BERT's Auxiliary Sentence focused on Word's Information for Offensiveness Detection

Fernando Sánchez-Vega and Adrián Pastor López-Monroy

Mathematics Research Center (CIMAT),
Jalisco s/n Valenciana, 36023, Guanajuato
{fernando.sanchez, pastor.lopez}@cimat.mx

Abstract. This paper describes the participation of the CIMAT-GTO team in the MeOffendEs 2021 competition. Our main goal is to evaluate an auxiliary sentence scheme for classification with BERT in the offensiveness identification task. The use of the auxiliary sentence has been shown to increase the efficiency of classifiers based on pre-trained BERT models in various tasks. We propose two new approaches to obtain the auxiliary sentence, the objective of the proposals is to remark the available information on the use of the words along the classes in the training corpus. The proposals S2KNNC and S2ChiN use techniques related to Nearest Neighbor and Attribute Selection by Chi-square, respectively, to construct the auxiliary sentence. Our results indicate that the auxiliary sentence scheme allows to improve the performance of the BERT-based classifier or even BERT classifier ensembles.

Keywords: Auxiliary sentence · BERT Ensembles · Offensiveness identification.

1 Introduction

The violence on social media is clearly manifested in widespread polarization [1] and it has direct repercussions that can manifest in cyber-bullying or even lead to suicide [2].

To promote the fight against this problem, Plaza-del-Arco et al [3] have set the task of identifying the offensive text in tweets within the framework of the MeOffendES 2021 competition at [4]. This competition has 4 sub-tasks:

1. Classification of aggressive text of generic Spanish into four classes
2. Classification of aggressive text of generic Spanish into four classes including additional contextual information
3. Identification of the aggressive text in Spanish of the Mexican variant
4. Identification of the aggressive text in Spanish of the Mexican variant including contextual information

IberLEF 2021, September 2021, Málaga, Spain.

The main goal of this paper is to evaluate an auxiliary sentence scheme for classification with BERT in the offensiveness identification task. The recent use of the auxiliary sentence in classifiers based on the fine tuning of pre-trained BERT models has shown increased efficiency in multiple tasks. We propose two new approaches to obtain the auxiliary sentence, these proposals seek to take advantage of the information available in the training corpus on the word´s uses. The S2KNNC approach takes back the training instances where the same words are used in close contexts to be compared and considered as auxiliary sentence for the classification model. The S2ChiN presents an instance version with refined information to the classifier as the auxiliary sentence in order to focus the attention on the most relevant words. Our results indicate that the auxiliary sentence scheme allows to improve the performance of the BERT-based classifier or even BERT classifier ensembles.

The remainder of the paper is structured as follows. Section 2 presents a brief description of the state of the art, it includes the aggressive text identification proposals that are closest to our approach and a brief description of the works that have recently been proposed to use the auxiliary sentence scheme in other tasks. Section 3 describes the three auxiliary sentence schemes explored. In Section 4 the experimental results are presented. In Section 5 some ethical issues that concern this work are discussed and Section 6 presents our main conclusions.

2 Related Work

In the same way as other NLP tasks, the initial approaches to the identification of offensiveness have proposed the use of classical schemes such as BoW with TFIDF [5], n-grams of words [6], n-grams of characters [7], however, the new deep neural network techniques have pushed the results to a new level of efficiency, in works such as those used CNN [8] or GRU [9] networks. Past editions of the aggressiveness identification competition [10] have clearly shown the predominance of transformer-based networks [11], the four approaches with the best results use some form of a transformer-based network.

In [12] an ensemble of BETO[1] classifiers with fine-tuning and the use of data augmentation is proposed, in [13] the use of automatic translation is proposed in order to be able to alternate the use of pre-trained transformers in Spanish, English or multilingual. The work of Villatoro-Tello et al. [14] proposes a classifier whose input is the output probability of the BETO network in addition to attributes obtained by a BoW and some metadata attributes that are specific to the nature of the social network used in the competition.

The results obtained by the different approaches show that the approaches based on BETO classifiers are the most effective, however, it has been seen some improvement by small assistance incorporating external information in the

[1] We note that BETO is a model very similar to BERT, pre-trained in Spanish [18] and made available by the authors at https://github.com/dccuchile/beto

training through data augmentation with external corpus or with modified instances versions of the same training corpus as adversary augmentation versions or augmentation with related words replaced [12].

In several text classification tasks that have been shown to obtain good results using transformers, such as sentiment analysis, topic classification, question classification and humor prediction, a new scheme has been proposed to help the transformer network by providing additional information. This approach focuses the BERT classification model to the new task, taking advantage of the fact that BERT has a 2-sentence input scheme. The second sentence, normally not used in text classification, introduces additional or auxiliary information [15–17].

2.1 Auxiliary sentence schemes

The idea behind the BERT auxiliary sentence schemes is to exploit the entire structure used in the BERT pre-training where two sentences are included. BERT simultaneously learns to predict the masked tokens and predict the relationship between the two sentences to infer whether they are consecutive or not (in the corpus in which the BERT language model is trained).

The two-sentence scheme was initially used for tasks where the classification problem requires two text sequences such as question answering (Q.A.) or semantic relationship between texts (STS), however, recently this additional entry is being used to give additional information or to guide the BERT model in single sentence classification tasks. In these schemes, it is important to define how the auxiliary sentence will be constructed in order to correctly orient the BERT model in the new task.

In [16] the auxiliary sentence scheme is used to perform sentiment analysis focused on specific aspects. The auxiliary sentences proposed are:

1. The use of the auxiliary sentence in the form of a question to guide BERT to focus on the specific aspect or classification's characteristic; the auxiliary sentence has the form: "what do you think of the <aspect>of <object of evaluation>" (e.g. "what do you think of the safety of location-1 ?").
2. The use of the auxiliary sentence in the form of a complete answer, including the class; the auxiliary sentence has the form: "the polarity of the aspect <aspect>of <object of evaluation>is <class>" (e.g. "the polarity of the aspect safety of location-1 is positive").
3. The use of an auxiliary sentence that only contains the keywords that BERT is required to focus on, in this case the sentence is not necessarily grammatically correct so it is called a pseudo-sentence; the auxiliary sentence has the form: "<aspect><object of evaluation><class>" (e.g. "safety location-1 positive").

In [15] in the spirit of introducing some generalization, an auxiliary sentence is constructed by means of data augmentation techniques in which certain words of the original sentence are replaced.

The auxiliary sentence scheme has allowed to provide additional task information to the pre-trained transformer networks (BERT, RoBERTa or XLNet), in order to improve efficiency in tasks where the fine-tuning is performed.

3 Proposed method

We propose the use of the auxiliary sentence scheme under the successful BETO fine-tuning classifier ensemble [12]. We explore three construction methods of the auxiliary sentence, one taken directly from the proposed schemes in the state of the art and two new approaches. In the following subsections, we describe the three methods to obtain the auxiliary sentences and the proposed BETO ensemble.

3.1 Auxiliary sentence baseline scheme: Basic Question (S2Q)

This strategy of auxiliary sentence construction was the direct adaptation of the characteristic question of the classification interest as it is proposed in [16]. The auxiliary sentence for the offensive language detection identification case is: "Es ofensivo?"; in English: "Is this offensive?".

3.2 Auxiliary sentence scheme based on Nearest Neighbor (S2KNNC)

Following a similar motivation to the use of auxiliary sentence with augmented instances as in [17], we propose a new scheme that provides to BETO classifier an auxiliary sentence that helps to compare the sentence to classify (S1) with other sentences in which several S1's words are used in similar context and that it is confident that their use and context make them offensive or non-offensive.

To obtain the auxiliary sentence, S2, for a sentence to be classified, S1, by the method based on the nearest neighbor (S2KNNC), we follow:

Given an input to the classifier, sentence S1, we obtain its auxiliary sentence S2 getting the k-th nearest neighbor sentence with class C (Positive or Negative) in the training corpus. Finally the input for the BETO fine-tuning in S2KNNC method is the pair S1, S2 .

The auxiliary sentence is chosen with a specific class (known because the sentence comes from the training corpus) and a specific neighbor number K. The selection of K and class C provides us with different alternatives for the auxiliary sentence S2 for the same input instance S1 as we shown in examples of the Table 1.

To obtain the closest neighbor we use a BoW representation with a TFIDF weighting scheme, removing stopwords and using the Euclidean distance. This

strategy allow us to obtain auxiliary sentences where the same relevant class words are used[2].

The intuitive idea of using the nearest neighbor-based auxiliary sentence is to assist the classifier in learning the task by offering a very close point of comparison where words are used in a particular class and take advantage of the S1 and S2 sentence relationship pre-training to the new task.

3.3 Auxiliary sentence scheme based on Relevant Information (S2ChiN)

In our second proposal for the use of the auxiliary sentence scheme, we construct a pseudo sentence from the most relevant words for the classification of the original sentence S1. The S2 auxiliary sentence is a version that filters out the least relevant words to focus the classifier's attention on the most informative words for class prediction purposes.

To obtain the auxiliary sentence, S2, for a sentence to be classified, S1, by the method based on Chi-squared information filtering (S2ChiN), we follow:

Given an instance of an input sentence to the classifier, sentence S1, we construct its auxiliary sentence S2 after eliminating all the words that are not in the N top of the ranking of the best attributes using chi-square test .

The auxiliary sentence constructed depends directly on the choice of the parameter N that is correlated with the amount of information and words that are preserved, therefore we can construct different alternatives of auxiliary sentence S2 for the same input instance S1 by choosing different thresholds value N.

The idea behind the use of the relevant information in to the auxiliary sentence is to help the classifier to focus on the words that provide (statistically) the highest confidence of the class. We expect to provide the neural network with a clue of the relevant attributes from the beginning and guide the network weight adjustment process to converge to an optimal point or at least a semi-optimal point better than when this additional information is not taken into account.

3.4 BETO classifier ensemble method

Taking into account the good performance of BETO ensembles [12], for each strategy used in the auxiliary sentence scheme, an ensemble of classifiers is generated by using fine-tuning pre-trained BETO model [18]. In our ensemble, the output probability of each fine-tuning BETO model is used as attributes input to a SVM[3] classifier.

[2] It should be noted that experimentally these characteristics achieve the best performance for classification with the Nearest Neighbors method.

[3] The linear kernel SVM from the Skit Learn library. This concatenation of methods is previously used in [9] and it obtains better experimental results than a voting weighing schemes.

The ensemble with the S2Q method is made up of 10 classifiers with randomly initialized of the last linear layer of the fine tuning BETO model, all classifiers are trained trained with the sames pairs S1, S2Q. The ensemble with the S2KNNC method integrates 18 classifiers, 3 for each parameter configuration (each one with linear layer random initialization) used in 6 parametric combinations with K = [1, 2, 3] and C = [Positive, Negative]. Finally in ensemble S2ChiN has 9 classifiers in the ensemble, 3 for each parameter configuration used with the parametric combinations with N = [400, 600, 800].

3.5 Auxiliary sentence examples

In this section we include an example of the auxiliary sentences generated for the same original sentence by the different methods. The Table 1 shows the example of the auxiliary sentences generated by the different alternatives of generation scheme for the same instance, we can see in S2KNNC with positive and negative classes that the auxiliary sentence give a good sample of the offensive and non-ofensive use of the word *"gorda"* (Fat), which is the word with a possible ambiguous use. In the auxiliary sentences generated by S2ChiN, we see that the word *"vista"* and the whole expression *"la vista gorda"* are selected, the word *"vista"* is the most important clue for know that the word *"gorda"* is part if an non-offensive expression *"hacerse de la vista gorda"*[4]. Therefore, S2ChiN gives the classifier an S2 with filtered key information to identify the non-offensiveness of the text.

Table 1. Examples of auxiliary sentence with S2KNNC method with two values of K and two classes and with S2ChiN method with two values of N.

Original sentence			
S1: *"Asi es, bien coludidos haciendose de la vista gorda"*			
Auxiliary sentences by method			
Method	Parameters		Auxiliary sentence (S2)
S2Q	None		*"Es ofensivo?"*
	kth NN	Class	
S2KNNC	1st	Positive	*"gorda pelotuda"*
S2KNNC	3rd	Positive	*"Tu madre es muy gorda"*
S2KNNC	1st	Negative	*"Últimamente me estoy haciendo de la vista gorda... Y de la cara, la cintura, los cachetes"*
S2KNNC	3rd	Negative	*"¡Eso es un secreto a voces! Solo las autoridades se hacen de la vista gorda..."*
	N threshold for chi-square		
S2ChiN	300		*"la vista"*
S2ChiN	1200		*"de la vista gorda"*

[4] Close to the meaning of "turn a blind eye" and nothing related to obesity

4 Results

In this section we describe the experiments designed to compare the proposed strategies. There are preliminary evaluations prior to those submitted in the MeOfendEs evaluation campaign [3] and the results of the proposals sent in the official submits.

4.1 Pre-competition results

We perform a stratified division on the MEXA3T training set [10] taking 72% of training, 8% of validation and 20% of test. The methods were evaluated with the three auxiliary sentence generation schemes as well as some evaluations of the individual components of the proposed methods and methods for comparison as the baseline of the ensemble without auxiliary sentence (Only S1), BETO-baseline. We evaluate the two ensemble strategies, the previously proposed weighted voting scheme and the use of linear SVM to weight the ensemble.

Table 2. Summary of pre-competition results.

Method	Parameters	Ensemble strategy	Size ensemble	F-Score
BETO-baseline	-	voted weighted	10	0.819
BETO-baseline	-	SVM	10	0.855
S2KNNC	K= 1; C= Positive	-	1	0.827
S2KNNC	K= 1; C= Negative	-	1	0.800
S2KNNC	K= 2; C= Positive	-	1	0.831
S2KNNC	K= 2; C= Negative	-	1	0.796
S2KNNC	K= 3; C= Positive	-	1	0.829
S2KNNC	K= 3; C= Negative	-	1	0.794
S2ChiN	N= 200	-	1	0.838
S2ChiN	N= 400	-	1	0.841
S2ChiN	N= 600	-	1	0.846
S2ChiN	N= 800	-	1	0.837
S2ChiN	N= 1000	-	1	0.853
S2Q	-	voted weighted	10	0.824
S2KNNC	K=[1,2,3]	voted weighted	18	0.833
S2ChiN	N=[400,600,800]	voted weighted	9	0.804
S2ChiN	N=[200,400,600,800,1000]	voted weighted	15	0.736
S2Q	-	SVM	10	0.856
S2KNNC	K=[1,2,3]	SVM	18	0.859
S2ChiN	N=[400,600,800]	SVM	9	0.862
S2ChiN	N=[200,400,600,800,1000]	SVM	15	0.856

In Table 2 we show the F-Score measure of the first harmonic (M-F1) obtained in the test set of our MEXA3T division. The experimental results show the different performances obtained by exploring specific parameters of each model and the size of the ensemble. From the Table 2 the relevance of the ensemble

strategy is observed, it is always better to use SVM. In general SVM allows to get results as good as the best individual component or better. Observing the use of a single auxiliary sentence (without ensemble) of the S2KNNC method, we find that the auxiliary sentence from the positive class always gets a better performance than the negative one, this phenomenon is probably due to the fact that the positive class is the minority class therefore is more difficult to learn for the classifier and the auxiliary sentences is a good help. From the components of S2KNNC it is surprising that the nearest neighbor (K= 1) is not the be the best component (only for negative class). In the comparison of the S2ChiN components, we see that apparently higher thresholds N allow better performance though, in the ensemble, better results were obtained in a more limited range of information filtering.

In general, it is observed that the addition of extra information included in the auxiliary sentence pushes up the results, however, not all information addition is equally good as evidenced by the fact that the best ensembles are not those with the greater number of components.

4.2 Official competition results

For the MeOffendEs competition all models were trained with the provided set for the competition and the predictions on the test set were submitted. In Table 3 we show the Precision, Recall and F-Score for the positive class reported by the organisers.

Table 3. Summary of official competition results.

Participant method	Precision	Recall	F1 score
CIMAT	0.760	0.653	**0.702**
NLP-CIC	0.755	0.640	0.693
DCCD-INFOTEC	0.673	**0.697**	0.685
CIMAT-GTO (S2ChiN)	**0.663**	**0.696**	**0.679**
UMUTeam	0.665	0.676	0.670
QU	0.743	0.606	0.668
CIMAT-GTO (S2KNNC)	**0.715**	**0.577**	**0.638**
Timen-autoBOT	0.600	0.608	0.604
CIC-DanHv	0.535	0.687	0.602
Dong	0.605	0.536	0.568
GDUFS$_D M$	0.888	0.342	0.493
Aomar	0.875	0.324	0.473
Sreelakshmi	**0.918**	0.314	0.468
Hugo.jair	0.270	0.270	0.270

From the Table 3 we can see the methods proposed by our CIMAT-GTO group, S2ChiN and S2KNNC. Those are in the fourth and seventh position in the rank of the F-score, the S2ChiN method obtains better performance than the

S2KNNC as in the preliminary results (Section 4.1). The S2ChiN method is 3.2% below the best performance "CIMAT" group. An interesting observation is that S2ChiN is the second method with the best Recall (with a value very close to the first place), an important factor for the offensive identification task because all suspected cases must be attended, as noted in the introduction, omissions could be costly.

5 Ethical issues

We find it necessary to note that this proposed system has not considered offensive expressions such as racism, sexism or other expressions that could offend or harm vulnerable groups in a more serious way. Similarly, it is necessary to note that the evaluation forum makes a distinction between offensive and vulgar expressions, therefore, groups that may be more sensitive to vulgarity may be offended by expressions not identified as offensive. Finally, we understand that language is a cultural expression and as such, it is always subject of interpretations that can always be different between different cultures. We warn that the criteria learned by the systems and used during the analysis should not be taken as a single reference, however, we believe that this reference is very useful for the tools development. Other reference criteria could be considered for the system if it is required in the future for a different context.

6 Conclusions

In this work, two new approaches were proposed for the application of auxiliary sentence schemes to the offensiveness identification. We found that the auxiliary sentences scheme help to BETO classifiers in the problem addressed, moreover, it was found that some ways of obtaining auxiliary sentences are better than others. The auxiliary sentence with the filtering of the relevant information were the best option but we found that the filtering parameter N must be selected in a narrow range to obtain a good balance between the relevant information concentration and the lost information. Finally, we discover that the auxiliary statement scheme is a good strategy to introduce additional information to the BETO classifier, however, better ways to build the auxiliary sentence need to be explored to achieve more significant improvements in the performance of the identification method.

Acknowledgments

The authors thank CONACYT, INAOE and CIMAT for the computer resources provided through the INAOE Supercomputing Laboratory's Deep Learning Platform for Language Technologies (*Laboratorio de Supercómputo: Plataforma de Aprendizaje Profundo*) with the project *"Identification of Aggressive and Offensive text through specialized BERT's ensembles"* and CIMAT Bajio Supercomputing Laboratory (#300832). Sanchez-Vega would like to thank CONACYT for

its support through projects *"Algoritmos de procesamiento del lenguaje natural para la modelación y análisis de la violencia textual con aplicación en documentos históricos"* (ID. BP-FP-20201015143044227-814705) and *"Ciencia de datos aplicado al análisis de expedientes de personas desaparecidas"*.

References

1. Danton Cetola: Why Social Media Makes Us More Polarized and How to Fix It, In: Scientific American. https://www.scientificamerican.com/article/why-social-media-makes-us-more-polarized-and-how-to-fix-it/ (October 15, 2015).
2. Luxton, David D.; June, Jennifer D.: Fairall, Jonathan M.: Social media and suicide: a public health perspective, In: American Journal of Public Health. 102 Suppl 2(Suppl 2):S195-S200. doi:10.2105/AJPH.2011.300608
3. Plaza-del-Arco, Flor Miriam, Casavantes, Marco and Escalante, Hugo Jair, Martín-Valdivia, M. Teresa, Montejo-Ráez, Arturo, Montes-y-Gómez, Manuel and Jarquín-Vásquez, Horacio, Villaseñor-Pineda, Luis: Overview of the MeOffendEs task on offensive text detection at IberLEF 2021, In: Procesamiento del Lenguaje Natural, V. 67, 2021.
4. Montes-y-Gomez, Manuel and Rosso, Paolo and Gonzalo, Julio and Aragón, Ezra and Agerri, Rodrigo and Álvarez-Carmona, Miguel Ángel and Álvarez Mellado, Elena and Carrillo-de-Albornoz, Jorge and Chiruzzo, Luis and Freitas, Larissa and Gómez Adorno, Helena and Gutiérrez, Yoan and Jiménez-Zafra, Salud María and Lima, Salvador and Plaza-de-Arco, Flor Miriam and Taulé, Mariona: Proceedings of the Iberian Languages Evaluation Forum, In: IberLEF 2021, CEUR Workshop Proceedings, 2021.
5. Mario Graff, Sabino Miranda-Jiménez, Eric Sadit Tellez, Daniela Moctezuma, Vladimir Salgado, José Ortiz-Bejar, Claudia N. Sánchez: INGEOTEC at MEX-A3T : author profiling and aggressiveness analysis in twitter using µTC and EvoMS., In: Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval2018), CEUR WS Proceedings (2018)
6. Marco Casavantes, Roberto López, Luis Carlos González-Gurrola: UACh at MEX-A3T 2020: Detecting Aggressive Tweets by Incorporating Author and Message Context. In: IberLEF@SEPLN 2020: 273-279
7. Mircea-Adrian Tanase, George-Eduard Zaharia, Dumitru-Clementin Cercel, Mihai Dascalu: Detecting Aggressiveness in Mexican Spanish Social Media Content by Fine-Tuning Transformer-Based Models, In: IberLEF@SEPLN 2020: 236-245
8. Victor Peñaloza: Detecting Aggressiveness in Mexican Spanish Tweets with LSTM + GRU and LSTM + CNN Architectures, In: IberLEF@SEPLN 2020: 280-286
9. María Guadalupe Garrido-Espinosa, Alejandro Rosales-Pérez, Adrián Pastor López-Monroy: GRU with Author Profiling Information to Detect Aggressiveness. In: IberLEF@SEPLN 2020: 246-251
10. Mario Ezra Aragón, Horacio Jesús Jarquín-Vásquez, Manuel Montes-y-Gómez, Hugo Jair Escalante, Luis Villaseñor Pineda, Helena Gómez-Adorno, Juan Pablo Posadas-Durán, Gemma Bel-Enguix: Overview of MEX-A3T at IberLEF 2020: Fake News and Aggressiveness Analysis in Mexican Spanish. In: IberLEF@SEPLN 2020: 222-235
11. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin: Attention is All you Need, In: NIPS 2017: 5998-6008

12. Mario Guzman-Silverio, Ángel Balderas-Paredes, Adrián Pastor López-Monroy: Transformers and Data Augmentation for Aggressiveness Detection in Mexican Spanish, In: IberLEF@SEPLN 2020: 293-302
13. Mircea-Adrian Tanase, George-Eduard Zaharia, Dumitru-Clementin Cercel, Mihai Dascalu: Upb at mex-a3t 2020: Detecting aggressiveness in mexican spanish social media content by fine-tuning transformer-based models, In: IberLEF@SEPLN 2020: 236-245
14. Esaú Villatoro-Tello, Gabriela Ramírez-de-la-Rosa, Sajit Kumar, Shantipriya Parida, Petr Motlícek: Idiap and UAM Participation at MEX-A3T Evaluation Campaign. In: IberLEF@SEPLN 2020: 252-257
15. Shanshan Yu , Jindian Su , Da Luo: Improving BERT-Based Text Classification With Auxiliary Sentence and Domain Knowledge, In: IEEE Access 7, 176600-176612, 2019
16. Chi Sun, Luyao Huang, Xipeng Qiu: Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence. In: NAACL-HLT (1) 2019: 380-385
17. Jian Ma, Shu-Yi Xie, Meizhi Jin, Lian-Xin Jiang, Yang Mo, Jian-Ping Shen: XSYSIGMA at SemEval-2020 Task 7: Method for Predicting Headlines' Humor Based on Auxiliary Sentences with EI-BERT. In: SemEval@COLING 2020: 1077-1084
18. J Canete, G Chaperon, R Fuentes, J Pérez: Spanish pre-trained bert model and evaluation data, In: PML4DC at ICLR 2020
19. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT (1) 2019: 4171-4186