tag{1}$$

[7] The dimension of a sentence matrix is $l \times d$, where l is the length of the longest text/sentence in the dataset and d is the dimension of word vector representation.

[8] The word vector representation is obtained with word2vec model [15] trained on the training and validation sets.

[9] Because there is 100 feature maps.

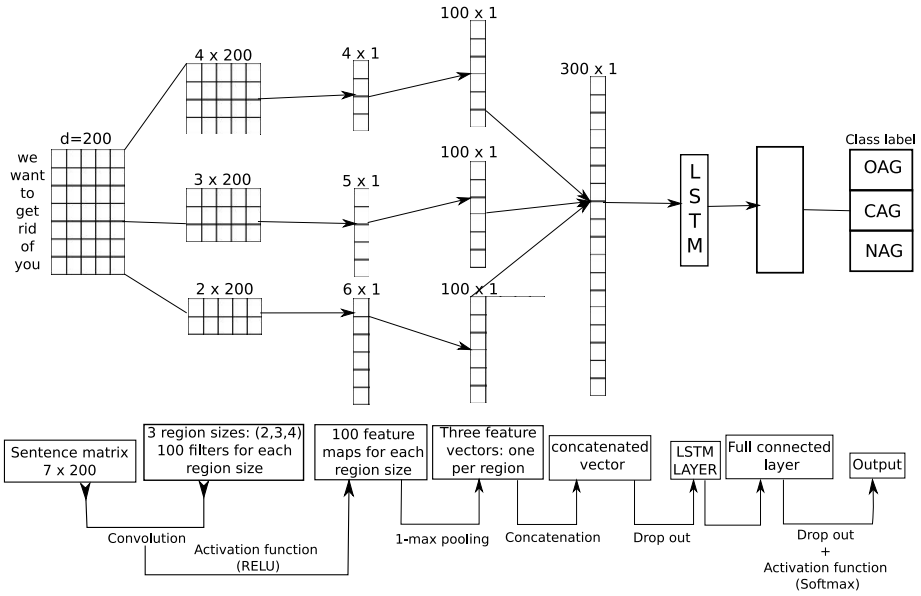

Fig. 1: Illustration of a CNN + LSTM architecture for aggression detection inspired from [26].

Number of	Train	Validation	Test	
			Facebook	Twitter
texts (=posts+comments)	11,999	3,001	916	1,257
Overt aggression	2,708	711	144	361
Covert aggression	4,240	1,057	142	413
No aggression	5,051	1,233	630	483

Table 2: Distribution of training, validation and testing data on TRAC 2018 data collection.

where $P = \frac{tp}{tp+fp}$ is the precision, $R = \frac{tp}{tp+fn}$ is the recall, tp denotes the true positives, fp the false positives, and fn the false negatives.

4.3 Results

Table 3 reports the results we obtained with the two models presented above. For comparison, we report also results obtained with the RF classifier only and with the LR classifier only. The baseline mentioned in the first row was given by the TRAC shared task organizers while the second row is the best result from participants in the TRAC workshop.

System	Weighted F1	
	Facebook	**Twitter**
Random Baseline	0.354	0.348
Saroyehun [2]	0.642	0.592
Trac-RF_LR	**0.581**	0.409
Trac-CNN_LSTM	0.559	**0.511**
Trac-RF_only	0.573	0.397
Trac-LR_only	0.569	0.452

Table 3: Results for the English (Facebook and Twitter) task. Bold value is the best performance for our approaches.

We can see that our two models outperform the baseline on both Facebook and Twitter subsets. Trac-RF-LR is better than Trac-CNN-LSTM on the Facebook collection while it is the opposite on the Twitter collection. This could be due to the train dataset which is only composed of texts crawled from Facebook. Indeed, we can observe the same behaviour for the other systems that participated to the challenge [9]. The only exception is for Saroyehun [2] system which performs better on the Twitter dataset.

5 Conclusion

In this paper, we presented two different supervised machine learning approaches for aggression identification on TRAC 2018 English collections (Facebook and Twitter based). The combination of random forest and linear regression classifiers based on a set of surface features and document vectorization leaded to the sixteenth ranked system out of thirty on the Facebook collection. The combination of CNN and Long Short-Term Memory was ranked fifteenth out of thirty systems.

To extend this work, we plan to update our models by adding new features such as bag of words or features more specific to the aggression. We also plan to apply feature engineering on the features we used in this paper in order to see which one are the most useful. On the other hand, feature selection could also be applied to build models that use features as less as possible [11,6]. Finally, an investigation on deep learning models will be conducted by using different architectures such as hierarchical attention network. We do believe that these tracks can help designing more performing models.

Ethical issue. While TRAC challenge has its proper ethical policies, detecting aggressive content from user's posts raises ethical issues that are beyond the scope of the paper.

Acknowledgement. This work has been partially funded by the European Union's Horizon 2020 H2020-SU-SEC-2018 under the Grant Agreement n°833115 (PREVISION project). This work has also been partially supported by the ***Ministère des Affaires étrangères et du Développement international*** under the scholarship ***EIFFEL-DOCTORAT 2017/ n°P707544H*** for Faneva Ramiandrisoa's PhD thesis.

References

1. Abdou Malam, I., Arziki, M., Nezar Bellazrak, M., Benamara, F., El Kaidi, A., Es-Saghir, B., He, Z., Housni, M., Moriceau, V., Mothe, J., Ramiandrisoa, F.: IRIT at e-Risk (regular paper). In: International Conference of the CLEF Association, CLEF 2017 Labs Working Notes. ISSN 1613-0073, vol. 1866. CEUR Workshop Proceedings, http://CEUR-WS.org (2017)
2. Aroyehun, S.T., Gelbukh, A.: Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018). pp. 90–97 (2018)
3. Caron, J., Light, J.: "social media has opened a world of 'open communication:'" experiences of adults with cerebral palsy who use augmentative and alternative communication and social media. Augmentative and Alternative Communication **32**(1), 25–40 (2016)
4. Chen, J., Xu, H., Whinston, A.B.: Moderated online communities and quality of user-generated content. Journal of Management Information Systems **28**(2), 237–268 (2011)
5. Décieux, J.P., Heinen, A., Willems, H.: Social media and its role in friendship-driven interactions among young people: A mixed methods study. YOUNG **27**(1), 18–31 (2019)
6. Déjean, S., Ionescu, R.T., Mothe, J., Ullah, M.Z.: Forward and Backward Feature Selection for Query Performance Prediction. In: ACM Symposium on Applied Computing (SAC). ACM : Association for Computing Machinery (2020)
7. Fuhr, N., Giachanou, A., Grefenstette, G., Gurevych, I., Hanselowski, A., Jarvelin, K., Jones, R., Liu, Y., Mothe, J., Nejdl, W., et al.: An information nutritional label for online documents. In: ACM SIGIR Forum. vol. 51, pp. 46–66. ACM (2018)
8. Kim, Y.: Convolutional neural networks for sentence classification. CoRR **abs/1408.5882** (2014)
9. Kumar, R., Ojha, A.K., Malmasi, S., Zampieri, M.: Benchmarking Aggression Identification in Social Media. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbulling (TRAC). Santa Fe, USA (2018)
10. Kumar, R., Reganti, A.N., Bhatia, A., Maheshwari, T.: Aggression-annotated corpus of hindi-english code-mixed data. arXiv preprint arXiv:1803.09402 (2018)
11. Laporte, L., Flamary, R., Canu, S., Déjean, S., Mothe, J.: Non-convex Regularizations for Feature Selection in Ranking with Sparse SVM. IEEE Transactions on Neural Networks and Learning Systems **25**(6), 1118–1130 (june 2014)
12. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. In: Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014. pp. 1188–1196 (2014)
13. Lipschultz, J.H.: Social media communication: Concepts, practices, data, law and ethics. Routledge (2017)
14. Marganski, A., Melander, L.: Intimate partner violence victimization in the cyber and real world: Examining the extent of cyber aggression experiences and its association with in-person dating violence. Journal of interpersonal violence **33**(7), 1071–1095 (2018)
15. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States. pp. 3111–3119 (2013)
16. Mishna, F., Regehr, C., Lacombe-Duncan, A., Daciuk, J., Fearing, G., Van Wert, M.: Social media, cyber-aggression and student mental health on a university campus. Journal of mental health **27**(3), 222–229 (2018)
17. Myers West, S.: Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. New Media & Society **20**(11), 4366–4383 (2018)

18. Osama, M., El-Beltagy, S.R.: A transfer learning approach for emotion intensity prediction in microblog text. In: Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2019, AISI 2019, Cairo, Egypt, 26-28 October 2019. pp. 512–522 (2019). https://doi.org/10.1007/978-3-030-31129-2_47
19. Priyadharshini, G.: A pragmatic supervised learning methodology of hate speech detection in social media (2019)
20. Raiyani, K., Gonçalves, T., Quaresma, P., Nogueira, V.B.: Fully connected neural network with advance preprocessor to identify aggression over facebook and twitter. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, TRAC@COLING, Santa Fe, New Mexico, USA. pp. 28–41 (2018)
21. Ramiandrisoa, F., Mothe, J.: Irit at trac 2018. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbulling, TRAC@COLING, Santa Fe, New Mexico, USA. pp. 19–27 (2018)
22. Ramiandrisoa, F., Mothe, J., Benamara, F., Moriceau, V.: IRIT at e-Risk 2018 (regular paper). In: Conference and Labs of the Evaluation Forum, Living Labs (CLEF 2018), Avignon, France, 10/09/2018-14/09/2018. p. (on line). CEUR-WS : Workshop proceedings (2018)
23. Rehurek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. Citeseer (2010)
24. Schmidt, A., Wiegand, M.: A Survey on Hate Speech Detection Using Natural Language Processing. In: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. SocialNLP@EACL 2017. pp. 1–10. Valencia, Spain (2017)
25. Trotzek, M., Koitka, S., Friedrich, C.M.: Linguistic metadata augmented classifiers at the CLEF 2017 task for early detection of depression. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017. (2017)
26. Zhang, Y., Wallace, B.C.: A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. CoRR **abs/1510.03820** (2015)