GuillemGSubies at IberLEF-2021 DETOXIS task: Detecting Toxicity with Spanish BERT

Guillem García Subies[1]

Instituto de Ingeniería del Conocimiento, Francisco Tomás y Valiente st., 11 EPS, B Building, 5th floor UAM Cantoblanco. 28049 Madrid, Spain
guillem.garcia@iic.uam.es

Abstract. This paper describes a system created for the DETOXIS 2021 shared task, framed within the IberLEF 2021 workshop. We present an approach mainly based in fine-tuned BERT models using a Grid-Search and Data Augmentation with MLM substitution. This approach only takes into account the textual data from the dataset to prove the power of language models. Our models far outperform the baselines and achieve results close to the state-of-the-art.

Keywords: Toxicity Detection · BERT · Transformers · Data Augmentation · BETO

1 Introduction

Polarization can be a very problematic issue in society, especially on social media. There are manual mechanisms to report these behaviors, however they can be slow and inefficient. To address this, we can use NLP to detect automatically these undesirable toxic behaviors. The DETOXIS (DEtection of TOxicity in comments In Spanish) [15] shared task proposes, during this third edition of the IberLEF [10] workshop, a corpus to detect toxicity level in comments on internet forums and newspapers discussions.

This article summarizes our participation in all the DETOXIS tasks. Given the success of Transformer-inspired language models [16], both in academia and industry [17], we decided to use already pre-trained BERT [5] models. Specifically, we will use BETO [4] with some extra transfer learning techniques for ordinal classification problems and a hyperparameters Grid-Search. To address the problem of small data, we will use Data Augmentation techniques.

In the next section, we will briefly see some previous work related to this topic. In Section 3 we will go through a brief description of the tasks and the corpus. Then, in Section 4, we will explain the main ideas behind the proposed models. In Section 5 we will present a summary of the experiments we carried out and the results we got. Finally, in Section 6 we will expose the main conclusions of our work and results and we will also propose some ideas for future work.

IberLEF 2021, September 2021, Málaga, Spain.

2 Related Work

There is an extensive bibliography on Sentiment Analysis and text classification in social networks, however not that much work has been done about identifying and classifying toxic behaviors until 2019.

Most of the toxicity detection datasets are focused on classifying what kind of toxicity is present in the text, instead of the level of toxicity. Basile et al. [2] propose a task for the identification of toxicity against women and migrant people in Spanish and English Tweets. Struß et al., [13] also focus on Twitter, but now with German tweets. Other corpora focus in a more multilingual emphasis like Kumar et al. [8] and Zampieri et al. [18].

However, all have a common denominator. The best results have been obtained with some kind of Transformers or BERT-based model. For instance, the best models in Struß et al., [13] are fine-tuned BERTs with German pretraining on general data or fine-tuned BERTs pre-trained with specific German tweets. The best results in Kumar et al. [8] also point at BERT and Transformers, specifically a bootstrap aggregation of BERT models. Finally, the best results in Zampieri et al. [18] are also some kind of Transformer based models and ensembles of them.

This is a clear indicator of the trends in the state-of-the-art for this topic.

3 Tasks Description

The main corpus consists of 3463 comments posted in Spanish online newspapers and forums for the train split and 890 for the test one. They were collected from August 2017 to July 2020. Furthermore, the articles were selected taking into account their potential toxicity and the number of comments in them (more than 50 comments).

For the first task, the comments are annotated into two categories; **toxic** and *not toxic*. The second task consists of further classifying that toxicity into four levels of toxicity; **toxicity_level_0=not toxic**, **toxicity_level_1=mildly toxic**, **toxicity_level_2=toxic** and **toxicity_level_3=very toxic**.

In addition to the classification labels, for every sample there are also annotation about the argumentation, constructiveness, stance, target, stereotype, sarcasm, mockery, insult, improper language, aggressiveness and intolerance. However none of these are public in the test set, so they will be ignored in this study. Finally, for every comment, there is also a label indicating if the comment is a response to another comment or not. This information will not be used either because we will only focus on the textual data.

The metrics used to evaluate the results are the F-measure for the task1 and the Closeness Evaluation Metric (CEM) [1] for the second task. This last metric is very useful for ordinal classification problems given that it takes into account the order of the classes using concepts from Measurement Theory.

Class	Nº of samples
not toxic	2317
mildly toxic	808
toxic	269
very toxic	69

Table 1. Distribution of Samples

In the table above we can see the distribution of the samples in the train split. The most notable fact is that 67% of the samples are not toxic, so the dataset is unbalanced. For the second task, we can also see a very notable unbalance.

In the table below, we can see some illustrative examples of the data and their labels:

Los detuvieron en ronda malaga, un saludo	not toxic
Loss mas valientes, los que mejor cortan nuestras cabezas, Para vosotros, socialistas, izquierdistas, y no racistas,	mildly toxic
Esto es lo que importas cuando los rescatas en lugar de hundirlos.	toxic
Está claro que vienen los mejores. Haced que pase putos rojos de mierda.	very toxic

Table 2. Examples of the different classes

4 Models

4.1 Data Preprocessing

We performed a simple preprocessing where we substituted some expressions with a more normalized form:

- Every URL was replaced with the token "[URL]" so we don't get strange tokens when the tokenizer tries to process and URL. Furthermore, no semantic information about toxicity can be inferred from a URL, the only information relevant for the model is that there is a URL in that token.
- Finally we normalized every laugh ("jasjajajajj" → "haha") so we minimize the noise of the misspellings, common in social networks.

4.2 Baselines

We created some baselines so we can compare our models properly. We selected a HashingVectorizer + RandomForest. This way, we can compare our models to a classic feature extraction model.

4.3 Language Models

We used BETO [4], a BERT model trained with the Spanish Unannotated Corpora (SUC) [3] that has proven to be much better than the multilingual BERT model.

We tried different training strategies given that the classes are related to each other:

- The simplest approach we tried is treating both tasks as the same one, with a multiclass classification model. For the first task, everything different from **not toxic** would be considered **toxic**.
- For the second approach, we first trained a binary classification model to distinguish between **not toxic** and **toxic** for the first task. Then we trained a multiclass model to classify between the three levels of toxicity.
- Similarly to the last approach, we then tried to transform the second task into three different binary classification problems; classifying between **not toxic** and the rest of the classes, **mildly toxic** and **toxic** or **very toxic**, and between **toxic** and **very toxic**. With this, we tried to have very specific models that can differentiate slight changes in toxicity.
- As there are not too many samples for the last models in the previous approach to learn correctly, we also tried with a transfer learning approach, similar to the one presented by Sun et al. [14]. Instead of using always the same BETO pretrained model for every finetuned model, we used the finetuned model from the step before, i.e. the model that classifies **mildly toxic** and **toxic** has as base model the one that classifies **not toxic** and **mildly toxic**.

In addition, for the fine-tuning process, we carried out a Grid-search optimization over the main parameters of the neural network: learning rate, batch size and dropout rate. The search was performed with a 5-fold stratified cross-validation with the following grid: Learning rate, $(1e-6, 1e-5, 3e-5, 5e-5, 1e-4)$; batch size, $(8, 16, 32)$ and dropout rate, $(0.08, 0.1, 0.12)$. The best parameters for both models were: learning rate, $1e-5$; batch size, 16 and dropout rate, 0.1.

4.4 Data Augmentation

As the dataset is relatively small, we decided to run Data Augmentation techniques. The selected strategy was the Data Augmentation through the masking of words with a Masked Language Model, BETO.

For every sample in the dataset, we randomly masked 15% of the tokens and used BETO to predict them, creating a modified sample. With this method, we obtained double the amount of the original samples.

5 Experiments and Results

5.1 Experimental Setup

We trained all the models with a NVIDIA Tesla P100-PCIE-16GB GPU and a Intel(R) Xeon(R) CPU E5-2640 v4 @ 2.40GHz CPU with 500GB of RAM memory.

The software we used was Python3.8, transformers 4.5.1 [17], pytorch 1.8.1 [11], scikit-learn 0.24.1 [12] and nlpaug [9] 1.1.3.

5.2 Results

In the Table 3 we can see the results for our models in the test set of the first task. Note that the ChainBOW baseline, Word2VecSpacy baseline and the SINAI_team (winner of the task) results are taken from the task Overview [15]. Our runs for this task are *BETO-multiclass* and *BETO-binary* both with and without data augmentation as explained in Section 4.3. Note that some of the results presented here were obtained after the labeled test set was published so we could analyze in depth our models.

We can see that there is almost no difference between the binary and multiclass models. This might happen because there is a great difference in the amount of samples and toxicity between the not toxic comments and the rest, which makes them easy to indentify in every situation. Finally we can see the the Data Augmentation strategy obtained around 0.02 points more than the models without augmentation. Our result was the second best in the competition, which proves that the simplicity of using BETO with some Grid-Search can yield really good results.

Model	**F-measure**
Word2VecSpacy	0.1523
ChainBOW	0.3747
HV+RF	0.4159
BETO-multiclass	0.5721
BETO-binary	0.5777
BETO-multiclass-aug	0.5981
BETO-binary-aug	0.6000
SINAI_team	0.6461

Table 3. Results for task1

For the second task, the results were similar to the ones obtained in the first task. In the Table 4 we can look at them in more detail. Again, ChainBOW baseline, Word2VecSpacy baseline and the SINAI_team results are taken from the task Overview [15]. We can see that our transfer learning approach (*BETO-transfer*) obtains better results than the other approaches and that there is

almost no difference between the simple multiclass approach and the one that first detects the not toxic comments (*BETO-2models-aug*). These results are in line with the ones in the first task, showing that adding the not toxic class to the models, will not make them worse.

This results are placed fifth among all the participating teams (24), which proves that our approach, given it's simplicity and the lack of any linguistic analysis, is very good.

Model	**CEM**
Word2VecSpacy	0.6116
HV+RF	0.6214
ChainBOW	0.6535
BETO-2_models-aug	0.6891
BETO-multiclass-aug	0.6913
BETO-3_models-aug	0.704
BETO-transfer	0.7172
BETO-transfer-aug	0.7189
SINAI_team	0.7495

Table 4. Results for task2

6 Conclusions and Future Work

Through this shared task, we have seen that NLP can be of great help in detecting and classifying unwanted toxic behavior in social networks and there is still a long way to go.

The results obtained by our systems are very promising given their great performance and their simplicity. This compilation of methods is very significant because it could lead to much better results when combined with other improvements from the state-of-the-art.

We believe that our results could improve a lot using specific language models trained with corpora from social networks like TWilBert [6]. Another interesting approach would be to use a general language model and further pre-train it with corpora from the same domain [14] as the final task. Finally, we have proven that good hyperparameters are also key for a good neural network so a better search, like the Population Based Training [7], would further improve the model.

Acknowledgments

This work has been partially funded by the Instituto de Ingeniería del Conocimiento (IIC) and the hardware used was also provided by the IIC.

References

1. Amigó, E., Gonzalo, J., Mizzaro, S., de Albornoz, J.C.: An effectiveness metric for ordinal classification: Formal properties and experimental results (2020)
2. Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F.M., 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. Association for Computational Linguistics, Minneapolis, Minnesota, USA (Jun 2019). https://doi.org/10.18653/v1/S19-2007, https://www.aclweb.org/anthology/S19-2007
3. Cañete, J.: Compilation of large spanish unannotated corpora (May 2019). https://doi.org/10.5281/zenodo.3247731, https://doi.org/10.5281/zenodo.3247731
4. Cañete, J., Chaperon, G., Fuentes, R., Pérez, J.: Spanish pre-trained bert model and evaluation data. In: to appear in PML4DC at ICLR 2020 (2020)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2018)
6. Ángel González, J., Hurtado, L.F., Pla, F.: Twilbert: Pre-trained deep bidirectional transformers for spanish twitter. Neurocomputing (2020). https://doi.org/https://doi.org/10.1016/j.neucom.2020.09.078, http://www.sciencedirect.com/science/article/pii/S0925231220316180
7. Jaderberg, M., Dalibard, V., Osindero, S., Czarnecki, W.M., Donahue, J., Razavi, A., Vinyals, O., Green, T., Dunning, I., Simonyan, K., Fernando, C., Kavukcuoglu, K.: Population based training of neural networks (2017)
8. Kumar, R., Ojha, A.K., Malmasi, S., Zampieri, M.: Evaluating aggression identification in social media. In: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying. pp. 1–5. European Language Resources Association (ELRA), Marseille, France (May 2020), https://www.aclweb.org/anthology/2020.trac-1.1
9. Ma, E.: Nlp augmentation. https://github.com/makcedward/nlpaug (2019)
10. Montes, M., Rosso, P., Gonzalo, J., Aragón, E., Agerri, R., Ángel Álvarez Carmona, M., Álvarez Mellado, E., de Albornoz, J.C., Chiruzzo, L., Freitas, L., Adorno, H.G., Gutiérrez, Y., Zafra, S.M.J., Lima, S., de Arco, F.M.P., Taulé, M.: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021). In: CEUR Workshop Proceedings (2021)
11. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc. (2019), http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
12. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
13. Struß, J.M., Siegel, M., Ruppenhofer, J., Wiegand, M., Klenner, M.: Overview of germeval task 2, 2019 shared task on the identification of offensive language.

In: Preliminary proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019), October 9 – 11, 2019 at Friedrich-Alexander-Universität Erlangen-Nürnberg. pp. 352 – 363. German Society for Computational Linguistics & Language Technology und Friedrich-Alexander-Universität Erlangen-Nürnberg, München [u.a.] (2019), https://nbn-resolving.org/urn:nbn:de:bsz:mh39-93197
14. Sun, C., Qiu, X., Xu, Y., Huang, X.: How to fine-tune bert for text classification? (2020)
15. 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. Procesamiento del Lenguaje Natural **67** (2021)
16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2017)
17. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. Association for Computational Linguistics, Online (Oct 2020), https://www.aclweb.org/anthology/2020.emnlp-demos.6
18. Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., Çöltekin, Ç.: SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In: Proceedings of the Fourteenth Workshop on Semantic Evaluation. pp. 1425–1447. International Committee for Computational Linguistics, Barcelona (online) (Dec 2020), https://www.aclweb.org/anthology/2020.semeval-1.188