SINAI at IberLEF-2021 DETOXIS task: Exploring Features as Tasks in a Multi-task Learning Approach to Detecting Toxic Comments

Flor Miriam Plaza-del-Arco, M. Dolores Molina-González, L. Alfonso Ureña-López, and M. Teresa Martín-Valdivia

Department of Computer Science, Advanced Studies Center in ICT (CEATIC)
Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
{fmplaza, mdmolina, laurena, maite}@ujaen.es

Abstract. This paper describes the participation of the SINAI research group at DETOXIS (DEtection of TOxicity in comments In Spanish) shared task at IberLEF 2021. The proposed system follows a Multi-task Learning approach where multiple tasks related to toxic comments identification are learned in parallel while using a shared representation. Specifically, we use the dataset features provided by the organizers as tasks along with the combination of polarity classification, emotion classification and offensive language detection tasks to explore if they help in the identification of toxic comments. Our proposal ranked first in both DETOXIS subtasks, toxicity detection and toxicity level detection.

Keywords: Multi-Task Learning · BERT · Toxic Features · Sentiment Analysis.

1 Introduction

Toxic comment classification is a field of research that has attracted increasing interest in the Natural Language Processing (NLP) community in recent years. In this task, the organizers defined a toxic comment as "a comment that denigrates, hates or vilifies, attacks, threatens, insults, offends or disqualifies a person or group of people based on characteristics such as race, ethnicity, nationality, political ideology, religion, gender and sexual orientation, among others". Therefore, toxicity term will be used as an umbrella term to include different definitions used in the literature to describe hate speech [5, 4], abusive [20], aggressive [14], and offensive language [28]. In fact, these different terms address different aspects of toxic language [24].

IberLEF 2021, September 2021, Málaga, Spain.

Detecting online toxicity may be difficult, as it can be expressed in different ways: explicitly (through insult, mockery or inappropriate humor) or implicitly (through sarcasm). Another aspect to take into account is the the presence of different levels of intensity in toxicity (from rude and offensive comments to more aggressive ones, the latter being those that incite hatred or even physical violence).

In this paper, we present the systems we developed as part of our participation in the DETOXIS (DEtection of TOxicity in comments In Spanish) shared task [26], at IberLEF 2021 [19] in both subtasks. The aim of DETOXIS is the detection of toxicity in comments posted in Spanish in response to different online news articles related to immigration. The DETOXIS task is divided into two related classification subtasks: (1) toxicity detection and (2) toxicity level detection. The first subtask consists of detecting whether or not a comment is toxic while the second one aims to categorize the comment according to four levels of toxicity (0: not toxic, 1: mildly toxic, 2: toxic, and 3: very toxic).

The rest of the paper is structured as follows. In Section 2 we explain the data used in our experiments. In Section 3, we describe our proposal to address the task. In Section 4 and 5, we present the experiment setup and results, respectively. Finally, the conclusion is presented in Section 6.

2 Corpora

To run our experiments, we used the Spanish dataset provided by the organizers of the DETOXIS task at IberLEF 2021. The DETOXIS dataset was collected from the NewsCom-TOX dataset. This dataset consists of 4,357 comments (approximately) posted in response to different articles extracted from Spanish online newspapers (ABC, elDiario.es, El Mundo, NIUS, etc.) and discussion forums (such as Menéame) from August 2017 to July 2020. These articles were manually selected taking into account their controversial subject matter, their potential toxicity and the number of comments posted (minimum 50 comments). A keyword-based approach was used to search for articles primarily related to immigration. Comments were selected in the same order in which they appear in the web timeline. The author (anonymous), date and time the comment was posted are also retrieved. The number of comments ranged from 65 to 359 comments per article. On average, approximately 30% of the comments are toxic. Each comment was annotated into two categories "toxic" and "non-toxic", and subsequently those annotated as "toxic" were assigned with different toxicity levels (non-toxic, slightly toxic, toxic, and very toxic). In addition, the following characteristics were also annotated: argumentation, constructiveness, stance, target, stereotype, sarcasm, mockery, insult, improper language, aggressiveness and intolerance. All of these characteristics (or categories) have a binary classification, except for the level of toxicity. Each comment was annotated by three annotators and, once all comments for each item were annotated, an inter-annotator agreement test was performed.

In addition, we used in our experiments other corpora corresponding to tasks that could be related to detection of toxicity from social media including polarity classification (InterTASS), emotion classification (EmoEvent and Universal Joy), HS identification (HatEval and HaterNet), and aggressiveness detection (MEX-A3T). The datasets are described below:

- **International TASS Corpus (InterTASS)** was released in 2017 [17] with Spanish tweets and updated in 2018 with texts written in three different variants of Spanish from Spain, Costa Rica and Peru [16]. In 2019, InterTASS was enlarged with new texts written in two new Spanish variants: Uruguayan and Mexican [10] and finally, it was completed with Chilean-Spanish Tweets in 2020 [13]. The corpus released in 2019 is the one used in this paper. At least three annotators annotated each tweet with its level of polarity, which could be labeled as positive, negative, neutral and none.
- **EmoEvent** [23] is a multilingual emotion dataset based on events that took place in April 2019. It focuses on tweets in the areas of entertainment, catastrophes, politics, global commemoration and global strikes. For the creation of the corpus, the authors collected Spanish and English tweets from the Twitter platform. Then, each tweet was labeled with one of seven emotions, six Ekman's basic emotions plus the "neutral or other emotions" label. Focusing on the Spanish language, a total of 8,409 were labeled by three Amazon Mechanical Turkers.
- **Universal Joy**[15] is a new data set of over 530k anonymized public Facebook posts across 18 languages. It was collected in October 2014 by searching for public Facebook posts with a Facebook "feelings tag", and labeled with five different emotions: *anger*, *anticipation*, *fear*, *joy*, and *sadness*. There is a wide variety in the amount of data per language, ranging from 284,265 posts for English, the most frequent language, to 869 posts for Bengali. We used the 31,326 Spanish posts.
- **HatEval** was provided by organizers in SemEval 2019 Task 5 [5]. The task consisted in detecting hateful content in Twitter posts, against two targets: women and immigrants. For the creation of the corpus, the data was collected using a different time frame. The majority of tweets against women were derived from an earlier collection made in the context of two earlier challenges on misogynistic speech identification, whose collection phase began on July 2017 and ended on November 2017 [12, 11]. The remaining tweets were collected from July to September 2018. The dataset contains tweets composed of an identifier, the text of the tweet and the mark of HS, which is 0 if the text is not hateful and 1 if the text is hateful speech against women or immigrants.
- **HaterNet** [22] was built for the intelligent system of the same name, used by the National Office against Hate Crimes of the Spanish Secretary of State for Security. For the creation of this corpus, over 2 million tweets originated in Spain on different random dates between February 2017 and December 2017 were collected. Subsequently, the tweets were filtered using six HS dictionaries and one dictionary containing generic insults. After this, only 6000

tweets were selected due to time restrictions, to be manually labeled by four experts with different backgrounds and in case of a tie a fifth person, cast the deciding vote. Finally, out of the 6000 tweets, 1,567 were labeled as hateful and 4,433 as non-hateful.

– **MEX-A3T** [3]. It was provided by the organizers in IberEval 2018: Authorship and aggressiveness analysis in Mexican Spanish tweets [1]. They built a corpus of tweets to detect aggressiveness from Mexican accounts collected from August to November of 2017. In order to extract the tweets, they selected a set of terms that served as seeds. Then, they used both words non-colloquial in the Dictionary of Mexicanisms and classified as vulgar. The hashtags were related to sexism, homophobia, politics and discrimination. They used Mexico City as the center and extracted all tweets that were within a radius of 500 km. Finally, two people labeled the collected tweets. The dataset contains tweets composed of an identifier, the text of the tweet, and the mark of aggressiveness, being 0 if the tweet is not-aggressive and 1 if the tweet is aggressive.

3 System overview

In this section, we describe the systems developed for the DEtection of TOxicity shared task in Spanish comments at IberLEF 2021.

We propose a Multi-Task Learning (MTL) system using the well-known Transformer-based model BERT which has been proven to be very successful in many natural language processing tasks.

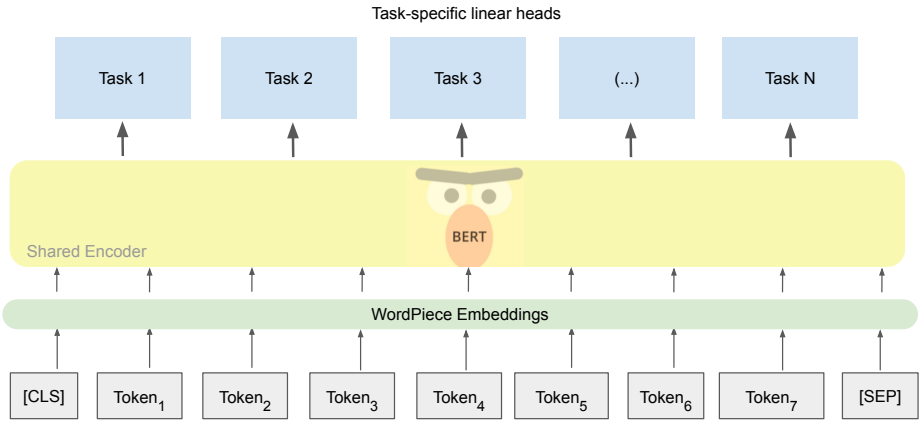
In the MTL scenario, the goal is to learn multiple tasks simultaneously instead of learning them separately in order to improve performance on each task [7]. These tasks are usually related, although they may have different data or features. By sharing representations across related tasks, we can allow our model to better generalize to our original task. In this study, we used tasks related to the toxicity comment detection task. These tasks include hate speech detection, offensive language identification, polarity classification, and emotion classification, sharing the same type of source: social media platforms (Twitter and Facebook). Moreover, we consider each of the features provided in the DETOXIS dataset (constructiveness, argumentation, mockery, sarcasm, positive stance, negative stance, target person, target group, stereotype, insult, improper language, aggressiveness, intolerance) as specific tasks to train our system. From now, we will refer to all these specific tasks (the ones provided in the DETOXIS dataset) as tasks related to toxicity comments features.

To develop the MTL system, we follow the most widely used technique in neural networks introduced by [7], the hard parameter sharing approach. It consists of a single encoder that is shared and updated between all tasks, while keeping a few task-specific layers to specialize in each task. [25].

The general architecture of the MTL-BERT model is shown in Figure 1. The shared layers are based on BERT [9]. Following Devlin et al., 2018, in the first step, all the inputs are converted to WordPieces [27], two additional tokens are

added at the start ([CLS]) and end ([SEP]) of the input sequence, respectively. In the shared layers, the BERT model first converts the input sequence to a sequence of embedding vectors. This semantic representation is shared across all tasks. Then, on top of the shared BERT layers, the task-specific output heads are created for each task, and task heads are attached to a common sentence encoder. Finally, the layers are fine-tuned according to the given set of downstream tasks.

Fig. 1. Proposed MTL system for the DETOXIS task.



4 Experimental setup

4.1 Dataset preprocessing

We perform a social media specific data cleaning in the corpora related to Twitter and Facebook (InterTASS, EmoEvent, Universal Joy, HatEval, HaterNet, MEX-A3T) before including the texts in the models. The following practices to prepare the text for deep learning experiments have been carried out using the ekphrasis module [6]:

- URLs, emails, users' mentions, percentages, monetary amounts, time and date expressions, and phone numbers are normalized.
- Hashtags are unpacked and split to their constituent words.
- Elongated words and repeated characters in words are annotated and reduced.
- Emojis are converted to its alias.

As the DETOXIS task dataset provided by the organizers includes responses to different articles extracted from Spanish online newspapers, we performed a different data cleaning that includes the following steps:

– Remove URLs, hashtags and users' mentions.
– Reduce words with more than 4 repeated characters to 3 repetitions.
– Remove multiple spaces.
– Remove texts with only numbers.

4.2 System settings

All the models were implemented using PyTorch, a high-performance deep learning library [21] based on the Torch library. The experiments were run on a single Tesla-V100 32 GB GPU with 192 GB of RAM.

During the evaluation phase, we train the model on the training set provided by the organizers, then we evaluate it on the test set.

Regarding our participation, we submitted five runs using the proposed MTL-based system. The details of the modules and the differences of the five settings are described below.

– **Run 1**. In order to establish a baseline in our study and compare the results with the MTL scenario, the first run correspond to our baseline, a single-task learning approach which involves only the DETOXIS dataset. For this setting, we use the well-known Transformer BERT.
– **Run 2**. In this setting, our goal is to train the MTL system on the tasks which are related to the identification of toxicity comments. Specifically, HS identification, offensive language detection and the toxicity comments features (constructiveness, argumentation, mockery...) explained in Section 3. Our assumption is that all these tasks are related to the inappropriate behavior on the web, therefore the knowledge share during training among these tasks may benefit to the task of toxic comments identification even if the texts correspond to different language registers from social media and newspapers.
– **Run 3**. This configuration includes run 2 but with the addition of a new task: polarity classification. Our goal is to leverage on the sentiment expressed in the posts to aid in the classification of toxic comments. Our assumption is that toxicity is associated with a negative polarity, then the knowledge share can help to detect easily toxic comments. For the polarity classification task, we use the InterTASS dataset.
– **Run 4**. This configuration includes run 2 but with the addition of a new task: emotion classification. In this setting, our goal is to leverage in the identification of emotion categories to aid in the classification of toxic comments. Our assumption is that negative emotions such as *anger*, *fear*, *sadness* and *disgust* could be related to toxicity while positive emotions are not. For the emotion analysis task, we use the EmoEvent and Universal Joy datasets.
– **Run 5**. In this setup, we have included the polarity and emotion classification tasks in run 2. Therefore, in this setting the MTL system is trained on the different tasks explainded above. We expect that the combination of all the tasks helps to identify toxic comments.

In all the runs, the DETOXIS training set has also been used to train the MTL system, and then we have evaluated the shared task using the DETOXIS test set.

Since the DETOXIS dataset is composed of Spanish texts, while training the MTL system we use the BETO model [8] trained on Spanish texts. We employ the following hyperparameters in the five runs: learning rate as 2e-05, batch size as 16, dropout probability as 0.01, the optimization algorithm Adamw, and maximum epoch as 3.

5 Results

In this section we present the results obtained by the different runs we have explored in both subtasks of the competition. In order to evaluate them we use the official competition metrics for subtask 1 (F-measure) and subtask 2 (Closeness Evaluation Metric (CEM) [2]). In addition, for the level detection subtask, the organizers has provided evaluation results with Rank Biased Precision (RBP) [18], Pearson coefficient, and Accuracy (Acc).

We evaluated our five runs on subtasks 1 and 2 of the DETOXIS shared task. The results obtained are shown in Table 1 and 2, respectively. As can be seen, in subtask 1, the different settings of the MTL system have outperformed our baseline BETO (Run 1). It should be noted that the best setting in both subtasks is Run 5 in which all tasks related to toxicity detection in comments are combined. Specifically, in subtask 1, run 5 outperforms with a substantial margin (3,77%) our baseline BETO. For subtask 1 it can also be observed that Run 3 achieves remarkable results, and Run 4 surpassed the baseline BETO, therefore we can confirm our hypothesis that sentiment analysis helps the task of detecting toxicity in comments. Regarding subtask 2, runs 4 and 5 surpasses the baseline BETO in terms of CEM score, which means that emotion analysis along the combination of HS identification, offensive language detection and the tasks related to toxicity comments features (argumentation, constructiveness, sarcasm, mockery...) could benefit the detection of toxicity comments.

It should be remarked that although the datasets of some of the tasks explored (sentiment analysis, hate speech detection and offensive language identification) include posts from social media, the MTL seems to be able to transfer the knowledge to a different language register employ in the comments posted in response to different articles extracted from online newspapers (DETOXIS dataset).

Finally, our results in the competition for both subtasks among the participants (Table 3 and Table 4) show the success of our proposed model achieving the first place in the ranking in both subtasks. The representations computed by the encoder embed the affective knowledge and the knowledge related to toxicity detection tasks (offensive language, constructiveness, sarcasm, among others) allows the MTL model to identify toxic comments more accurately.

Table 1. Results in subtask 1 on the test set of DETOXIS shared task. Best result is marked in bold.

Run	F-measure
1	0.6084
2	0.6172
3	0.6406
4	0.6125
5	**0.6461**

Table 2. Results in subtask 2 on the test set of DETOXIS shared task. Best result is marked in bold.

Run	CEM	RBP	Pearson	Acc
1	0.7421	0.2722	0.5065	0.7419
2	0.7344	0.3425	0.4638	0.7441
3	0.7389	0.2499	0.4580	0.7396
4	0.7425	0.2361	0.4892	0.7553
5	**0.7495**	0.2612	0.4957	**0.7654**

Table 3. Ranking of participants' systems in subtask 1 of DETOXIS shared task.

Ranking	Team	F-measure
1	**SINAI (run 5)**	**0.6461**
2	GuillemGSubies	0.6000
3	AI-UPV	0.5996
12	ToxicityAnalizers	0.4562
-	BOWClassifier	0.1837
31	JOREST	0.0246

Table 4. Ranking of participants' systems in subtask 2 of DETOXIS shared task.

Ranking	Team	CEM	RBP	Pearson	Acc
1	**SINAI (run 5)**	**0.7495**	**0.2612**	**0.4957**	**0.7654**
2	Team Sabari	0.7428	0.2670	0.5014	0.7464
3	DCG	0.7300	0.3925	0.4544	0.7329
11	ToxicityAnalizers	0.6332	0.0709	0.1805	0.6139
-	BOWClassifier	0.6318	0.1657	0.1688	0.7329
24	JosepCarles_LNR	0.5376	0.0705	0.0072	0.4949

6 Conclusion

This paper presents the participation of the SINAI research group at the DEtection of TOxicity in comments in Spanish shared task at IberLEF 2021. Our proposal explores how transferred knowledge from tasks related to the identi-

fication of toxicity language (polarity classification, emotion classification, hate speech detection, offensive language detection, constructiveness, argumentation, sarcasm, mockery, etc.) may help in a text classification task like DETOXIS. Experiments conducted show the efficacy of our proposed approach in achieving convincing performance in both subtasks. Further exploration on how and which of the features we use in our MTL approach (constructiveness, argumentation, sarcasm, mockery, etc.) helps to the identification of toxic comments are left as future work, and we welcome the community to contribute.

Acknowledgement

This work has been partially supported by a grant from European Regional Development Fund (FEDER), LIVING-LANG project [RTI2018-094653-B-C21], and Ministry of Science, Innovation and Universities (scholarship [FPI-PRE2019-089310]) from the Spanish Government.

References

1. Álvarez-Carmona, M.Á., Guzmán-Falcón, E., Montes-y-Gómez, M., Jair-Escalante, H., Villaseñor-Pineda, L., Reyes-Meza, V., Rico-Sulayes, A.: Overview of MEX-A3T at ibereval 2018: Authorship and aggressiveness analysis in mexican spanish tweets. In: Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018. CEUR Workshop Proceedings, vol. 2150, pp. 74–96. CEUR-WS.org (2018)
2. Amigó, E., Gonzalo, J., Mizzaro, S., Carrillo-de Albornoz, J.: An effectiveness metric for ordinal classification: Formal properties and experimental results. arXiv preprint arXiv:2006.01245 (2020)
3. Aragón, M.E., Jarquín-Vásquez, H.J., Montes-y-Gómez, M., Escalante, H.J., Villaseñor-Pineda, L., Gómez-Adorno, H., Posadas-Durán, J.P., Bel-Enguix, G.: Overview of MEX-A3T at IberLEF 2020: Fake News and Aggressiveness Analysis in Mexican Spanish. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain, September 23th, 2020. CEUR Workshop Proceedings, vol. 2664, pp. 222–235. CEUR-WS.org (2020)
4. Plaza-del Arco, F.M., Molina-González, M.D., Ureña-López, L.A., Martín-Valdivia, M.T.: Comparing pre-trained language models for spanish hate speech detection. Expert Systems with Applications **166**, 114120 (2021)
5. 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

6. Baziotis, C., Pelekis, N., Doulkeridis, C.: Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). pp. 747–754. Association for Computational Linguistics, Vancouver, Canada (August 2017)
7. Caruana, R.: Multitask learning. Machine learning **28**(1), 41–75 (1997)
8. Cañete, J., Chaperon, G., Fuentes, R., Ho, J.H., Kang, H., Pérez, J.: Spanish pre-trained bert model and evaluation data. In: PML4DC at ICLR 2020 (2020)
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2018)
10. Díaz-Galiano, M.C., García-Vega, M., Casasola, E., Chiruzzo, L., García-Cumbreras, M.Á., Martínez-Cámara, E., Moctezuma, D., Montejo-Ráez, A., Sobrevilla-Cabezudo, M.A., Sadit-Tellez, E., et al.: Overview of TASS 2019: One More Further for the Global Spanish Sentiment Analysis Corpus. In: IberLEF@ SEPLN. pp. 550–560 (2019)
11. Fersini, E., Nozza, D., Rosso, P.: Overview of the evalita 2018 task on automatic misogyny identification (ami). EVALITA Evaluation of NLP and Speech Tools for Italian **12**, 59 (2018)
12. Fersini, E., Rosso, P., Anzovino, M.: Overview of the task on automatic misogyny identification at ibereval 2018. In: Rosso, P., Gonzalo, J., Martínez, R., Montalvo, S., de Albornoz, J.C. (eds.) Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018). CEUR Workshop Proceedings, vol. 2150, pp. 214–228. CEUR-WS.org (2018)
13. García-Vega, M., Díaz-Galiano, M.C., García-Cumbreras, M.Á., Plaza-del-Arco, F.M., Montejo-Ráez, A., Jiménez-Zafra, S.M., Martínez-Cámara, E., et al.: Overview of TASS 2020: Introducing emotion detection. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020), Málaga, Spain, September 23th, 2020. CEUR Workshop Proceedings, vol. 2664, pp. 163–170. CEUR-WS.org (2020)
14. 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 (2020)
15. Lamprinidis, S., Bianchi, F., Hardt, D., Hovy, D.: Universal joy a data set and results for classifying emotions across languages. In: Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. pp. 62–75 (2021)
16. Martínez-Cámara, E., Almeida-Cruz, Y., Díaz-Galiano, M.C., Estévez-Velarde, S., García-Cumbreras, M.Á., García-Vega, M., Gutiérrez, Y., Montejo-Ráez, A., Montoyo, A., Muñoz, R., Piad-Morffis, A., Villena-Román, J.: Overview of TASS 2018: Opinions, health and emotions. In: Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN, TASS@SEPLN 2018. CEUR Workshop Proceedings, vol. 2172, pp. 13–27. CEUR-WS.org (2018)
17. Martínez-Cámara, E., Díaz-Galiano, M.C., García-Cumbreras, M.A., García-Vega, M., Villena-Román, J.: Overview of TASS 2017. Proceedings of TASS pp. 13–21 (2017)
18. Moffat, A., Zobel, J.: Rank-biased precision for measurement of retrieval effectiveness. ACM Transactions on Information Systems (TOIS) **27**(1), 1–27 (2008)
19. Montes, M., Rosso, P., Gonzalo, J., Aragón, E., Agerri, R., Álvarez Carmona, M., Álvarez Mellado, E., Carrillo-de Albornoz, J., Chiruzzo, L., Freitas, L.,

Gómez Adorno, H., Gutiérrez, Y., Jiménez-Zafra, S.M., Lima, S., Plaza-del-Arco, F.M., Taulé, M. (eds.): Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) (2021)
20. Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., Chang, Y.: Abusive language detection in online user content. In: Proceedings of the 25th international conference on world wide web. pp. 145–153 (2016)
21. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in neural information processing systems. pp. 8026–8037 (2019)
22. Pereira-Kohatsu, J.C., Quijano-Sánchez, L., Liberatore, F., Camacho-Collados, M.: Detecting and monitoring hate speech in twitter. Sensors **19**(21), 4654 (2019)
23. Plaza-del-Arco, F., Strapparava, C., Ureña-Lopez, L.A., Martin-Valdivia, M.T.: EmoEvent: A Multilingual Emotion Corpus based on different Events. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 1492–1498. European Language Resources Association, Marseille, France (May 2020), https://www.aclweb.org/anthology/2020.lrec-1.186
24. Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., Patti, V.: Resources and benchmark corpora for hate speech detection: a systematic review. Language Resources and Evaluation pp. 1–47 (2020)
25. Ruder, S.: Neural transfer learning for natural language processing. Ph.D. thesis, NUI Galway (2019)
26. 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)
27. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al.: Google's neural machine translation system: Bridging the gap between human and machine translation. CoRR **abs/1609.08144** (2016)
28. 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). arXiv preprint arXiv:2006.07235 (2020)