Topic detection in Spanish and Catalan short texts: exploring embeddings and classification techniques in a low-resource setting

Anonymous COLING submission

Abstract

This paper explores the right pipeline to accurately perform text classification in a very specific task. The task consists on classifying short texts with very informal language coming from teenagers. Classification categories are abusive topics including aggression or violence, disorders such as anxiety, depression or distress, use of substances such as alcohol, drugs or tobacco, sexuality, and none of the previous. While text classification of short sequences has been widely studied for English, in this paper we are exploring the task in Spanish and Catalan using a small and unbalanced dataset. We test different possible pipelines exploring preprocessing and classification techniques. Best results are achieved using our proposed preprocessing without stemming and the multilingual version of BERT with the default model for sequence classification for most of the tasks, except for one where using BERT plus Support Vector Machines slightly improves results.

1 Introduction

The purpose of this paper is to design a preprocessing and modelling pipeline based on selecting most adequate state-of-the-art techniques to achieve a high performance in a specific text classification task. The task consists on automatically detecting the topic in short posts and tweets from children and teenager's conversations written in Spanish and Catalan. Classification categories are abusive topics including aggression-violence, anxiety-depression-distress, sexuality, substance and neutral (none of the previous). This task is trained on a database created in the framework of the project "Safeguarding children online" (Moreno et al., 2018). The database has the following annotated sentiments: aggression, anxiety, depression, distress, substances, sexuality and violence. For similarity reasons as well as low representation of certain classes, we are grouping together aggression and violence as well as anxiety, depression and distress.

As modelling techniques we consider several machine learning classifications techniques: Support Vector Machines (SVMs) (Hearst, 1998) and Neural Networks classifiers. Beyond this, we explore several strategies for word and sentence vectorization: tf-idf (Sammut and Webb, 2010), doc2vec (Le and Mikolov, 2014) and BERT (Devlin et al., 2018a). Since our evaluation dataset is highly unbalanced, we evaluate the results using a modified f1 score as suggested in (Forman and Scholz, 2010), which is proven as an adequate way to compare unbalanced datasets. Best results are achieved with the BERT default model for sequence classification, obtaining 74 % accuracy in a binary classification with any of the sentiments or neutral, improving the majority voting by 13 %.

The rest of the paper is organised as follows. Section 2 describes in detail the proposed task and related work. Section 3 describes the general algorithms used in this work (embeddings and classifiers) as well as the final methodology to implement our system. Section 4 describes the experiments including data, preprocessing and model parameters details. Finally, section 5 reports results and conclusions.

2 Proposed task and related work

Our task consists on classifying short texts into different categories including aggression and violence, anxiety, depression and distress, substance or neutral. For this task we have a supervised dataset of approximately 220,000 statements both in Catalan and Spanish, which can be considered low-resource. The same dataset is detailed in (Moreno et al., 2018) and it has been previously used in a sentiment analysis work (Costa-jussà et al., 2020). In particular, the work was using the same database, but with information about the degree of the sentiment, and focusing on detecting statements that reflected abuse. The main difference in our case is that we are disregarding the degree of the sentiment or abuse and only targeting at labelling the statement by topic.

Our task falls in the field of supervised topic modeling or text classification. While both topic modeling and text classification has been largely addressed from an unsupervised point of view, there is also quite a few amount of works from the supervised point of view, including statistical (Blei and McAuliffe, 2007; Korshunova et al., 2019) and machine-learning approaches (Alsmadi and Gan, 2019; Huang et al., 2018). While the former are effective techniques, the latter can better deal with complex data sets and do not require restraining assumptions like linearity and normality (Kumar and Bhattacharya, 2006). Moreover, machine learning allows to further scale to include large data sets or pretrained models (Devlin et al., 2018b), which can hugely benefit in low-resource frameworks. In this work, we are exploring several representative techniques within the machine learning category to know which is the best option for our particular case. We explore the right pipeline for preprocessing, embedding and classification. For preprocessing, we propose standard procedures like tokenization, deleting punctuation and stopwords, lowercasing and stemming combined with filtering of vocabulary by frequency of appearance. In the embedding case, we are exploring either standard word embeddings and their generalization to sentence embeddings (doc2vec) or latest proposals on contextual word embeddings (Devlin et al., 2018b). These embeddings are compared to the results of the tf-idf. Finally, we compare the performance of classifiers using Support Vector Machines (SVMs) or Neural Networks. The entire procedure is summarized in figure 1.

3 Background and proposed methodology

This section describes the main existing techniques explored in the context of our task, including different vectorization forms (tf-idf and embeddings) as well as two different classifiers SVMs and fine-tuning on BERT with a fully-connected network (softmax).

TF-IDF Term Frequency - Inverse Document Frequency is a vectorizing method often used in text mining and information retrieval. It is a statistical measure used to evaluate how important a word is to a corpus, and it basically transforms text to feature vectors that can be used as input to an estimator. The importance increases proportionally to the number of times a word appears in the document, but is inversely proportional to the frequency of the word in the corpus (Sammut and Webb, 2010).

Doc2Vec Doc2Vec is aimed to create a numeric representation of a text document, regardless of its length, that can be used for many purposes, such as document retrieval, web search, topic modeling or spam filtering. This concept was presented by Le and Mikolov (Le and Mikolov, 2014) as an extension to Word2Vec (Mikolov et al., 2013).

BERT Bidirectional Encoder Representations from Transformers model (Devlin et al., 2018a) is a model that applies the bidirectional training of Transformer (an attention mechanism that learns contextual relations between words in a text) to language modelling. As opposed to directional models (which read the input sequentially), the encoder reads the entire sequence at once without an specific direction, which allows the model to learn the context of a word based on its entire surroundings. As input it accepts a sequence of tokens, which are first embedded into vectors and then processed in the neural network, and as output it gives a sequence of vectors corresponding to the input tokens. BERT uses two training strategies: masked language models and next sentence prediction. Moreover, BERT has available a multilingual training which is trained with 104 languages.

Classifiers Machine learning approaches offers a large variety of classifiers. For their success in short text classification e.g. (Yin et al., 2015), we explore and contrast, in our particular task, Support Vector Machines (SVMs) and Neural-based with BERT (Ma, 2019). SVM classification algorithm (Hearst,

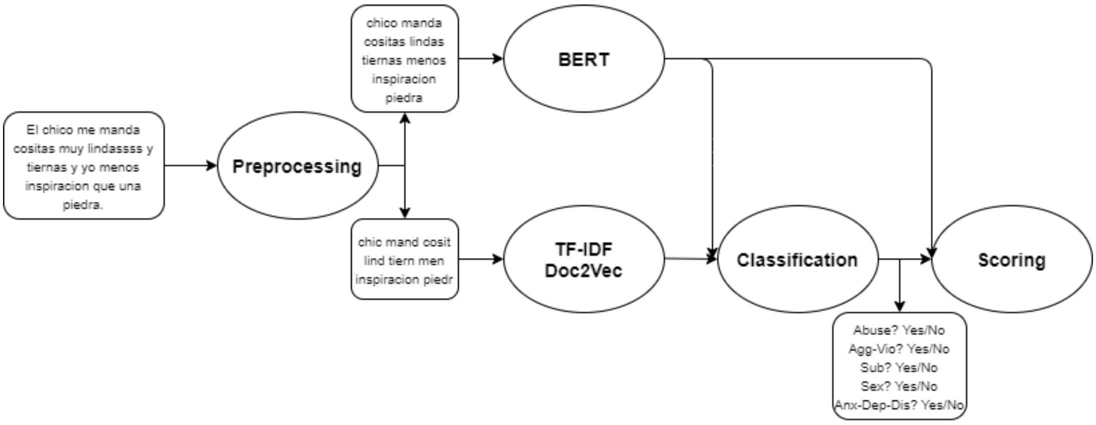
chico manda cositas lindas tiernas menos inspiracion piedra
El chico me manda cositas muy lindassss y tiernas y yo menos inspiracion que una piedra.
Preprocessing
BERT
chic mand cosit lind tiern men inspiracion piedr
TF-IDF Doc2Vec
Classification
Scoring
Abuse? Yes/No
Agg-Vio? Yes/No
Sub? Yes/No
Sex? Yes/No
Anx-Dep-Dis? Yes/No

Figure 1: Block Diagram

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. An alternative to SVM is using the final hidden state of the first word ([CLS]) from BERT as input to a fully connected layer to perform softmax. This is a fine-tuning approach.

4 Experiments

Data and Evaluation The dataset (Moreno et al., 2018) includes classification for short texts in Catalan and Spanish for 7 different categories including: aggression, violence, anxiety, depression, distress, sex and substance. These different categories are in turn tagged in 1-to-5 degrees, where 1 is neutral (no concern) and 5 is high quantity of concern. In this work, we simplify the classification in two directions. First, we group aggression and violence in one category (violence) and disorders such as anxiety, depression and distress also in one category (disorders). Second, we do not consider degrees and use the classification of neutral when there is a 1 degree, and the specific category when degree ranges from 2-to-5. Note that one short text can be classified within several categories. The number of posts in each category is summarized in Table 1. We randomly extract our evaluation test set (20% of the total). Since our dataset is highly unbalanced, we evaluate the results using a modified f1 score as suggested in (Forman and Scholz, 2010), using true positives (TP), false negatives (FN) and false positives (FP) as follows: $F1_{tp,fp,fn} = (2 * TP)/(2 * TP + FP + FN)$.

Catalan	Spanish	violence	substance	sex	disorders	neutral
109141	111719	42227	6209	29621	43615	121747

Table 1: Total number of posts by language and category.

Preprocessing For this kind of tasks, a solid preprocessing is key in order to properly prepare the data for training. The difference between training a classifier using raw data and training it using processed data is astonishing. Preprocessing gets rid of the majority of unimportant words, leaving only the words that contain the highest information either in terms of what they mean or in terms of the context they offer. The stemming step was avoided in all the experiments involving BERT. Table 2 shows a complete example of our preprocessing. We also did a manual tagging of the less frequent words in the doc2vec experiment in order to try and get the system to focus even more on the more important parts. To do this, we used an extensive list of key abusive words in order to avoid deleting those words which can be infrequent but are relevant.

Action	Sentence
Original	El chico me manda cositas muy lindassss y tiernas y yo menos inspiracion que una piedra.
Tok.	El chico me manda cositas muy lindassss y tiernas y yo menos inspiracion que una piedra .
No-punct	El chico me manda cositas muy lindassss y tiernas y yo menos inspiracion que una piedra
Lowercase	el chico me manda cositas muy lindassss y tiernas y yo menos inspiracion que una piedra
Normaliz.	el chico me manda cositas muy lindas y tiernas y yo menos inspiracion que una piedra
No-stopwrd	chico manda cositas lindas tiernas menos inspiracion piedra
Stemming	chic mand cosit lind tiern men inspiracion piedr

Table 2: Preprocessing example.

Implementation and Parameters We used the sklearn toolkit[1] and in particular the sklearn TfidfVectorizer() function with the default parameters to extract the tf-idf. For the SVM classification task, we used the LinearSVC() function. For doc2vec embeddings, we used the gensim doc2vec library [2]. The parameters used include a vector size of 200 dimensions, use of distributed bag of words and we trained for 70 epochs. For BERT, we have used the BERT-Base, Multilingual Uncased model that can be found in their github page[3], with 102 languages, 12-layer, 768-hidden, 12-head and 110M parameters, which is the default configuration. In the case of implementing BERT-SVM, we have used our own fine-tuned models for extracting the BERT embeddings using the huggingface sentence_transformers library[4].

5 Results and Discussion

Table 3 shows results doing binary classifications to detect between each category in its own or neutral and finally to detect any of the categories (general abuse) or neutral. Our systems are compared with the Majority Voting (MV) baseline system.

Experiment	violence	substance	sex	disorders	general
MV	32.23	5.16	23.49	33.19	61.95
TFIDF-SVM	19.64	3.46	8.52	46.00	70.91
doc2vec-SVM	37.31	50.40	37.32	31.36	68.86
BERT	**65.20**	73.45	**66.02**	**50.74**	**74.93**
BERT-SVM	64.38	**73.83**	64.93	47.20	74.82

Table 3: Classification accuracy.

Discussion. Improvements vary depending on the techniques used, and some of them are consistent in all classification tasks, while others are not. Using doc2vec as opposed to tf-idf as preprocessing for SVM improves on 3 out of 5 tasks. On the other hand, using BERT increases consistently in all tasks compared to using tf-idf or doc2vec with SVMs. Finally, globally, it is better to use default BERT classifier token with a fully connected layer to perform softmax than using this token as input to a SVM classifier. We have tested SVMs as opposed to other fancy Neural Networks classifiers such as non-linear layers, recurrent or convolutional, because these ones were proven without success in previous works (Ma, 2019). Given the task and motivated by the social implications of succeeding in it (e.g. detecting abuse in teenagers environments), we are planning to continue our research work in several directions. We will consider using contextual topic identification taking advantage of combined topic modeling techniques such as Latent Dirichlet Allocation as well as making use of extra data crawled from specific websites.

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

[2] https://radimrehurek.com/gensim/models/doc2vec.html

[3] https://github.com/google-research/BERT/blob/master/multilingual.md

[4] https://github.com/huggingface/transformers

References

Issa Alsmadi and Keng Hoon Gan. 2019. Review of short-text classification. *International Journal of Web Information Systems*, 01.

David M. Blei and Jon D. McAuliffe. 2007. Supervised topic models. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, page 121–128, Red Hook, NY, USA. Curran Associates Inc.

Marta R. Costa-jussà, Esther González, Asuncion Moreno, and Eudald Cumalat. 2020. Abusive language in spanish children and young teenager's conversations: data preparation and short text classification with contextual word embeddings. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1533–1537, Marseille, France, May. European Language Resources Association.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018a. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018b. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

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

Minghui Huang, Yanghui Rao, Yuwei Liu, Haoran Xie, and Fu Lee Wang. 2018. Siamese network-based supervised topic modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4652–4662, Brussels, Belgium, October-November. Association for Computational Linguistics.

I. Korshunova, H. Xiong, M. Fedoryszak, and L. Theis. 2019. Discriminative topic modeling with logistic lda. In *Advances in Neural Information Processing Systems 33*.

Kuldeep Kumar and Sukanto Bhattacharya. 2006. Artificial neural network vs linear discriminant analysis in credit ratings forecast: A comparative study of prediction performances. *Review of Accounting and Finance*, 5:216–227, 07.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, page II–1188–II–1196. JMLR.org.

Guoqin Ma. 2019. Tweets classification with bert in the field of disaster management. In *Stanford Report*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

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

Claude Sammut and Geoffrey I. Webb, editors, 2010. *TF–IDF*, pages 986–987. Springer US, Boston, MA.

C. Yin, J. Xiang, H. Zhang, J. Wang, Z. Yin, and J. Kim. 2015. A new svm method for short text classification based on semi-supervised learning. In *2015 4th International Conference on Advanced Information Technology and Sensor Application (AITS)*, pages 100–103.