Aggressive Language Detection using VGCN-BERT for Spanish Texts *

Errol Mamani Condori[1[0000−0001−8481−0366]] and
José Ochoa-Luna[1[0000−0002−8979−3785]]

Department of Computer Science
Universidad Católica San Pablo, Arequipa, Peru
{errol.mamani,jeochoa}@ucsp.edu.pe

Abstract. The increasing influence from users in social media has made that aggressive content disseminates over the internet. To tackle this problem, recent advances in Aggressive Language Detection have demonstrated a good performance of Deep Learning techniques. Recently Transformer based architectures such as Bidirectional Encoder Representations from Transformer (BERT) outperformed previous aggressive text detection baselines. However, most of the Transformers-based approaches are unable to properly capture global information such as language vocabulary. Thus, in this work, we focus on aggressive content detection using the combination of Vocabulary Graph Convolutional Network (VGCN) to capture global information and BERT to model local information. This combined approach called VGCN-BERT allows us to improve the feature level representation in Spanish aggressive language detection. Our experiments were performed on a benchmark called MEX-A3T aggressiveness dataset which is composed of aggressive and non-aggressive Tweets written in the Mexican Spanish variant. We report 86.46% in terms of F1-score using this VGCN-BERT approach which allows us to obtain comparable results with the current state-of-the-art, ensemble BERT, so as to detect aggressive content regarding the track MEX-A3T 2020.

Keywords: Aggressiveness Detection · BERT · Graph Convolutional Networks.

1 Introduction

The exponential growth of social media such as Twitter, community forums, and blogging platforms has revolutionized communication and content publishing, but it has also increased the dissemination of hate speech [6, 32]. Thus, nowadays offensive and aggressive language is pervasive in social media. The

* This research was supported by National Fund for Scientific and Technological Development and Innovation (Fondecyt-Perú) within the framework of the “Project of 50 E038-2019-01-BM Improvement and Expansion of the Services of the National System of Science, Technology and Technological Innovation” [Grant 028-2019-FONDECYT-BM-INC.INV].

content which has profanity, abusive, aggressive, or any kind of words that disparages a person or a group is considered hate speech. Seemingly, the language is considered aggressive if encourages any kind of hostility or intention to be harmful, conveying a violent message.

Social media platforms and technology companies have heavily invested in ways to cope with this offensive language to prevent abusive behavior in social media [29]. One of the first approaches to tackle this problem was human control on text content. This manual filtering is very time-consuming as it can cause post-traumatic stress disorder-like symptoms to human annotators. Thus, the most effective strategy is to use computational methods to identify offense, aggression, and hate speech in user-generated content. This topic has attracted significant attention as evidenced in recent publications [28, 11].

In this context, we address the automated task of aggressive language identification in Spanish texts. To do so, we use the combination of BERT model and Vocabulary Graph Convolutional Network which improves local information encoded in BERT embeddings by adding global information between words and concepts (Vocabulary GCN). This combination has been previously applied to other tasks [20, 30].

We argue that VGCN-BERT is better suited to identify aggressive language in Spanish because it allows integrating the global notion of a specific domain language. That is, in addition to being able to capture the context of a word within a sentence with BERT, we can also catch the semantic relationships between words which can be replaced with other words in the same category using VGCN.

Basically, we incorporate the feature of a word and its relationship to other words that may replace it in meaning. For example, the phrase "read a book " would be replaced with "study a book" or " understand a book". It is worth noting that some aggressive words are composed by two or more tokens. One single word rarely appears as aggressive in the corpus, and most of the deep learning models cannot even find the word in their pre-trained vocabulary. Whereby, adding an extra feature of the connected "read a book" to "study a book" meaning will add a piece of global context information to the model.

Connecting those expressions more explicitly with their meanings allows to take into account the knowledge about the language (vocabulary). For instance, let consider the following tweet of aggressive content:

> *Las fans de odisseo se ven bien bonitas en sus fotos de twitter y estan bien feas en persona.*
> (**Odisseo fans look really pretty in their twitter photos and they are pretty ugly in person.**)

The sentence "they are pretty ugly" can be related to "disgusting" and "nasty" through the connections in a GCN graph. In this example, both positive and negative words appear in the sentence and the component "Odisseo fans look really pretty" denotes a strong positive opinion. Thus, the overall text would be wrongly classified as non-aggressive due to the first part overweights the classifier.

On the contrary, if we connect explicitly the last part of the sentence "pretty ugly" with the meaning of "really nasty" taking into account the knowledge of the aggressive tweet content (vocabulary), then it would be classified more accurately.

This last correlation between words and concepts in GCN has demonstrated to have a good performance capturing global information [30]. In brief, the combination of the local (embedding) and global information (graph) through self-attention is feed into a BERT model. In this sense, the word and graph embeddings interact with each other through the attention mechanism while learning the classifier.

In our experiments using VGCN-BERT, we report 0.8646 in terms of macro F1-score. This result outperforms slightly the state-of-the-art ensemble BERT model and other deep learning techniques [15], including the baselines reported during the Mexican Spanish Aggressive detection (MEX-A3T) task [3].

Overall, the contribution of the paper can summarized as:

- Combining the local an global information using VGCN-BERT without any external word embeddings or knowledge which is suitable and novel to detect the aggressiveness words with a global notion for the Spanish language.
- The combination of VGCN with a BERT model (a Spanish pre-trained version called BETO [9]) allows to obtain comparable results regarding the ensemble BERT models in Spanish Aggressiveness detection.

The paper is organized as follows. Related works are presented in Section 2. The methodology is described in Section 3. Experiments and conclusions are presented in Sections 5 and 6.

2 Related Work

The English Offensive Language detection problem has been increasingly researched in the last years [11, 31], this task is closely related to Spanish Aggressive identification.

Although many classification approaches have been applied to hate speech [5], offensive language, and aggressive, there still very few approaches applied to the Spanish aggressive language [2] compared to English [31]. Among them, the first attempts to detect offensive language were mostly based on the surface features techniques such as *uni-grams* and a larger *n-grams*. Later Badjatiya et al. [6] found out that the character n-grams had better performance than word tokens for hate speech detection.

Unlike feature extraction approaches, classification methods for Offensive Language detection have been predominantly supervised [23]. First works focused on manual feature engineering used for Machine learning algorithms such as Support Vector Machine (SVM) [11], Naive Bayes [11], and Logistic Regression [29]. Recent Deep Learning approaches [21] have demonstrated that automated feature learning representation allows to obtain better results. Also, pre-trained Word Embeddings have been applied successfully [6].

Nowadays, the best approach for the task and related NLP tasks has been the Bidirectional Encoder Representation from Transformer called BERT [12]. Although all of those techniques have been applied to the English language, they can also be "transferred" to other languages such as Spanish. Recently, to encourage the NLP community to develop this task for the Spanish language, IberEval and IberLEF (Iberian Languages Evaluation workshops) released the Aggressive identification task[1]. Related works on Spanish are mostly based on SVM and Convolutional Neural Network (CNN) [2]. It is also worth noting that good results were obtained using SVM and Bag of Words [4] and Long Short Term Memory (LSTM) and Gate Recurrent Unit [14].

2.1 Transfer Learning with BERT

With the recent emergence of the attention mechanism in 2017 [26], many improvements began to be noted in the field of Natural Language Processing (NLP). The same could be observed with the proposal of the revolutionary BERT [12]; that proved to be good and obtained outstanding results for 10 NLP downstream tasks.

Using BERT for identifying offensive language in the English language has proven to be effective [31], obtaining results that far exceed traditional models. However, since the transformers architecture and the attention mechanism appeared, attempts to combine BERT with one extra layer of attention were proposed [22]. In more recent related works multi-task BERT and ensemble BERT models showed good results to the detection of aggressive, offensive, and misogyny language in English, as reported by TRAC in ACL[2] [18].

Seemingly, the last MEX-A3T 2020 recently reported the use of BERT transformer-based models for aggressive language in Spanish. The best results were obtained using a pre-trained BERT version for the Spanish called BETO. Fine-tuning BETO for the aggressive has proved to be useful especially if we use an ensemble of different BETO[9] models for classification. Furthermore, BETO has been integrated with AutoEncoder representation [27], and a XGBoost classifier [10]. Tanase et al. [25] tried to fine-tune BETO with different dataset and multilingual transformer models. Additionally, as evidenced in the last MEX-A3T at IberLEF 2020, 9 teams used BERT, especially BETO, and 5 obtained the best results [3].

2.2 Graph Convolutional Networks (GCN)

Throughout the development of research in natural language processing, graphs have served as a representation of documents and texts. Very recently, this idea inspired to search for the relationship of words in a language. There are many studies combining graphs with neural networks which is commonly called Graph

[1] MEX-A3T: Authorship and aggressiveness analysis in the Mexican Spanish case study

[2] Association for Computer Language conference

Neural Networks (GNN) [7]. Graphs proved to be good at capturing general knowledge about words in a language. Thus, numerous studies and variations of GNNs have been proposed and applied to text classification tasks [17, 24].

One of the most outstanding recent works showed the use of Graph Convolutional Networks (GCN) based on the spectrum of graph theory [17]. Kipf built a symmetric adjacency matrix denoting relationships in graphs. Thus, the representation of a node is also affected by its neighbors and its relationship in the graph during the convolution. Based on Kipf's work, Yao et al.[30] recently proposed a special case of GCN for text classification called text GCN. The authors denoted words and documents as nodes in a graph and their relationships as edges. Weights in edges can be calculated using three alternatives: the co-occurrence relationship between the words, tf-idf between documents and words and the similarity between documents. The work showed that GCNs are good at convolving global information from an input graph. [20] would later demonstrate the effectiveness of combining BERT with GCN using the co-occurrence measure and the relationship between words which is called Vocabulary Graph Convolutional Networks (VGCN) at the feature level.

3 Methodology

The goal of this paper is to identify whether a Spanish tweet content is aggressive or not. To do that, we carefully fine-tune the bidirectional encoder representation from transformers (BERT-Spanish version) and combine that with the VGCN.

Our design choice for BERT is due to its good performance in the English offensive detection [31]. Thus, to identify aggressive language in Spanish we have used the BERT model. However, we have noted this model is not suitable to lead with Spanish vocabulary which took us to explore other ways to improve it.

BERT is good at capturing the words' order and the local context information, but it also unable to capture global information. This missing information could be relevant to recognize aggressiveness specially in long text sequences. Starting from this point, we have already seen that GCNs for text classification are good at extracting global information [30]. Intuitively we thought of combining both approaches to join local and global text information. Fortunately, Yao et al. [30] focused on the feature level of BERT adding GCN features. In this sense, to address the missing global information in BERT, an additional embeddings graph would be needed.

Having said that, in this work we improve the BERT (BETO) model with a GCN. We aim to capture global notion and semantic correlation between words using the graph as a vocabulary of words following the previous proposals for text classification [16]. Thereby, the GCN-BERT architecture based on a vocabulary graph was deemed into the pipeline of our system.

The Vocabulary graph enables the lexical relationship between words in a language. The graph is built using the co-occurrence of words with documents such as Lu did [20]. The local information of a tweet is captured by BERT and is enriched by adding a graph embedding. We also must highlight that the construction of the vocabulary graph is based on the vocabulary of BERT, in such a way that the size is reduced. At the same time, the added embedding can have an efficient interaction and improve local context information of BERT.

By doing so, the relevant part of the global vocabulary graph is selected according to the input sentence and transformed into an embedding representation. Then VGCN-BERT uses multiple layers of attention mechanism to concatenate both embeddings representations allowing the interaction between BERT embeddings and graph embeddings. Finally, the classifier uses the fully connected layers, more details are depicted in Figure 1.

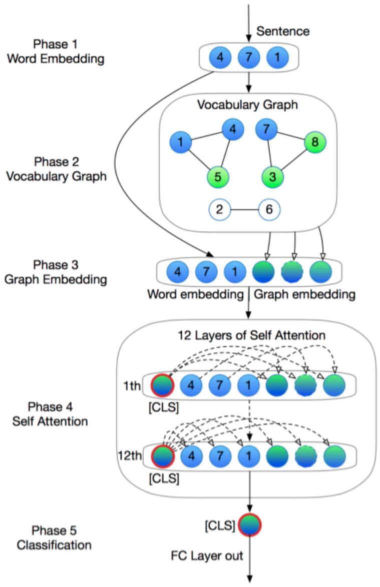

Fig. 1. Illustration of VGCN-BERT. (Blue) represents the input words, (green) are the related words in the graph, and the (blue-green) blend is the convolutional representation. The embeddings of the input sentence (Phase 1) are combined with the vocabulary graph (Phase 2) to produce a graph embedding, which is concatenated to the input sentence (Phase 3). Phase 2 produces three (hyperparameter) graph embeddings which have the same shape as the word embeddings. Note that from the vocabulary graph, the relevant part from input is extracted and embedded. In this case (3,5,8) will be convoluted but (2,6) will not. (Phase 4), several layers of self-attention are applied to the concatenated representation, allowing interactions between word embeddings and graph embeddings. The final embedding at the last layer is fed in a fully connected layer (Phase 5) for classification[20].

By applying the VGCN-BERT we aim to add carefully the global information of a language using vocabulary (of BERT) graph embedding to the classical embedding BERT (Spanish pre-trained BETO) in the features level. The system we used obtained good results improving the detection of the aggressiveness words in contrast to the other ensemble models using multiple BERT models at the same time. The Vocabulary graph knowledge from global information and its background concepts of the words will be added to the local information throughout multiple self-attention mechanisms. The VGCN and the integration to BERT will further explained in next sections.

3.1 VGCN

Our system follows the implementation and the model proposed by Lu [20], which uses a vocabulary graph based on the word co-occurrences in documents. This graph is built using normalized point-wise mutual information (NPMI)[8] to measure the relationship between words in a document. Then, they create an edge between two words if their NPMI is larger than a threshold and the performance is better when the threshold is between 0.0 and 0.3.

The GCN is generally based on Kipf's model [17], i.e., it has two multilayer neural networks that convolve directly on the graph and create embedding vectors of nodes based on the neighborhoods. In this case, the convolution is made from related words instead of documents and the GCN is constructed in the vocabulary. Thus, for a single document, assuming the document is a row vector x consisting of words in the vocabulary, a layer of convolution is defined as (1).

$$h = (Ax^T)^T W = xAW \tag{1}$$

where $A^T = A$ is the vocabulary graph. xA extracts the part of vocabulary graph relevant to the input sentence x. W holds the weights of the hidden state vector for the single document, whose dimension is $|v|h$. We could define the two layers VGCN with ReLU as follows:

$$VGCN = ReLU(X_{mv}A_{vv}W_{vh})W_{hc} \tag{2}$$

where m is the number of documents in mini-batch size. v is the vocabulary size, h is the hidden layer size, c the class size or sentence embedding size [17].

3.2 Integrating VGCN into BERT

In this section, we want to integrate VGCN embeddings with BERT and take advantage of its attention mechanism [26]. When using the attention mechanism one gets weighted vectors that encode the context information. Thus, instead of using only word embeddings of the input sentences, we feed both the vocabulary graph embedding obtained with the equation (2) and the sequence word embeddings to the BERT transformer.

In this sense, not only the order of the words in the sentence is retained (local context information), but also the background information obtained by VGCN (global information of the language). After that, the local and global embedding are fully integrated through layers interacting them with 12-layers and 12-heads of the self-attention encoder.

4 Dataset

Since our goal is to identify the aggressiveness content in social media for Spanish, we worked with the Mexican Spanish dataset MEX-A3T track [3]. We also noted that this track on detecting aggressive language is not new; however, it is still a challenge for the Spanish language as shown in the previous workshops: TRAC2020 [4], iberLEF2020[5] and the track at iverEval2018.

The dataset was collected considering Twitter as a main source media since it is open and its anonymity allows people to write judgments about others. While building the corpus they use some rude words and controversial hashtags to narrow the search that ranged from August until November of 2017. They also used around 143 terms that served as a seed to extract tweets [13]. Additionally, the MEX-A3T contains 10.475 tweets 3143 for testing and 7332 for training (see Table 1) and we also noted that the test set does not contain labels in the dataset.

It is worth noting that the data (tweets) contains a dictionary of Mexican words "Mexicanisms" with at least one of vulgar or insulting words; then they were manually labeled with "Aggressive" and "Non Aggressive". The criteria used while tagging was on the promise that offensive content is disparaging or humiliating a person or a group of persons [3]. The linguistic criteria are the approach of using "vulgar, Aggressive and offensive" as an identifier. Diaz et al. [13] used a new annotation scheme based on the linguistic characteristics and intent of the message.

Table 1. MEX-A3T tweets with Aggressive data set distribution of classes.

Class	Train Corpus	Test Corpus
Non Aggressive	5222	2238
Aggressive	2110	905
Total	**7332**	**3143**

The aggressive content samples are shown below, we use # symbol to hide the aggressive word in purpose to avoid any offensive words for the reader.

Aggresssive Samples:
"Sólo a las Pu#..aS MOCOS#S GORDAS Y feas les gusta ese al%man xd"
"Profe hijo de las mil pu#..as 6 de calificación como es posible. "

Non Aggressive Samples:
"Put#s Madres ahora comprendo todo, tu tan lava y yo tan frio "
" Segunda vez que me pasa. Estoy hasta la madre"

[3] mex-a3t site: https://mexa3t.wixsite.com/home
[4] II Trolling Aggressive and Cyberbylling workshop
[5] Iberian Languages Evaluation Forum 2020

We follow [15] to split the data, i.e., 80% for training and 20% for testing.

The picture below describes this distribution. We can see that the amount of non-aggressive content is greater which means our data is imbalanced (Figure 2), but also it reflects the true distribution of data (more detailed in [1]).

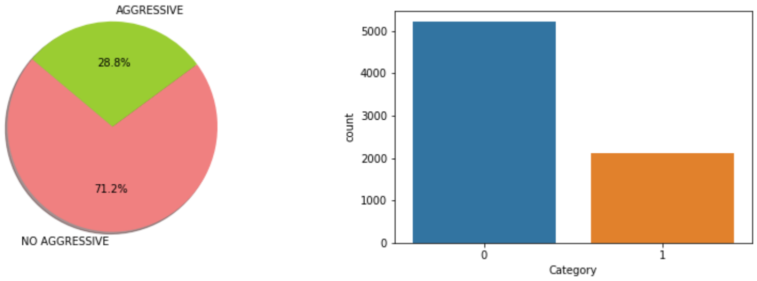

Fig. 2. MEX-A3T DataSet with 5222 labeled distribution, left :28.8% (green)AGGRESSIVE, 71.2% NO AGGRESSIVE (pink), right: bar shows the majority class of the "non-aggressive" (blue) class 0 and 2110 for "aggressive"(orange) class 1.

5 Experiments

This section presents the baselines models, pre-processing techniques, experiments, evaluation methodology, and results. We also discuss how we address the issue of vocabulary performance within Aggressive identification.

5.1 Baselines and Task

In order to investigate the Aggressiveness detection for Mexican Spanish Language, we studied the previous techniques and research lines. According the literature and the best performance models, we re-implement the current state of the art model proposed during the MEX-A3T 2020 track [15]. We also took other baselines [3]. Besides BERT there are others neural network models that we will detail below:

- **BOW with SVM** (base-line): The approach is presented as a traditional machine learning approach. The model uses a bag-of-words representation with TF and TF-IDF weights. One of their good results is obtained when applying a SVM and other neural network [4].

- **BI-LSTM** (base-line): This deep learning model approach is a bi-directional GRU model that uses words as inputs. The model also combines the predictions of gender and occupation of users obtained by a reference model, it uses a simple concatenation and one-hot-encodings [14].
- **Transformers** (State-of-the-art): the approach uses the novel pre-trained BERT and performs fine-tuning to accomplish the detection of aggression for Spanish. The BERT variation used is the Spanish transformer-based BERT called BETO [6]. There are also BETO improvements using multiple ensembles of these models, weighting vote schemes and different Data augmentation techniques [15].

Most of the models above work well; however, we also explore deep learning classification models in other languages. Thus, we found a good performance of the VGCN-BERT for English text classification task. Of course, there are other attempts to combine VGCN with BERT (see Section2). We will briefly describe the vanilla model we based on:

- **vanilla-VGCN-BERT**: Our system is mainly based on this model which combines BERT and VGCN. First, the model produces two separate representations through BERT and GCN. Both features are concatenated.Then, fully connected layer and Rectified Linear Unit (ReLU) are applied to classify [16].

There is a large number of models that obtained good results for the task of aggressive content detection on Twitter. We were also able to describe some other scopes that were given for the classification of text in the English language. However, we noted that multi-tasking models or combined models have not yet been proposed to improve the detection of aggressive language especially for Spanish Language. Thereby, we took as a baseline the works described above for the aggressive language detection for Mexican Spanish.

5.2 Pre-processing and Model setting

When using a dataset that came from social media usually has a lot of noise and every tweet needs a pre-processing technique. To implement this work we first noted that there are URL strings to be removed as well as "@", i.e, we retain word text content and remove all unnecessary symbols. After cleaning, we also convert all the text into lower case and tokenized using social media *tweetTokenizer* library from NLTK [7]. We also use the BERT Tokenizer to split the text, so the vocabulary of GCN is always a subset of the BERT vocabulary.

We fine-tune VGCN-BERT model and train the model for the Spanish Aggressive language detection because we want to capture the global context. To get the best performance we are using VGCN with NPMI for measuring the

[6] **BETO** is a pre-trained BERT on a large Spanish corpus [9]

[7] http://www.nltk.org/api/nltk.tokenize.html

correlation between words in the vocabulary such we reduce the range of the meaningful relationship up to 0.2.

The graph embedding size used in the VGCN is set 16 and the hidden graph embedding dimension is 128. Whereas, the pre-trained BERT model use 200 length max of the sequence. Then, the whole model is trained in 8 epoch with a dropout of 0.2. Furthermore, due to the limitation of our GPU memory we use mini batches of 8 to seed the data and a learning rate of 8e-6, L_2 loss weight decay of 0.01.

It is worth to mention that all these parameters were defined based on our experimental test results. We also follow the parameter setting of L_2 loss weight and the learning rate according to the paper [20]. We also opted to use cross-entropy to optimize the classification. On the other hand, Adam optimizer and class weights were estimated to handle the unbalanced data using *compute_class_weight* [8] .

5.3 Evaluation Metrics

As we are going to compare VGCN BERT against baselines approaches, we adopt the widely used metric F1-score for text classification, which was also used by [13]. The weighted average F1-score and the macro F1-score were defined as follows [19].

$$F1 = 2 * \frac{precision * recall}{precision + recall} \tag{3}$$

$$\text{Weigthed avg F1} = \sum_{i=1}^{C} F1_{ci} * W_{ci}, \qquad \text{Macro F1} = \frac{1}{C}\sum_{i=1}^{C} F1_{ci} \tag{4}$$

where C: is the each individual different class.

5.4 Experimental Results

In this section, we show the results of our VGCN-BERT system and to compare it against the best models for the Mexican Spanish aggressive detection task. We carefully selected the previous outstanding approaches considered as baselines for this task. We also highlight the implementation of the BERT (multi-language) combined with VGCN comparing to the best (SOTA) Spanish pre-trained BERT (BETO). We also found that 5 of the best winning models were based on BETO and clearly outperformed traditional deep learning methods [3]. Those methods are references to compare results under the same criteria. However, to the best of our knowledge, there are very few BERT models combined with other models. Also, the BERT multi-lingual was not much used during the MEX-A3T aggressiveness analysis track at iberLEF 2020 competition. Thereby, our BERT with

[8] https://scikit-learn.org/.../sklearn.util..._weight.compute_class_weight.html

VGCN system becomes a good alternative for the detection of aggressiveness, since our results show that it clearly outperformed the baselines models as shown in Table 2.

The table shows the “macro F1-score” evaluated with the test set for every model. To compare our proposal against the others, we use the average F1-macro score. We have also shown the scores for the aggressiveness and non-aggressiveness detection tasks. In addition, we also report the best model performed using ensembled BETO models. We noted that the closest score to ours is the BERT ensemble (20 BETOS). Another good classifier is BETO+XGBoost referred to as BETO+msg. On the other hand, we also re-implemented the single BERT-multi-language, the single BERT (Spanish pre-trained BETO), and the aforementioned ensembled BETO following Guzman et al. paper’s [15]. This last model varies from 1 BETO to 20 BETOs and was used a voting scheme to compare them against our results.

Table 2. Results of the Aggressive Detection using F1-score in test set.

Model	F1 Aggressive	F1 non-agressive	F1 macro
BoW-SVM	0.6760	0.8780	0.7770
BI-GRU	0.7124	0.8841	0.7983
BETO+msg	0.7720	0.9042	0.8381
bert (**multi***)	0.7809	0.9094	0.8452
bert (**1 BETO**)	0.7998	0.9195	0.8596
bert (20 BETOs)	0.7994	0.92.23	0.8608
VGCN-BERT *	**0.8124**	**0.9169**	**0.8642**

* meas our developed system model.

Those results demonstrate that our system which combined VGCN with the fine-tuned BERT (*BETO-uncased*) effectively identifies whether a given tweet contains Aggressive content or not. Furthermore, we see that all models (except ours) have low values when they classify the minority aggressive class.

It is worth emphasizing that even in the case we are using a multi-language BERT model(Not BETO) with VGCN, referred to as **bert multi***, we are still able to outperform BETO+XGBoost. That shows the improvement over pre-trained BERT model for Spanish text. We also noted that our system obtained a high F1-score on the aggressive class. We obtained the third score on the non-aggressive class. We argue that those results are due to the global information which adds attention to the aggressive class with the vocabulary. One interesting thing that we found is that the VGCN-BERT, without any additional external embedding, obtained comparable results to BERT (20 BETOs).

Preliminary studies for English text classification did not show impressive improvements using VGCN and BERT. However, in this paper we have shown that this model is suitable and fits better in the multi-language scenario. Despite all apparent disadvantage, we have shown that Global information captured using VGCN can contribute to the classification task. Likewise, if we add global to the local context information from BERT then the result is very good at detecting Aggressive language.

6 Conclusion

In this paper, we have explored the tight combination of the transfer learning fine-tuning BERT (Spanish pre-trained BETO) model and the Graph Convolutional Network to identify aggressive content for Spanish tweets. We have shown that adding a vocabulary graph, even in a small set of vocabulary for Spanish BERT, could improve the context local information. That was possible since the GCN has proven to be good at capturing the global information of a language. Our experimental results allowed us to state that using VGCN-BERT could be beneficial to detect Aggression content, specially for the Spanish language.

On the other hand, although our system used the original classic BERT version, with disadvantages compared to ensembled BETOs, it is still able to perform well when combined with another neural network such as GCN. Thus, BERT could be improved using the Vocabulary Graph Convolutional Network as our promising results show.

We have also demonstrated that it is still possible to improve the original BERT with models that increase the global knowledge information about the domain-specific language.

Based on our results, we also argue that, for detecting the aggressiveness in tweets, a vocabulary and global information of a language is a path for improving the BERT model as a classifier. Furthermore, we believe that a word within a sentence could contribute a lot if its relationship within a tweet contributes to detect aggressiveness. We also leave open another way to explore GCN for text classification.

References

1. Álvarez-Carmona, M.Á., Guzmán-Falcón, E., Montes-y Gómez, M., Escalante, H.J., Villasenor-Pineda, L., Reyes-Meza, V., Rico-Sulayes, A.: Overview of mex-a3t at ibereval 2018: Authorship and aggressiveness analysis in mexican spanish tweets. In: Notebook Papers of 3rd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL), Seville, Spain. vol. 6 (2018)
2. Aragón, M.E., Álvarez-Carmona, M.Á., Montes-y Gómez, M., Escalante, H.J., Villasenor-Pineda, L., Moctezuma, D.: Overview of mex-a3t at iberlef 2019: Authorship and aggressiveness analysis in mexican spanish tweets. In: Notebook Papers of 1st SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Bilbao, Spain (2019)
3. Aragón, M., Jarquín, H., Gómez, M.M.y., Escalante, H., Villaseñor-Pineda, L., Gómez-Adorno, H., Bel-Enguix, G., Posadas-Durán, J.: Overview of mex-a3t at iberlef 2020: Fake news and aggressiveness analysis in mexican spanish. In: Notebook Papers of 2nd SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Malaga, Spain (2020)

4. Arce-Cardenasa, S., Fajardo-Delgadoa, D., Álvarez-Carmonab, M.Á.: Tecnm at mex-a3t 2020: Fake news and aggressiveness analysis in mexican spanish
5. 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)
6. Badjatiya, P., Gupta, S., Gupta, M., Varma, V.: Deep learning for hate speech detection in tweets. In: Proceedings of the 26th International Conference on World Wide Web Companion. pp. 759–760. International World Wide Web Conferences Steering Committee (2017)
7. Battaglia, P.W., Hamrick, J.B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., et al.: Relational inductive biases, deep learning, and graph networks. arXiv preprint arXiv:1806.01261 (2018)
8. Bouma, G.: Normalized (pointwise) mutual information in collocation extraction. Proceedings of GSCL pp. 31–40 (2009)
9. Canete, J., Chaperon, G., Fuentes, R., Pérez, J.: Spanish pre-trained bert model and evaluation data. PML4DC at ICLR **2020** (2020)
10. Casavantes, M., López, R., González, L.: Uach at mex-a3t 2020: Detecting aggressive tweets by incorporating author and message context. In: Notebook Papers of 2nd SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Malaga, Spain (2020)
11. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Eleventh International AAAI Conference on Web and Social Media (2017)
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
13. Díaz-Torres, M.J., Morán-Méndez, P.A., Villasenor-Pineda, L., Montes, M., Aguilera, J., Meneses-Lerín, L.: Automatic detection of offensive language in social media: defining linguistic criteria to build a mexican spanish dataset. In: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying. pp. 132–136 (2020)
14. Garrido-Espinosa, M., Rosales-Pérez, A., López-Monroy, A.: Gru with author profiling information to detect aggressiveness. In: Notebook Papers of 2nd SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Malaga, Spain (2020)
15. Guzman-Silverio, M., Balderas-Paredes, A., López-Monroy, A.: Transformers and data augmentation for aggressiveness detection in mexican spanish. In: Notebook Papers of 2nd SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Malaga, Spain (2020)
16. Jeong, C., Jang, S., Park, E., Choi, S.: A context-aware citation recommendation model with bert and graph convolutional networks. Scientometrics **124**(3), 1907–1922 (2020)
17. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
18. Kumar, R., Ojha, A.K., Lahiri, B., Zampieri, M., Malmasi, S., Murdock, V., Kadar, D.: Proceedings of the second workshop on trolling, aggression and cyberbullying. In: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying (2020)
19. Lever, J., Krzywinski, M., Altman, N.: Classification evaluation (2016)

20. Lu, Z., Du, P., Nie, J.Y.: Vgcn-bert: augmenting bert with graph embedding for text classification. In: European Conference on Information Retrieval. pp. 369–382. Springer (2020)
21. Park, J.H., Fung, P.: One-step and two-step classification for abusive language detection on twitter. arXiv preprint arXiv:1706.01206 (2017)
22. Samghabadi, N.S., Patwa, P., Srinivas, P., Mukherjee, P., Das, A., Solorio, T.: Aggression and misogyny detection using bert: A multi-task approach. In: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying. pp. 126–131 (2020)
23. Schmidt, A., Wiegand, M.: A survey on hate speech detection using natural language processing. In: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. pp. 1–10 (2017)
24. Shang, J., Ma, T., Xiao, C., Sun, J.: Pre-training of graph augmented transformers for medication recommendation. arXiv preprint arXiv:1906.00346 (2019)
25. Tanase, M.A., Zaharia, G.E., Cercel, D.C., Dascalu, M.: Detecting aggressiveness in mexican spanish social media content by fine-tuning transformer-based models
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. arXiv preprint arXiv:1706.03762 (2017)
27. Villatoro-Tello, E., Ramırez-de-la Rosa, G., Kumar, S., Parida, S., Motlicek, P.: Idiap and uam participation at mex-a3t evaluation campaign. In: Notebook Papers of 2nd SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Malaga, Spain (2020)
28. Waseem, Z., Davidson, T., Warmsley, D., Weber, I.: Understanding abuse: A typology of abusive language detection subtasks. arXiv preprint arXiv:1705.09899 (2017)
29. Waseem, Z., Hovy, D.: Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In: Proceedings of the NAACL student research workshop. pp. 88–93 (2016)
30. Yao, L., Mao, C., Luo, Y.: Graph convolutional networks for text classification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 7370–7377 (2019)
31. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R.: Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). arXiv preprint arXiv:1903.08983 (2019)
32. Zhang, Z., Luo, L.: Hate speech detection: A solved problem? the challenging case of long tail on twitter. Semantic Web (Preprint), 1–21 (2018)