Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1533–1537
Marseille, 11–16 May 2020

Abusive language in Spanish children and young teenager's conversations: data preparation and short text classification with contextual word embeddings

Marta R. Costa-jussà, Esther Gonzalez, Asuncion Moreno, Eudald Cumalat
TALP Research Center
Universitat Politècnica de Catalunya, Barcelona
marta.ruiz@upc.edu,esthergonzalezbenitez@gmail.com
asuncion.moreno@upc.edu, eudaldcumalat@gmail.com

Abstract

Abusive texts are reaching the interests of the scientific and social community. How to automatically detect them is one question that is gaining interest in the natural language processing community. The main contribution of this paper is to evaluate the quality of the recently developed "Spanish Database for cyberbullying prevention" for the purpose of training classifiers on detecting abusive short texts. We compare classical machine learning techniques to the use of a more advanced model: the contextual word embeddings in the particular case of classification of abusive short-texts for the Spanish language. As contextual word embeddings, we use Bidirectional Encoder Representation from Transformers (BERT), we show that it mostly outperforms classical techniques. Far beyond the experimental impact of our research, this project aims at planting the seeds for an innovative technological tool with a high potential social impact and aiming at being part of the initiatives in artificial intelligence for social good.

1. Introduction

Many studies show that the unappropriated behaviour of children and young teenagers is increasing, due to their easy access to Internet information, among other reasons (Saba, 2018). Facebook, Twitter or Instagram are some of the many applications that they are using nowadays, without taking into account how other people may feel about their posts, making cyberbullying one of the problems for young people. There are many studies focusing on automatic abusive or cyberbullying text detection (Salawu et al., 2017). Approaches vary from using classical machine learning algorithms (Van Hee et al., 2018) to using neural networks models and specifically using different word embedding features (Chen et al., 2017), among others. Additionally, there is recent work (Zhu et al., 2019) using contextual word embeddings (Devlin et al., 2018) which is reaching great success in many downstream Natural Language Processing (NLP) applications. While most of the research is done for English, there is a recent work for Spanish abusive text classification (Mercado et al., 2018) focusing on large texts and using Support Vector Machines.

Differently from previous work, the purpose of this study is to automatically analyse the teenager's mood according to their short texts for Spanish and using latest NLP techniques (BERT on its multilingual version). In turn, we want to know if we can achieve this objective by training on the database created in the framework of the project "Safeguarding children online" where project partners were: d-LAB, SafeToNet (London, UK), Orange and the Innovation and Technology Center of the Universitat Politècnica de Catalunya (CIT-UPC).

The database has the following annotated sentiments: aggression, anxiety, depression, distress, substances, sexuality and violence. While having all these sentiments annotated, we have a very low representation on each. Therefore, in this first study, we limit our scope to obtain a binary result for each analysed entry and determine either a correct or misbehaviour in order to take action helping the teenager. Basically, the program is able to detect if an entry is promoting the cyberbullying and the teenager needs some kind of parental help. Results reach up to and F1 score of 80%. Additionally, and as an approximation to the fully fine-grained classification, we provide an aditional classification taking into account 2 different classes which are: aggression-violence and distress-anxiety-depression.

2. Background

In this section, we briefly describe the existing techniques that we are using for our particular task in the experimental section. We have chosen Support vector machine (Hearst, 1998) and Random forest (Breiman, 2001) techniques because they have achieved satisfactory results in classification tasks. These two techniques are the contrastive techniques with a more recent one, Bidirectional Encoder Representations from Transformers model (Devlin et al., 2018), that is just emerging for classification purposes as well.

2.1. Support vector machine

Support vector machine (SVM) classification algorithm (Hearst, 1998) uses a hyperplane to separate the data into different classes. This hyperplane is set in the middle of the support vectors, which are the nearest points of each class. All the new inputs are mapped to the same space and the new points are determined to one class or the other depending on the side of the gab they are assigned to. Nevertheless, for most of the real-life problems, it is not possible to set a hyperplane as a linear separator. For that reason, Kernel functions are used to convert a non-linear problem from a determined space to a linear one by projecting the space into a higher dimensional one.

2.2. Random forest

A decision tree is the method that uses a tree-like model to evaluate inputs and extract results depending on the probability of decisions. The main goal is to split all the information to other subsections to analyze it as deep as possible. Having it into account, Random Forest (RF) (Breiman, 2001) is composed of multiple decision trees that are working as a group. Separately, each tree is splitting the information using different conditions. For the classification model, the final prediction is the class that appeared in the majority of the trees.

2.3. Bidirectional Encoder Representations from Transformers model

BERT (Devlin et al., 2018) is a model that pre-trains language understanding models based on some given large text. It comes in different flavours, including a model available for many languages, because all the input data for the unsupervised learning process is taken from the web, so many languages can be taken into account. BERT reads the input data as an entire sequence of words and this data is analysed as an entire block, so the model can learn about the context and surroundings. The input information corresponds to a sequence of tokens embedded into vectors, and, afterwards, it is processed in a neural network. Contrary to that, the output data is a sequence of vectors, where each one corresponds to the same index input token. BERT can be trained on the following strategies:

Masked language model (MLM) It consists of replacing some random words (close to 15% of each sequence) with a mask token. These sequences with masked parts are the input data in the process. The model tries to predict the mask value based on the context information provided in the rest of the sequence. Even if the MLM provides information about the sentence context, next sentences are not taken into account.

Next sentence prediction (NSP) The BERT model receives sentence pair as input and it learns to predict if the second sentence is related to the first one. Approximately, half of the training inputs have related sentences and the other half have a randomly chosen sentence. The model will learn to know if the following sentence of a particular input has a useful context for improving predictions' accuracy. There is a sentence embedding to indicate both sentences and, finally, a positional embedding is added to indicate the word position in the sequence.

3. Database generation for Spanish Abusive Short Texts

In this section we are briefly describing the original database as well as the preparation that we have done for the current study.

3.1. Description

The database used to train and test the system is the "Spanish Database for Cyberbullying Prevention" (Moreno et al., 2018). The database was collected and annotated jointly a Catalan database with the same specifications. Each database consists of 140,000 posts. Posts were selected from sources such as twitter, teenagers' chats, blogs, forums, medical consultation web sites, etc. Prior to annotation, selection of appropriated sources was carried out to verify the data was mainly coming from children and young teenagers between 7 and 14 years old. 100,000 post were selected from non-supervised sources that could not be censured. Other 40.000 posts were selected from sources either supervised or with non-expected abusive language (sports associations, school chats, . . .). Data were either manually or automatically downloaded, cleaned, formatted and selected. Posts were chosen to have a minimum number of characters (without counting –not discarding-@names and internet addresses) of 50 and a total maximum of 280. Spanish as spoken in Latin America posts were discarded when possible. Posts were labeled according to their content in 7 categories: aggression, anxiety, depression, distress, sexuality, use of substances, and violence. For each category, 5 levels of concern have been established. 1: no concern post, 5: extremely concern post. Annotation includes for each post, if a User, Perpetrator, Victim or Witness wrote it. Annotation was carried out by 16 bilingual (Spanish and Catalan) people, 4 males and 13 females. Each post, out of the 100.000 possible abusive posts per language, was annotated by two raters. Assignment was done in blocks of 1000 posts and at the end of the project, each rater had worked with all other raters in at least one block. The remaining 40.000 posts per language were rated once, just to verify that this subset was a 'non-concern' set. At the beginning of the project, all annotators took one training week. Each week, a follow-up procedure open to all the annotators was established with the following information: number of posts labelled during the week, graphical information about their dissimilarity in the use of labels and levels, and inter-rate agreement indexes for each pair of annotators. The following indexes were calculated to show the quality of the annotation on a binarized database (concern/no-concern) per category: Accuracy, Cohen's kappa, Cronbach‘s alpha and Pearson index. This information was very helpful to prepare monthly re-training sessions. Main annotation difficulties were to discriminate between violence and aggression, as well as between distress, depression and anxiety.

Table 1 shows the gender (code F/M), background, and the mean Cohen's kappa and Cronbach‘s alpha indexes of each annotator. Due to his low kappa index, rater M03 participated in the annotation of the 'non-concern' set during the third part of the project.

Code	Background	Cohen κ	Cronbach alpha
F01	Philosophy student	0.73	0.72
F02	Teacher degree	0.65	0.72
F03	Psychology student	0.78	0.80
F04	Criminology student	0.74	0.75
F05	Biomedicine student	0.70	0.71
F06	Psychology student	0.72	0.66
F07	Social Communication student	0.67	0.71
F08	Psychologist degree	0.67	0.72
F09	Biology and Neuroscience student	0.76	0.76
F10	Statistics and economy student	0.72	0.64
F11	Children's education	0.66	0.68
F12	Psychology student	0.72	0.68
F13	Engineering student		
M01	Engineering student	0.73	0.69
M02	Psychology student	0.66	0.70
M03	Physical activity and sports degree	0.40	0.64
M04	Engineering student	0.74	0.71

Table 1: Annotator's code, background and inter-rate agreement indexes. (F13 annotated non-concern post only)

Table 2 shows a few Spanish examples, with the annotated sentiment and also with the English translation (only for illustration).

3.2. Preparation

Even if texts on the database are available in Spanish and Catalan, current study only takes into account Spanish texts. So, the first step is to dismiss the information in other languages as well as additional cleaning (e.g. sentences with only emoticons, with only one word...), but reserving them for future implementations. Concretely, the Spanish database contains 111700 samples.

The labels are tagged in a range from "1" to "5", where "1" is no abuse and "5" is high quantity of abuse. However, to simplify the final task, each label value is facilitated with a binary range: "0" label corresponds to the "1" and "2" values in the "1" to "5" scale, and the "1" on the binary abusive label represents values from "3" to "5". Furthermore, the number of labels (which initially was 7) was reduced to 4 labels which are the following:

1. **Aggresion and violence (agg-vio):** 10989 samples (9.83%)
2. **Distress, anxiety and depression (dis-anx-dep):** 14452 samples (12.94%)
3. **Sexuality (sex)** (violations, period or reproduction)**:** 3315 samples (2.97%)
4. **Substances (sub)** (drugs, alcohol or tobacco)**:** 2847 samples (2.55%)

As shown in the percentages above, some of the labels are under 5% of occurrence. For that reason, we have considered one further simplification which is a binary classification where the evaluation is the difference between containing abusive behaviour or not. If one sentence contains an abusive value in any label, the "abusive label" is marked as 1, and it is identified with 0 otherwise. In this case, there is a 24,77% of the sentences with abusive behaviour.

Finally, we splitted the database into training, validation and test sets. The data division percentage is 70% for training, 15% for validation and 15% for test. The final percentage of "abusive label" for each data set is shown in Table 3.

4. Experiments

In this section, we detail the experimentation (parameters and implementation), main results and discussion about them.

4.1. Parameters and Implementation

For SVM and RF implementation, we used the Scikit-learn toolkit[1]. We used a Linear SVC method for SVM. We used 10 trees with a maximum depth of 2 for RF. For the BERT model, we used the TensorFlow implementation available from github[2]. In particular, we use the BERT-Base and multilingual cased. We used a batch size of 32, a maximum sequence length of 128 and trained for 1 epoch.

[1] https://scikit-learn.org/

[2] https://github.com/google-research/bert

Spanish Example	Sentiment							English Translation
	Agg	Anx	Dep	Dis	Sex	Sub	Vio	
No duermo, me obligo a comer, tengo siempre un nudo en el estómago, pienso que mi vida es una mierda, no soy feliz y me siento más sola que nunca.	1	4	5	5	1	1	1	I don't sleep, I force myself to eat, I always have a knot in my stomach, I think my life is crap, I'm not happy and I feel more alone than ever.
Te puedo decir q desde hace ocho meses q lo conozco él consume los sabados cocaína pero a esto hace unas semanas se ha añadido heroína	1	1	1	1	1	5	1	I can tell you that for eight months I know him he consumes cocaine on Saturdays, but to this, heroin has been added a few weeks ago
Te voy a partir la cara maldita hija de perra.	5	1	1	1	1	1	4	I'm going to break your damn bitch daughter's face.
Mi novio me obliga a ver demasiado porno y no me gusta.	1	1	1	1	4	1	3	My boyfriend forces me to watch too much porn and I don't like it.

Table 2: Database examples with annotated sentiments: aggression (agr), anxiety (anx), depression (dep), distress (dis), sexuality (sex) and substances (sub)) We provide the English translation just for illustration.

	Training	Val	Test
Agg-vio	9.81	10.02	9.81
Anx-dep-dis	13.05	12.53	12.95
Sex	2.96	3.13	2.91
Sub	2.52	2.60	2.63

Table 3: Percentage of positive label occurrence for multilabel classification

Sys	agg-vio	dis-anx-dep	BC
MV	**90.19**	87.05	75.23
SVM	85.60	*88.78*	*79.96*
RF	86.25	83.82	68.21
BERT	90.00	***89.11***	***80.73***

Table 4: Percentage of F1-score for the classifications agg-vio and dis-anx-dep, and the binary classification (bc, abusive text or not). In bold best results, in italics results above the majority vote (MV).

Sys	agg-vio	dis-anx-dep	BC
MV	17.18	22.93	39.71
SVM	15.42	***42.58***	***53.28***
RF	5.42	5.67	16.10
BERT	***32.38***	*32.78*	*48.69*

Table 5: Percentage of $F1_{tp,fp,fn}$ for the classifications: agg-vio and dis-anx-dep, and the binary classification (bc, abusive text or not). In bold best results, in italics results above the majority vote (MV).

4.2. Evaluation and Results

Table 4 shows results for the binary classification (BC) of "abuse" or not "abuse" statement as well as the classification between two of the categories described in section 3.2., which consists on providing a binary classification: either the specific class (e.g. agg-vio, anx-dep-dis) is shown in the short text or not. We do not include "sex" or "substance" because their representation was too low (under 5%); this unbalanced classifications are left for further research.

Note that we are facing a highly imbalanced dataset. This implies that a naive method that simply voted for the majority vote class (MV), which is not having any type of abuse, would already report high results. This naive method is also reported in the results in Table 4 where a 75% F1 score is achieved for the binary classification (BC) and over 90% for agg-vio. Therefore, we use the F1 score measure, both computed as simply the harmonic mean of precision and recall ($F1$) (Table 4) or as suggested in (Forman and Scholz, 2010), using true positives (TP) and false negatives (FN) and false positives (FP) as follows: $F1_{tp,fp,fn} = (2 * TP)/(2 * TP + FP + FN)$ which is the most adequate way to compare imbalanced datasets (Table 5).

From Table 4 we can see that not all the proposed methods are able to beat the majority vote technique, except for BERT, which is always better and only slightly worse in the case of aggression-violence. In this case we observe that for all classifications, the BERT model achieves higher accuracy than any of the other models (MV, SVM or RF). From Table 5 conclusions vary a little bit and we see that SVM and BERT have best performances depending on the classification.

4.3. Discussion

One of the most important objectives of this study was to test several classifiers, including latest approaches based on pre-trained powerful models like BERT, in a challenging unbalanced database of short texts. Given the nature of the task, detecting abusive "comments", we are highly concerned on the number of False Negatives (FN) of our binary classification, because these are the cases were "abusive" comments are missed by the system. Additionally, we show the number of False Positives (FP) ("non-abusive" comments are detected as "abusive"), which much less alarming, may also lead to unnecessary attention. These are shown in Table 6.

5. Conclusions and Future Work

This paper reports an experimental research on short text classification for abusive language in Spanish.

System	FP	FN
SVM	1122	2236
RF	2983	3288
BERT	970	2503

Table 6: False Positives (FP) and False Negatives (FN) for the Binary Classification of Tables 4 and 5.

The main contribution of this paper is constrating the performance of ”classic” classifiers like SVM or RF with latest pre-trained models, BERT in a challenging unbalanced dataset of short texts.

Our best results are either obtained when using BERT model in most of the cases, while SVMs still were competitive in few cases. One of the main motivations of using BERT is that we are interested in extending our work to classify comments in several languages at the same time, and the multilingual version of BERT allows specifically for this option. Therefore, we can extend our initial current work on different directions. We have to research on how to classify extremely unbalanced categories like ”sex” or ”substance”. We can further exploit our data using both Spanish and Catalan as sources of information in the multilingual version of BERT. In a different direction, and concerned about gender fairness in NLP (Costa-jussà, 2019), we can use the evaluator’s name to provide information about its gender, we can study if men and women have different opinions regarding abuse sentiments.

Acknowledgments

This work is also supported in part by the Spanish Ministerio de Economía y Competitividad, the European Regional Development Fund through the postdoctoral senior grant Ramón y Cajal.

References

Breiman, L. (2001). Random forests. *Mach. Learn.*, 45(1):5–32, October.

Chen, H., McKeever, S., and Delany, S. J. (2017). Abusive text detection using neural networks. In *AICS*.

Costa-jussà, M. R. (2019). An analysis of gender bias studies in natural language processing. *Nature Machine Intelligence*, 1(11).

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Forman, G. and Scholz, M. (2010). Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement. *SIGKDD Explor. Newsl.*, 12(1):49–57, November.

Hearst, M. A. (1998). Support vector machines. *IEEE Intelligent Systems*, 13(4):18–28, July.

Mercado, R. N. M., Faustino, H., and Gutierrez, E. G. C. (2018). Automatic cyberbullying detection in spanish-language social networks using sentiment analysis techniques.

Moreno, A., Bonafonte, A., Jauk, I., Tarrés, L., and Pereira, V. (2018). Corpus for cyberbullying prevention. In *Proc. IberSPEECH*, pages 170–171.

Saba, N. (2018). The rise of bullying as a public health issue. *Law School Student Scholarship. 945.*

Salawu, S., He, Y., and Lumsden, J. (2017). Approaches to automated detection of cyberbullying: A survey. early online, 10. Copyright 2017 IEEE - All rights reserved.

Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G., Daelemans, W., and Hoste, V. (2018). Automatic detection of cyberbullying in social media text. *PLoS One. 2018;13(10):e0203794. Published 2018 Oct 8. doi:10.1371/journal.pone.0203794*, 10.

Zhu, J., Tian, Z., and Kübler, S. (2019). Um-iu@ling at semeval-2019 task 6: Identifying offensive tweets using bert and svms. In *SemEval*, 04.