Alejandro Mosquera at DETOXIS 2021: Deep Learning Approaches to Toxicity Detection in Spanish Social Media Texts

Alejandro Mosquera López[1[0000-0002-6020-3569]]

Broadcom Corporation, 1320 Ridder Park Drive San Jose, 95131 California, USA
alejandro.mosquera@broadcom.com

Abstract. This paper presents the system submitted to the DETOXIS 2021 challenge for detecting toxicity in Spanish social media texts. The chosen approach relies on an ensemble of different neural network architectures including thread and topic features as side information. For sub-task 1, we have also applied machine translation in order to reuse linguistic resources from other languages such as English. Our best submission scored 0.569 F1 in the test set, ranking 6th out of 31 competing teams.

Keywords: Toxicity detection · Spanish · Social Media · Machine translation · Text Normalization · Deep learning · Capsule networks.

1 Introduction

News websites allow million of users to share and discuss their opinions publicly in near real-time every day. Such large reach and constantly increasing user base present challenges for content moderation teams, which not only need to fight affiliate and cyber-crime operators but also less traditional forms of messaging abuse such as the spread of hate, propaganda and fake news.

While social media platforms are under increasingly pressure to swiftly deal with the spread of toxic content, the use of over-aggressive filtering models and the under-representation of certain user groups in the training data can also have negative consequences if false positives happen at large scale [23].

Because of the aforementioned reasons, the automatic detection of toxic language in social media has received growing attention from the NLP research community in the last few years, which is also reflected in the number of public evaluations and resources recently focused on this area: e.g. HASOC [14] for hate speech and aggressive content, TRAC [8] for identifying aggression, HatEval [1]

IberLEF 2021, September 2021, Málaga, Spain.

for detecting hate speech against women and immigrants, OffensEval-2019 [24] and OffensEval-2020 [25], both for identifying and categorizing offensive language.

This paper evaluates our participation in the shared task DETOXIS [22] of IberLEF-2021 for the subtask 1: Toxicity detection of Spanish comments posted in response to news articles related to immigration, using an ensemble of neural networks. The rest of the document is organised as follows: In section 2, related work is reviewed. In Section 3 we describe our system and approach. In Section 4 we evaluate the obtained results. Finally, in Section 5 we draw our conclusions and outline potential future work.

2 Related Work

Best performing approaches for toxicity detection follow the recent advances in neural networks for NLP: Liu et al. [10] and Zhu et al. [26] leveraged bidirectional transformers by fine-tuning BERT [5] embeddings. Earlier architectures such as convolutional neural networks (CNN) and bidirectional LSTMs (bi-LSTMs) can also obtain strong results [12] when paired with pre-trained embeddings such as FastText [2], GloVe [19] or word2vec [15]. Finally, combining different models and features helps reducing bias and variance, examples are voting ensembles [21] and stacked generalization [13].

3 System Description

Since the first sub-task was only focused on determining if a comment is either toxic or not, we have treated it as a binary classification problem.

3.1 Pre-processing

Social media texts usually contain informal lexical variants and out-of-vocabulary words which can be difficult to understand not only for humans but also for NLP tools and applications [18]. For this reason, we have applied a text normalization filter in order to reduce out-of-vocabulary words (OOV) by using a lexical normalization dictionary which is recursively combined with shortening and lengthening rules [17].

3.2 Data Augmentation

Data augmentation is a popular technique that can increase the volume and diversity of the training data for many applications including NLP [9]. While we have only used the NewsCom-TOX dataset provided by the organization for training purposes, in order to reuse publicly available pre-trained resources for the English language we have also generated a parallel dataset in English by using the Google Translate API.

3.3 Models

The list of models that our system comprises of is as follows:

- **capsule_es** Neural network with a capsule network architecture [20] using SBWC [3] i25 GloVe Spanish embeddings.
- **capsule_en** Neural network with a capsule network architecture using GloVe 840B-300d English embeddings.
- **detox_orig** Detoxify original [6], a pre-trained BERT model that detects toxicity in English texts.
- **detox_unb** Detoxify unbiased, a pre-trained RoBERTa [11] model that recognizes toxicity in English texts and minimizes unintended biases with respect to mentions of identities.
- **detox_multi** Detoxify multilingual, a pre-trained XLM [4] model that detects toxicity in English texts.
- **detox_multi_es** Detoxify multilingual, a pre-trained XLM model that detects toxicity in Spanish texts.
- **spacylr** Logistic regression model trained using Spacy [7] Spanish embeddings.

3.4 Side Information

In addition to the actual comments, non-textual metadata was made available as part of the training dataset such as topic, thread_id, comment_id and reply_to. These were used in order to engineer extra features for the stacking model as side information:

- **topic_words_max** The maximum word-wise toxicity score in a comment after averaging all the capsule_es model probabilities of the individual words across the training data by topic.
- **topic_words_avg** The average word-wise toxicity score in a comment after averaging all the capsule_es model probabilities of the individual words across the training data by topic.
- **avg_group_tox** The average toxicity score determined by the capsule_es model for all the comments with the same thread_id.

Since there was no topic information in the test dataset, we have considered it as a separate topic when computing the features above. Although inaccurate (the test data had comments from the same set of topics as train) it did not impact negatively in the final results.

3.5 Stacking Model

Due the relatively small amount of training data in the NewsCom-TOX corpus (less than 4000 samples, only 1147 positive) we went for an stacked generalization strategy, where the soft probabilities calculated from several models are used as features with the original labels against the stacking model. This not only reduces the computing resources needed in order to tune hyperparameters and perform cross-validation, but can also achieve competitive results even with just pre-trained models [16].

Our stacking model was logistic regression with a custom threshold of 0.32, which was determined via cross-validation. The latter was required because of the class imbalance and the unusual evaluation metric used in this sub-task (F1 of the toxicity class rather than micro or macro averages) which favours aggressive models towards the positive class.

The most important features by considering the regression coefficients can be seen at Figure 1. From there we can determine that capsule networks and avg_group_tox are the strongest features for detecting the toxic and non-toxic class respectively.

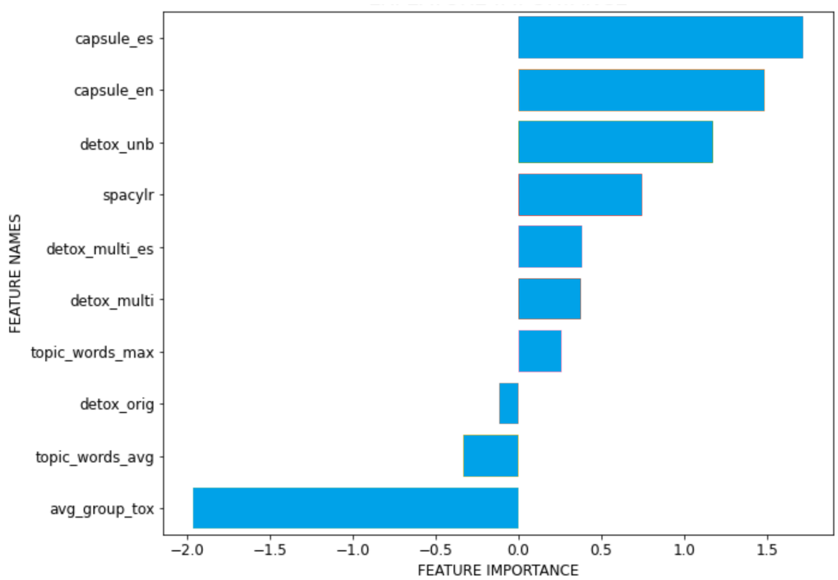

Fig. 1. figure
Feature importance based on the LR coefficients of the stacking model.

4 Results

Our toxicity detection system obtained promising results as shown in Table 1: It ranked 6th/31, with a difference in F1 of only 0.077 when compared against the winning system. It is also worth mentioning that only 16 systems (out of 31) achieved better F1 score than the AllToxic benchmark, which highlights the difficulty of this sub-task for the chosen evaluation metric.

System	F1 Toxic	Model	F1 Toxic	Model	F1 Toxic
SINAI (best)	**0.6461**	capsule_es	0.5040	capsule_es	0.5168
Alejandro Mosquera	0.5691	capsule_en	0.4872	capsule_en	0.5299
AllToxic	0.4231	spacylr	0.4833	spacylr	0.5156
RandomClassifier	0.3760	detox_unb	0.4117	detox_unb	0.4671
ChainBOW	0.3746	detox_multi	0.4053	detox_multi	0.4430
BOWClassifier	0.1837	detox_multi_es	0.3887	detox_multi_es	0.4209
		detox_orig	0.3542	detox_orig	0.4237
				Alejandro Mosquera	0.5813

Table 1. Partial results table for the test set (left) results of individual models in the ensemble for the test set (middle) and out-of-fold validation scores for the train set (right).

With regards to our individual models, we can observe that they are weaker, only 3 out of 7 would beat the AllToxic baseline, and exhibit higher variance between train and test scores than the final stacking ensemble. However, a post-workshop analysis showed that removing the weakest models would have not improved the final score.

5 Conclusions

In this paper we describe the system for detecting toxicity in Spanish social media texts engineered for DETOXIS 2021 sub-task 1. Since the amount of training data was relatively small, different strategies were applied in order to overcome this limitation, such as performing data augmentation through machine translation and leveraging pre-trained models using larger toxicity datasets. Our best submission was a logistic regression ensemble using neural network predictions and side information features extracted from thread and topic metadata.

References

1. Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F.M.R., Rosso, P., Sanguinetti, M.: 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. pp. 54–63 (2019)
2. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017), `https://www.aclweb.org/anthology/Q17-1010`
3. Cardellino, C.: Spanish Billion Words Corpus and Embeddings (August 2019), `https://crscardellino.github.io/SBWCE/`
4. CONNEAU, A., Lample, G.: Cross-lingual language model pretraining. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019), `https://proceedings.neurips.cc/paper/2019/file/c04c19c2c2474dbf5f7ac4372c5b9af1-Paper.pdf`

5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2018)
6. Hanu, L., Unitary team: Detoxify. Github. https://github.com/unitaryai/detoxify (2020)
7. Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A.: spaCy: Industrial-strength Natural Language Processing in Python (2020). https://doi.org/10.5281/zenodo.1212303, https://doi.org/10.5281/zenodo.1212303
8. Kumar, R., Reganti, A.N., Bhatia, A., Maheshwari, T.: Aggression-annotated Corpus of Hindi-English Code-mixed Data. In: chair), N.C.C., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., Tokunaga, T. (eds.) Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (May 7-12, 2018 2018)
9. Li, Y., Li, X., Yang, Y., Dong, R.: A diverse data augmentation strategy for low-resource neural machine translation. Information **11**(5) (2020). https://doi.org/10.3390/info11050255, https://www.mdpi.com/2078-2489/11/5/255
10. Liu, P., Li, W., Zou, L.: NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In: Proceedings of the 13th International Workshop on Semantic Evaluation. pp. 87–91. Association for Computational Linguistics, Minneapolis, Minnesota, USA (Jun 2019). https://doi.org/10.18653/v1/S19-2011, https://www.aclweb.org/anthology/S19-2011
11. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach (2019), http://arxiv.org/abs/1907.11692, cite arxiv:1907.11692
12. Mahata, D., Zhang, H., Uppal, K., Kumar, Y., Shah, R.R., Shahid, S., Mehnaz, L., Anand, S.: MIDAS at SemEval-2019 task 6: Identifying offensive posts and targeted offense from twitter. In: Proceedings of the 13th International Workshop on Semantic Evaluation. pp. 683–690. Association for Computational Linguistics, Minneapolis, Minnesota, USA (Jun 2019). https://doi.org/10.18653/v1/S19-2122, https://www.aclweb.org/anthology/S19-2122
13. Malmasi, S., Zampieri, M.: Challenges in Discriminating Profanity from Hate Speech. Journal of Experimental & Theoretical Artificial Intelligence **30**, 1–16 (2018)
14. Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., Patel, A.: Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In: Proceedings of the 11th Forum for Information Retrieval Evaluation. pp. 14–17 (2019)
15. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. p. 3111–3119. NIPS'13, Curran Associates Inc., Red Hook, NY, USA (2013)
16. Mosquera, A.: Amsqr at SemEval-2020 task 12: Offensive language detection using neural networks and anti-adversarial features. In: Proceedings of the Fourteenth Workshop on Semantic Evaluation. pp. 1898–1905. International Committee for Computational Linguistics, Barcelona (online) (Dec 2020), https://www.aclweb.org/anthology/2020.semeval-1.250

17. Mosquera, A., Lloret, E., Moreda, P.: Towards facilitating the accessibility of web 2.0 texts through text normalisation. In: Proceedings of the LREC Workshop: Natural Language Processing for Improvign Textual Accessibility (NLP4ITA). pp. 9–14 (2012)
18. Mosquera, A., Moreda, P.: The use of metrics for measuring informality levels in web 2.0 texts. In: Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology (2011), https://www.aclweb.org/anthology/W11-4523
19. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543. Association for Computational Linguistics, Doha, Qatar (Oct 2014). https://doi.org/10.3115/v1/D14-1162, https://www.aclweb.org/anthology/D14-1162
20. Sabour, S., Frosst, N., Hinton, G.E.: Dynamic routing between capsules. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 3859–3869. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)
21. Seganti, A., Sobol, H., Orlova, I., Kim, H., Staniszewski, J., Krumholc, T., Koziel, K.: NLPR@SRPOL at SemEval-2019 task 6 and task 5: Linguistically enhanced deep learning offensive sentence classifier. In: Proceedings of the 13th International Workshop on Semantic Evaluation. pp. 712–721. Association for Computational Linguistics, Minneapolis, Minnesota, USA (Jun 2019). https://doi.org/10.18653/v1/S19-2126, https://www.aclweb.org/anthology/S19-2126
22. Taulé, M., Ariza, A., Nofre, M., Amigó, E., Rosso, P.: Overview of the detoxis task at iberlef-2021: Detection of toxicity in comments in spanish. In: Procesamiento del Lenguaje Natural, Vol. 67 (2021)
23. Vincent, J.: Youtube brings back more human moderators after ai systems overcensor (Sep 2020), Online.https://web.archive.org/web/20210515080450/https://www.theverge.com/2020/9/21/21448916/youtube-automated-moderation-ai-machine-learning-increased-errors-takedowns
24. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R.: SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In: Proceedings of The 13th International Workshop on Semantic Evaluation (SemEval) (2019)
25. Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., Çöltekin, c.: SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In: Proceedings of SemEval (2020)
26. Zhu, J., Tian, Z., Kübler, S.: UM-IU@LING at SemEval-2019 task 6: Identifying offensive tweets using BERT and SVMs. In: Proceedings of the 13th International Workshop on Semantic Evaluation. pp. 788–795. Association for Computational Linguistics, Minneapolis, Minnesota, USA (Jun 2019). https://doi.org/10.18653/v1/S19-2138, https://www.aclweb.org/anthology/S19-2138