UMUTeam at MEX-A3T’2020: Detecting Aggressiveness with Linguistic Features and Word Embeddings

José Antonio García-Díaz[a], Rafael Valencia-García[a]

[a]*Facultad de Informática, Universidad de Murcia, Campus de Espinardo, 30100, Spain*

Abstract

Social networks have become a dangerous place used by some people to harass others by taking advantage of the anonymity that the Internet provides. The consequences of this behaviour lead to long-term harm to victims, or in extreme cases, they can lead them to commit suicide. Due to the large volume of post created daily, manual supervision to prevent this harmful content becomes impossible. In this paper we describe our participation in MEX-A3T focused on aggressiveness identification in tweets written in Mexican Spanish. Our proposal is grounded in the combination of linguistic features and pre-trained word embeddings. In our first run, we use Support Vector Machines to train a combination of linguistic features and sentence embeddings; whereas in the second and third run the linguistic features are combined with two deep-learning models: a Convolutional Neural Network and a Bidirectional Long-Short Term Memory. Although our proposal does not beat the baseline, almost achieves the same results providing at the same time an interpretable model.

Keywords

Sentiment-Analysis, Text Classification, Deep-learning

1. Introduction

Some people use social networks as a means to harass others by taking advantage of anonymity or the perceived social distance that the Internet provides [1]. The consequences of this behaviour should not be taken as a joke, since they can cause long-term harm to victims and, in extreme situations, lead them to severe isolation and even to commit suicide [2]. Due to the large volume of posts created every day, manual supervision of harmful material becomes very complicated, if not impossible, in some cases [3]. In order to prevent this harmful behaviour, researchers have taken advantage of modern Natural Language Processing (NLP) techniques for analysing and therefore improving automatic hate-speech detectors through the analysis of large amounts of data. In line with this, many NLP workshops have proposed the creation of hate-speech detectors. In HatEval [4], for example, the objective consisted in multilingual detection of hate speech against immigrants and women. Other tasks have focused in specific groups such as women. In this sense, AMI (Automatic Misogyny Identification) is a task focused on spotting several misogynist traits proposed in [5] and [6] in available datasets composed

Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)

EMAIL: joseantonio.garcia8@um.es (J.A. García-Díaz); valencia@um.es (R. Valencia-García)
ORCID: 0000-0002-3651-2660 (J.A. García-Díaz); 0000-0003-2457-1791 (R. Valencia-García)

CEUR Workshop Proceedings (CEUR-WS.org)

English, Spanish, and Italian.

One of the reasons why hate-speech identification is hard is because is heavily cultural and background dependant; even in countries that share the same language [7]. Therefore, in this paper we describe our participation in the Aggressiveness Identification Track of MEX-A3T'2020 [8] which involved the identification of tweets written in Spanish-Mexican labelled as aggressive. Our proposal is grounded on the usage of linguistic features and different forms of pre-trained word embeddings. The rest of the paper is organised as follows. First, in Section 2, our proposal and the different runs submitted are described. Then, in Section 3, we describe the achieved results. Finally, the lessons learned and further work are described in Section 4.

2. System description

In the bibliography, several approaches can be found regarding Hate Speech identification. Some of them are oriented towards specific forms of hate-speech, such as cyber-bullying [9, 10, 11], misogyny identification [12, 13], or detecting hate-speech towards immigrants [14]. Others, as the proposed by HateEval [4], included several target groups.

The corpus consisted in tweets compiled from Mexico City. The training split is composed by 7332 tweets where 2110 of them were labelled as *aggressive*, and the remaining were labelled as *non-aggressive*. We encode the tweets as lowercase, strip multiple white spaces, and remove the hashtag symbol. Then, we extract the linguistic features with UMUTextStats [15], a tool inspired in Linguistic Inquiry and Word Count (LIWC) [16]. Although LIWC is available in Spanish [17, 18], it does not consider some Spanish linguistic phenomena such as grammatical gender or a deep classification of part-of-speech, morphemes or suffixes. UMUTextStats manages a total of 311 linguistic features organised into: (1) grammatical features, (2) morphological features, (3) spelling and stylistic errors, (4) figurative language [19], (5) statistics regarding the sentence type, (6) punctuation symbols, (7) topics, and (8) a great variety of positive and negative feelings.

In our proposal, we combine linguistic features with word embeddings [20] and sentence embeddings [21]. Word embeddings can be trained from news sites and public encyclopedias to convey general semantic rules. For the first run, we use Weka [22] to train a Support Vector Machine (SVM) with the combination of linguistic features and sentence embeddings. Specifically, we use the open-source library LibLinear [23], based on efficient Support Vector Classifier (SVC) with linear kernels. Sentence embeddings were obtained from pre-trained word embeddings based on the Spanish model of fastText [24]. The model was trained with the provided training dataset and evaluated with 10 fold-cross validation. For the second and third run we use the functional API of Keras [25]. In both runs we combined the linguistic features with pre-trained word embeddings with a Bidirectional Long-Short Term Memory (BiLSTM) for the second run and a Convolutional Neural Network (CNN) for the third run. BiLSTM was selected because it can handle long semantic dependencies and CNN was selected because it can spot local information regardless their position, obtaining syntactic and semantic information [26]. Both runs were trained with 10 epochs and a batch size of 32. In a nutshell, the layer architecture can be described as follows: linguistic features are the input of a hidden layer with a dimensionality of 10 and its output is concatenated to the output of CNN or BiLSTM (depending on the run). Then, the combination of the outputs are connected sequentially to

Table 1
Comparison of our runs with the two base-lines and the winner of the task

Model	F1-OFF	F1-NON	F1 macro	P	R	ACC
best-result	0.7998	0.9195	0.8596	0.8605	0.8588	0.8851
baseline1	0.7124	0.8841	0.7983	0.7988	0.7988	0.8348
baseline2	0.6760	0.8780	0.7770	-	-	0.8228
run2	0.6727	0.8706	0.7716	0.7744	0.7691	0.8145
run3	0.6516	0.8771	0.7644	0.7644	0.7503	0.8183
run1	0.5892	0.8430	0.7161	0.7223	0.7112	0.7728

two more hidden dense layers (both with Rectified Linear Unit -ReLU- as activation function) until the final output layer for binary prediction with a sigmoid as activation function. In case of the run 2, the architecture of BiLSTM consists in a Bidirectional layer with a dropout of 0.2 and a recurrent dropout of 0.2 connected to a hidden layer with a dimensionality of 10 and softmax as activation function. In case of the run 3, the architecture of CNN consisted in a one-dimensional convolutional layer (Conv1D) with an output of 128 output, an 1D-window size of 5, and a rectified linear unit -ReLU- as activation function. This layer is connected to a GlobalMaxPool1D layer, and this is connected to two more hidden layers with a dimensionality of 10.

3. Results

The organisers of MEX-A3T'2020 ranked the participants by using the F1 measure on the *aggressive* class. They also created two baseline models. The first one is based on a Bag of Words (BoW) model trained with a SVM, which achieves an F1 measure of the offensive class of 0.6760; whereas the second baseline was trained with a Bidirectional Gated Unit Network (BiGRU), which achieves an F1 measure of the offensive class of 0.7124. The comparison of our three runs with the two baseline models and winner run are shown in Table 1. This table contains the F1 measure both of the *offensive* (F1-OFF) and the *non-offensive* (F1-NON) classes, as well as the macro F1 measure (F1 macro), the precision (P), recall (R) and accuracy (ACC).

Our best result was achieved by the *run2* (BiLSTM+LF) which almost equals the second baseline for the F1-offensive (0.6727 vs 0.6760) and *run3* drops the F1-offensive slightly (0.6516 vs 0.6760); however, both runs where far from the first baseline. In order to understand which are the most discriminating features, we calculated the Information Gain (IG) of the training dataset (not showed). We observed that linguistic features related to negative sentiments are the most discriminating ones, which includes *offensive-language*, *sadness*, *anger*, or *anxiety* among others. Other discriminatory features were *swear-words* (vulgar expressions, but not necessarily offensive). *Twitter mentions* and verbs in first person are also discriminating features, which suggests that some of the menaces and offensive expressions that appear in the texts are made in first person and towards specific people. It is worth noting that demonyms appear on the top twenty linguistic features, which suggest that some of the tweets refers to people belonging to specific places or ethnic groups.

4. Conclusions

In this paper we have described our participation in the MEX-A3T task regarding aggressiveness identification with the experiments that rely on linguistic features and different combinations word embeddings; however, our best results do not outperform the baseline results for a minimal difference. After an analysis of the results and the discriminating linguistic features, we achieve the following insights: (1) aggressiveness in social networks are characterised by the usage of a strong offensive language as well as misspelled words and linguistic errors; (2) the number and relevance of verbs in first person in singular indicates that the threats are commonly performed directly; (3) the number of twitter mentions and the appearance of demonyms do not make clear if the targets are either ethnics groups or individual target; so a more detailed analysis of the results is required.

However, these findings should be taken with caution. As we observed, our proposal does not improved any of the baselines proposed. Furthermore, we consider two main actions to follow. First, as our proposal was designed with Spanish-European language in mind, it should be adapted to handle different specific languages to manage cultural background. Second, we will include state-of-the-art NLP techniques, such as BERT or ELMo, in order to evaluate the classification accuracy.

Acknowledgments

This work has been supported by the Spanish National Research Agency (AEI) and the European Regional Development Fund (FEDER/ERDF) through projects KBS4FIA (TIN2016-76323-R) and LaTe4PSP (PID2019-107652RB-I00). In addition, José Antonio García-Díaz has been supported by Banco Santander and University of Murcia through the Doctorado industrial programme.

References

[1] N. Lapidot-Lefler, A. Barak, Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition, Computers in Human Behavior 28 (2012) 434 – 443. URL: http://www.sciencedirect.com/science/article/pii/S0747563211002317. doi:https://doi.org/10.1016/j.chb.2011.10.014.

[2] G. S. O'Keeffe, K. Clarke-Pearson, et al., The impact of social media on children, adolescents, and families, Pediatrics 127 (2011) 800–804.

[3] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, E. Gilbert, You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech, Proceedings of the ACM on Human-Computer Interaction 1 (2017) 1–22.

[4] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. R. Pardo, P. Rosso, M. Sanguinetti, Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 54–63.

[5] E. Fersini, P. Rosso, M. Anzovino, Overview of the task on automatic misogyny identification at ibereval 2018., in: IberEval@ SEPLN, 2018, pp. 214–228.

[6] E. Fersini, D. Nozza, P. Rosso, Overview of the evalita 2018 task on automatic misogyny identification (ami), EVALITA Evaluation of NLP and Speech Tools for Italian 12 (2018) 59.

[7] A. M. Croom, Spanish slurs and stereotypes for mexican-americans in the usa: A context-sensitive account of derogation and appropriation, Pragmática Sociocultural/Sociocultural Pragmatics 8 (2014) 145–179.

[8] M. E. Aragón, H. Jarquín, M. Montes-y Gómez, H. J. Escalante, L. Villaseñor-Pineda, H. Gómez-Adorno, G. Bel-Enguix, J.-P. Posadas-Durán, Overview of mex-a3t at iberlef 2020: Fake news and aggressiveness analysis in mexican spanish, in: Notebook Papers of 2nd SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Malaga, Spain, September, 2020.

[9] H. Rosa, N. Pereira, R. Ribeiro, P. C. Ferreira, J. P. Carvalho, S. Oliveira, L. Coheur, P. Paulino, A. V. Simão, I. Trancoso, Automatic cyberbullying detection: A systematic review, Computers in Human Behavior 93 (2019) 333–345.

[10] M. A. Al-garadi, K. D. Varathan, S. D. Ravana, Cybercrime detection in online communications: The experimental case of cyberbullying detection in the twitter network, Computers in Human Behavior 63 (2016) 433–443.

[11] A. López-Martínez, J. A. García-Díaz, R. Valencia-García, A. Ruiz-Martínez, Cyberdect. a novel approach for cyberbullying detection on twitter, in: International Conference on Technologies and Innovation, Springer, 2019, pp. 109–121.

[12] E. Shushkevich, J. Cardiff, Automatic misogyny detection in social media: A survey, Computación y Sistemas 23 (2019). doi:10.13053/cys-23-4-3299.

[13] T. Lynn, P. T. Endo, P. Rosati, I. Silva, G. L. Santos, D. Ging, A comparison of machine learning approaches for detecting misogynistic speech in urban dictionary, in: 2019 International Conference on Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA), IEEE, 2019, pp. 1–8.

[14] A. Ben-David, A. M. Fernández, Hate speech and covert discrimination on social media: Monitoring the facebook pages of extreme-right political parties in spain, International Journal of Communication 10 (2016) 27.

[15] J. A. García-Díaz, M. Cánovas-García, R. Valencia-García, Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in latin america, Future Generation Computer Systems 112 (2020) 614–657. doi:https://doi.org/10.1016/j.future.2020.06.019.

[16] Y. R. Tausczik, J. W. Pennebaker, The psychological meaning of words: Liwc and computerized text analysis methods, Journal of language and social psychology 29 (2010) 24–54.

[17] M. del Pilar Salas-Zárate, M. A. Paredes-Valverde, M. Á. Rodríguez-García, R. Valencia-García, G. Alor-Hernández, Automatic detection of satire in twitter: A psycholinguistic-based approach, Knowl. Based Syst. 128 (2017) 20–33. URL: https://doi.org/10.1016/j.knosys.2017.04.009. doi:10.1016/j.knosys.2017.04.009.

[18] M. del Pilar Salas-Zárate, E. López-López, R. Valencia-García, N. Aussenac-Gilles, Á. Almela, G. Alor-Hernández, A study on LIWC categories for opinion mining in spanish reviews, J. Inf. Sci. 40 (2014) 749–760. URL: https://doi.org/10.1177/0165551514547842. doi:10.1177/0165551514547842.

[19] M. del Pilar Salas-Zárate, G. Alor-Hernández, J. L. Sánchez-Cervantes, M. A. Paredes-

Valverde, J. L. García-Alcaraz, R. Valencia-García, Review of english literature on figurative language applied to social networks, Knowledge and Information Systems (2019) 1–33.

[20] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013).

[21] C. Zhang, S. Sah, T. Nguyen, D. Peri, A. Loui, C. Salvaggio, R. W. Ptucha, Semantic sentence embeddings for paraphrasing and text summarization, CoRR abs/1809.10267 (2018). URL: http://arxiv.org/abs/1809.10267. arXiv:1809.10267.

[22] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, The weka data mining software: an update, ACM SIGKDD explorations newsletter 11 (2009) 10–18.

[23] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin, Liblinear: A library for large linear classification, Journal of machine learning research 9 (2008) 1871–1874.

[24] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, T. Mikolov, Learning word vectors for 157 languages, arXiv preprint arXiv:1802.06893 (2018).

[25] F. Chollet, et al., Keras, https://github.com/fchollet/keras, 2015.

[26] M. A. Paredes-Valverde, R. C. Palacios, M. del Pilar Salas-Zárate, R. Valencia-García, Sentiment analysis in spanish for improvement of products and services: A deep learning approach, Scientific Programming 2017 (2017) 1329281:1–1329281:6. URL: https://doi.org/10.1155/2017/1329281. doi:10.1155/2017/1329281.