Deep Learning-based Natural Language Processing Methods Comparison for Presumptive Detection of Cyberbullying in Social Networks

Diego A. Andrade-Segarra[1], Gabriel A. León-Paredes[2]
Grupo de investigación en Cloud Computing Smart Cities & High Performance,
Universidad Politécnica Salesiana,
Cuenca, Ecuador, 010102

Abstract—**Due to TIC development in the last years, users have managed to satisfy many social experiences through several digital media like blogs, web and especially social networks. However, not all social media users have had good experiences with these media. Since there are malicious users that are able to cause negative psychological effects over people, this is called cyberbullying. For this reason, social networks such as Twitter are looking to implement models based on deep learning or machine learning capable of recognizing harassing comments on their platforms. However, most of these models are focused on the use of English language and there are very few models adapted for Spanish language. This is why, in this paper we propose the evaluation of an RNN+LSTM neural network, as well as a BERT model through sentiment analysis, to perform the detection of cyberbullying based on Spanish language for Ecuadorian accounts of the social network Twitter. The results obtained show a balance between execution time and accuracy obtained for the RNN + LSTM model. In addition, evaluations of comments that are not explicitly offensive show a better performance for the BERT model, which outperforms its counterpart by 20%.**

Keywords—***Bidirectional Encoder Representations from Transformers, BERT; Cyberbullying; Natural Language Processing; Recurrent Neural Network + Long Short Term Memory; RNN+LSTM; Sentiment Analysis; Semantics; Spanish Language Processing***

I. Introduction

Nowadays, the constant information and communication technologies (ICTs) development has enabled interpersonal interaction, allowing real experiences to be transferred to a virtualized medium such as social networks (SSNs) [1]. Within this environment, SSNs users can establish real time conversations to exchange ideas [2]. This satisfies the need for affection and integration through positive reinforcers and facilitates socialization among its users [3].

However, there are not only positive reinforcements for the use of social networks, but there are also risks such as, emotional distancing, loss of limits in communication, sexting, cyber addiction, or cyberbullying [4]. In cyberbullying, malicious users take advantage of situations of vulnerability to attack others through offensive comments using digital media. Although cyberbullying does not produce physical injuries, it causes negative psychological effects and disorders in users who are victims of these practices [5]. For this reason, social platforms such as Twitter offer methodologies based on social systems for cyberbullying recognition, such as Social Filter or the Theory of Planned Behavior, have become increasingly popular [6]. These methodologies are able to identify stalkers based on the allegations in their messages and the behavior of the account. However, these techniques tend to be ineffective in the face of occasional or unidentifiable comments.

Similarly, different countries take preventive measures through campaigns or laws that limit and sanction this type of practice. In the case of Ecuador, the Ministry of Education agreed on an operational definition for this type of practices within educational environments, described as violent acts that are frequently carried out intentionally between students of an educational institution, in a relationship of imbalance of power. Through it, a bully seeks to assert superiority in a group. [7]. For its part, the National Agency for Intergenerational Equality establishes that the competent authorities and entities must create an inter-institutional strategy to prevent, detect and address all forms of harassment within educational institutions (*bullying*, *ciberbullying*, sexual harassment) [8]. One of the most common forms of harassment in Ecuador, is cyberbullying which has a 7.46% share of the total, which puts it ahead of the total number of cases where victims were beaten in the same period with 7.02% [7]. It is worth mentioning that these statistics are limited to cyberbullying in educational environments.

II. Background

Following the ideas on [9], [10], [11] they carry out campaigns through images, talks, marches and trends in SSNs on tolerance and social respect to prevent cyberbullying. Despite this, the campaigns carried out, being a preventive method, are not able to act on a perpetuated act or event of cyberbullying. For this reason, there are methodologies based on deep learning and machine learning capable of identifying comments with potential cyberbullying [12]. Based on this resource, it is possible to take the necessary measures against the users involved. In this regard, deep learning offers cyberbullying prediction models based on Support Vector Machine (SVM), Linear Regression (LR) and Naive Bayes (NB) [13], [14], [15], [16].

However, the studies carried out are limited to work in the English language, which results in greater vulnerability for Spanish-speaking user groups in SSNs.

For this reason, works such as [17] and [18] test SVM, LR and NB models to perform sentiment analysis based on Spanish language. Antagonistically, these evaluations are focused on traditional methods for sentiment analysis tasks with NLP. In this sense, traditional methodologies show a good level of accuracy when performing sentimental analysis.

However, methods based on Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), Memory Networks (MemNNN), Long Short-Term Memory (LSTM) and Recursive Neural Network (RecNN) it present a more robust and efficient result in the area of Natural Language Processing (NLP) [19], that represent an area of study for cyberbullying cases in Spanish language.

For this reason, in recent years, neural networks are dedicated for sentimental analysis tasks in SSNs; becoming a focus of research interest, offering models capable of classifying various sentimental topics from texts of limited length [20]. At the same time, these works are focused on empirical experiments whose adjustment seeks to avoid overfitting, to allow the implementation of the models in SSNs.

One such model is found within [21], where a sentimental analysis based on a convolutional recursive neural network (RvNN) coupled to pretrained word vectors in the upper layer for English language is proposed. Thus, it uses the RvNN as an alternative to the CNN clustering layer, thus reducing the loss of local information and allowing the model to efficiently capture long-term dependencies based on its tree structure. Consequently, a classification accuracy of 89.1%, as it extracts sentimental representations at word level considering syntactic and semantic relations, similar to what was achieved with techniques such as Latent semantic analysis (LSA) proposed in [22], [23].

On the other hand, [24] proposes a CNN-based text classification model for a dataset of verbal aggression in English language. The experiment combines the LSTM and a CNN to a 2-dimensional embedding Term Frequency Inverse Document Frequency (TF-IDF) layer. The obtained result reaches 91—% accuracy for the LSTM+2D TF-IDF and 92% accuracy for the CNN+2D TF-IDF model. Thus, it increases the accuracy achieved by the LSTM with 72% and CNN with 83% when using a preprocessed embedding layer such as Word2Vec [25].

In 2018, [26] proposes a pre-training method based the deep bidirectionality of BERT. Thus, his model is trained using the ”masked language model” (MLM) but without the ”next sentence prediction” (NSP) task and a left context model that is trained using a standard left-to-right LM (LTR), instead of an MLM. Then, it manages to improve the accuracy of traditional BERT by 2.5%. In this context, [27] proposes a model that uses a single linear neural network layer for classification. In its experimentation, it achieves an accuracy of 81% when using the Wikipedia Corpus.

Also, [20] conducts an evaluation of BERT for the detection of cyberbullying on social media: Twitter, Wikipedia, and FormSpring. The purpose of their evaluation is to refine the parameters of BERT when using the [26] Corpus to achieve better results. Therefore, in evaluating the model, it achieves an accuracy of 90% when using the Wikipedia Corpus. However, its best performance was achieved with a Twitter Corpus with which it obtains 94% detection accuracy. While, an RNN+LSTM model evaluated with the same Corpus only achieves 85% accuracy.

As a Spanish language case, we found that [17] performs the detection of violence against women in SSNs, based on Opinion mining techniques, Document Term Matrix (DTM), such as *bolsa de palabras* (bag of words), together with NB, Decision Trees (DT) and SVM algorithms for the social network Twitter. It´s most relevant result shows an accuracy of 90.35% when using the DT algorithm.

Thus, [18] presents a semi-automatic and presumptive detection system of cyberbullying; based on NB, SVM and LG techniques for the social network Twitter in Spanish language, called Spanish Cyberbullying Prevention (SPC). Thus, [18] uses sentiment analysis, machine learning and NLP to analyze phrases in Spanish from a group of trends and specific users from Ecuador, in order to identify the existence of a semantic set related to cyberbullying with up to 91% of precision, when using SVM.

In this context, we continue the work presented in [18], where a comparison is made between two models of deep learning; the first one based on RNN+LSTM and the second one on BERT for the presumptive detection of cyberbullying in Spanish language. Our goal is to obtain accurate and efficient models in terms of execution time, using bidirectional models with memory capabilities to train the models with a Spanish language Corpus.

The rest of the document is presented as follows: Section III describes the methodology used for the proposed method. Section IV analyze the results obtained. Finally, section V presents the conclusions of this paper.

III. Methodology

After reviewing several proposals of deep learning works for sentimental analysis focused on cyberbullying, we propose the implementation represented in Fig 1, composed of 3 stages: Dataset extraction, deep learning training and testing the trained models.

A. *Dataset Extraction*

Prior to training deep learning models, it´s important establish a dataset according to the subject matter. Therefore, it is necessary to generate a customized Corpus that fits the needs of the research. For this purpose, Twitter´s API and its advantages were used to form a Corpus through the extraction, processing and analysis of tweets through several implemented scripts; which were subsequently used for the proper training of the proposed models. Scripts with NLP algorithms are used to semi-automatically identify the sentiments of each tweet since its content is related to cyberbullying situations and a correct labeling of the data depends on it.

Main Corpus elaboration starts establishing a set of phrases and keywords that suggest some kind of insult or harassment; as well as others that show a more pleasant type of feeling, as shown in Table I. Thus, a set of weightings is established to determine the choice of an appropriate label for a comment. The label can be between 0 and 1, where the first value indicates that a comment is negative or has harassment and the second one when it is a positive or clean one. In turn, the Corpus used here has been implemented and developed by [18].

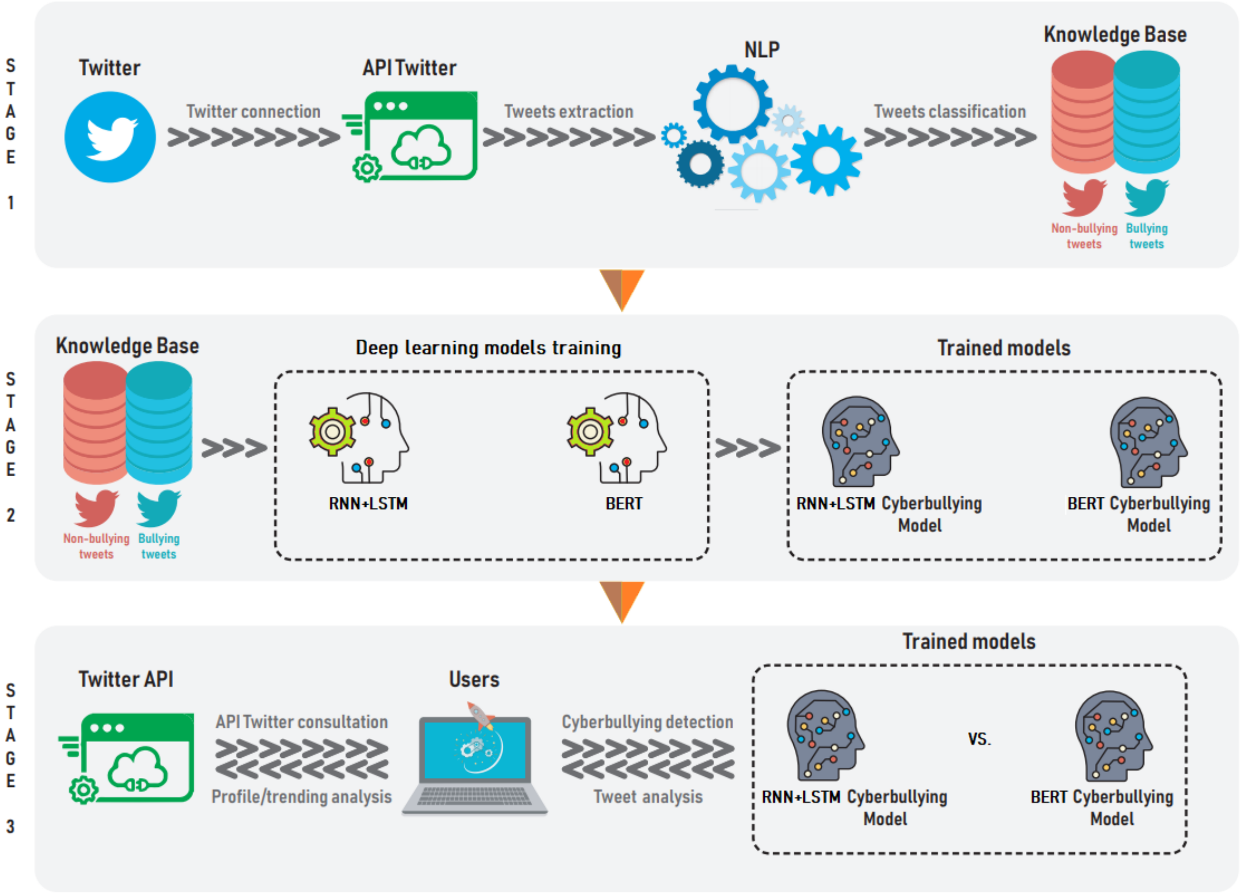

Fig. 1. Implementation Stages of Deep Learning Models.

TABLE I. Sample Corresponding to 6.25% of Phrases used for Tweets´s Automatic Classification According to a Positive or Negative Intent Focus

Frases Positivas	Frases Negativas
amar	desgraciado
eres increíble	gay
genial	hija de put*
gracias	indio
muy amable	vales verg*

However, to establish a margin of error, we have developed a second manually classified corpus. This is how a comparison is established between the effectiveness of an automatic classification method and one with personal criteria. The second Corpus represents 1.36% of information respect to automated version Corpus. The result is a difference of 20% less in the interpretation of the data considered as "harassment" by the automatic model compared to the manual model. Subsequently, a data cleaning process was applied to remove mentions, emoticons and retweets because they affect the content of the comments and, consequently, could generate false negatives or positives in the system. As a result, a Corpus was formed with 881503 tweets, where 471432 of them correspond to comments with some type of harassment.

It should be noted that, as described in [18], a limitation of the Corpus is the impossibility of analyzing irony, since both verbs and contextual structures are not taken into consideration.

B. *Deep Learning Training*

To detect comments with presumed cyberbullying, our designed system uses a training corpus composed of 53.48% of comments or phrases with bullying. From this, the use of two models of deep learning is considered to detect and classify negative sentiments in Spanish language.

The proposed system has been developed from two methods: RNN + LSTM and BERT, because they are currently the deep learning models with the highest level of accuracy and development in prediction and classification tasks in the area of sentiment analysis.

1) RNN+LSTM Model: Sequence classification is defined as a predictive modeling problem where we have an input sequence over space or time and our goal is to predict a category where to assign that sequence [28]. Hence, a classification model development starts with the implementation of a block that allows text preprocessing, so that a deep cleaning is applied to the content of the original comments.

The stipulated cleaning includes special characters, punctuation symbols and repeated words elimination.

Also, it is necessary to apply *"Stemming"* or *"Lemmatization"* techniques that allow delete word endings to obtain their base structure, all that to avoid an excessive amount of identifiers, which may result in an inefficient classification system.

On the other hand, word representations are a fundamental component of many NLP systems, so it is common to represent words as indexes in a vocabulary [29]. However, this fails to capture the corresponding lexical structure. Hence, the importance of text preprocessing and the use of techniques that allow a mathematical fitting of a one-dimensional word space to a continuous vector space with fewer dimensions.

For the case of RNN, there is *Word Embedding*, known as a set of techniques for language modeling and NLP learning techniques. Its use represents an efficient way to clean, coding and vectorize words.

Thereby, the technique allows us know the spatial position of a word when it´s part of a vector, its characteristics and surrounding words. Here, a Word Embedding technique *"Word2Vec"* was proposed by [25] to delimit the number of most common words in a dataset, as well as their length provided by this technique. For this reason, with a greater features number defined, a better level of interpretation will be reached during training stage. Also, it should be considered that the system will take as reference only the words defined as the most common ones, since the rest of the words are labeled with a 0. This process is carried out automatically during the Tokenization process, where the larger text strings are divided into smaller parts or Tokens; in this way the larger text samples can be converted into sentences and these in turn can be tokenized into words. From Tokenization, it is possible to analyze the number of times each word appears in the dataset, as shown in Table 2.

TABLE II. DATASET´S MOST USED WORDS SAMPLE

Top	Word	Top	Word
1	hac	**11**	grac
2	buen	**12**	amar
3	quer	**13**	sol
4	dec	**14**	mejor
5	put	**15**	dia
6	ir	**16**	pas
7	pod	**17**	verg
8	amig	**18**	tan
9	ser	**19**	gust
10	ver	**20**	hol

Based on the most used words, a dictionary was defined to match the variety of words in the comments with the tokenization of the vocabulary. After that, the comments were transformed into integer sequences. These sequences are truncated or filled in so that they all have the same length, prior to assigning the data to Train and Test blocks. Subsequently, the Train and Test blocks were defined in portions of 25% and 75% respectively. After padding the sequences, two matrices were obtained; the first one corresponding to X-Train of (220375x500) and the second one assigned to X-Test of (661126x 500), where 500 corresponds to the maximum length sequences. On the other hand, the labels in Y-Train and Y-Test are directly extracted from the original data set.

On the other hand, our RNN+LSTM model main feature is to use of a Bidirectional LSTM layer. Its use implies a remarkable improvement compared to a standard LSTM layer. A single LSTM layer is only capable of retaining information from a past state because it only considers one input related to past state. In contrast, a bidirectional LSTM layer can handle past and future states combined and thus give a better understanding context to the model to be trained.

Our RNN+LSTM model is composed as follows: an input layer, followed by an Embedding layer, commonly used in models with text. Next, a Batch normalization layer is placed, to adjust the dimensions of the data in the Embedding layer to the rest of the layers. Subsequently, a Bidirectional layer is applied, to provides our RNN model with a two-way LSTM memory; this adds the BERT´s main feature to the RNN+LSTM, bidirectionality.

2) BERT Model: Preprocessing of the dataset for the BERT model was performed using*stopwords* as a technique for removing special characters and punctuation symbols not relevant to the meaning of a sentence. [30]. Subsequently, as in RNN+LSTM, a tokenization process was applied. However, BERT has it´s own library to perform preprocessing and tokenization tasks called *BERT Tokenizer*. The library, by itself, splits text into tokens and at the same time converts those tokens into tokenizing vocabulary indexes. In addition, it allows to limit or equalize the sentences to equal length and to create an attention mask; where the last parameter is a necessary component for the training of the model. Through this, sentences were constrained into vectors with a maximum length of 64 identifiers of tokens each. Thus, as in the RNN+LSTM model, the dataset was divided into Train and Test blocks in portions of 25% and 75%, respectively.

By using a pre-trained model, BERT presents an efficient prediction accuracy when compared to other methods of deep learning. However, it´s possible increase the model´s accuracy level through Fine-Tunning techniques. This thanks to the *Transformers by Hugging Face* option in the *PyTorch* class. This do a settings, tokenization and architecture optimizing since base model; all thanks because Transformers provide a general-purpose architectures for natural language understanding with about 32 pre-trained models in more than 100 languages.

In addition, BERT is a robust model that consumes a large amount of system resources on which it will be run. One way to optimize its performance during a training and save memory is the use *PyTorch DataLoader* class. On the other hand, *PyTorch*, is a deep learning project development library whose main feature is the use of GPU acceleration. Thanks to this, *PyTorch* allows a dynamic behavior change of a neural network on the go. Finally, for our BERT model, the *BERT base* configuration was chosen for its trade-off between number of parameters vs. execution time. The base BERT model differs from the *BERT large* model by 2.8% and its runtime increases substantially; for this reason it will not be considered in this study.

Base BERT is configured by 768 hidden layers and 12 headers. However, it´s possible set a number of improvements to the model through AdamW optimizers, including a learning rate between $5e^{-5}$, $3e^{-5}$, $2e^{-5}$.

TABLE III. Several Studie-cases Harassment and Non-Harassment Comments Examples

Harassment free Tweets	Harassment Tweets
si te toca estar en mesa. si sólo tienes que ir a votar y desaparecer.	Todos se cagaron, porque empezaron a decir, "ahí viene la mujer de ese man", nos escuchó gritar "cachudo",
Que viva Guayaquil de mis amores!	jajajaja cerraran las bien puertas y ventanas que pasa el lelo y su mafia.

C. Testing the Trained Models

Once the models are trained, we use different datasets to establish two case studies with which to corroborate their classification accuracy. For this paper, we used two accounts belonging to Ecuadorian users whose contents have different focuses: the first account specializes in hurtful comments (Study Case 1) and the other is focused on political issues (Study Case 2).

Based on this, the operating characteristics are evaluated through accuracy parameters of the trained models.

Thereby, an analysis is made of the phrases or words that, according to the models, suggest some type of harassment or insult.

In addition, the models make it possible determine the percentage of harassment on each analyzed account has in relation to extracted information.

Table III shows one example of a comment with harassment and one without harassment, from which the Study Cases Corpus are composed.

IV. Analysis of Results

To validate our proposed models, we used the Table IV parameters for RNN+LSTM model and Table V parameters for BERT model.

Both models were trained and evaluated on a computer with 12 processing threads at 3.9GHz coupled with 48GB of RAM. In addition, since the models allow the use of vector graphics processing for their execution, a GPU with 1920 CUDA cores and 6GB of VRAM was used.

TABLE IV. RNN+LSTM Model Structuring

Settings	RNN+LSTM
Epochs	6
Embedding neurons	128
Bidirectional Lstm neurons	128, with "batch normalization"
Dropout	128
Batch size	32
Dense layer	1, with "Sigmoid" activation

A. Operating Characteristics Analysis

Fig. 2 shows the curves with operating characteristics (ROC) of RNN+LSTM and BERT models.

The curve corresponding to RNN+LSTM model, in blue, shows an area under the curve (AUC) with a value of 0.97. On the other hand, BERT´s curve model, in red, shows an AUC with a value of 0.98.

TABLE V. BERT Model Structuring

Settings	BERT
Epochs	4
Hidden layers	768
Heads	12
Batch size	32
Parameters	768
Learning rate	$5e^{-5}$
Epsilon value	$1e^{-8}$

Thereby, AUC is the probability that a negative random comment is considered as negative and a positive one as positive. Therefore, the value of the AUC ranges from 0 to 1.

Our implemented models provide optimal performance overall, where BERT model outperforms the RNN+LSTM model by 1%.

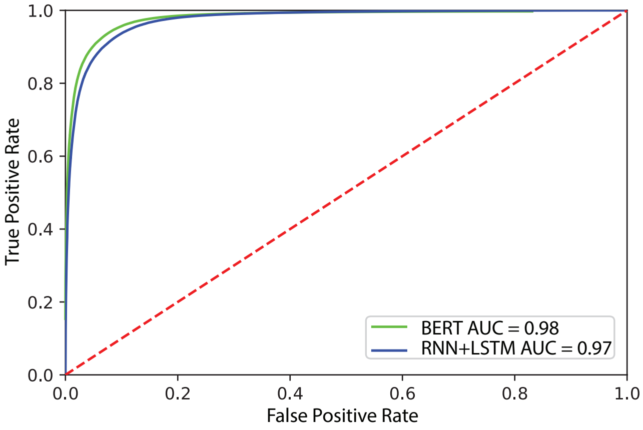

Fig. 2. RNN+LSTM & BERT Model´s Operating Characteristics.

B. Performance Analysis

As shown in Table VI, the BERT model reflects the highest accuracy with respect to its counterpart with RNN+LSTM and even surpasses by far the result obtained with SVM. In [18]. However, the RNN+LSTM model offers very close performance to that of BERT, with only a fraction of its run time.

TABLE VI. Results from Trained Models

Parameter	RNN+LSTM	BERT	SVM	NB	LR
Accuracy [%]	91.82	92.82	87	83	85
Execution time [min]	78	270	-	-	-
AUC	0.97	0.98	-	-	-
Number of Tweets	220000	220000	230000	230000	230000

In terms of Epoch, RNN+LSTM model shows a decreasing trend with respect to the precision value, as shown in Fig. 3. Thus, the higher the number of Epochs the accuracy tends to

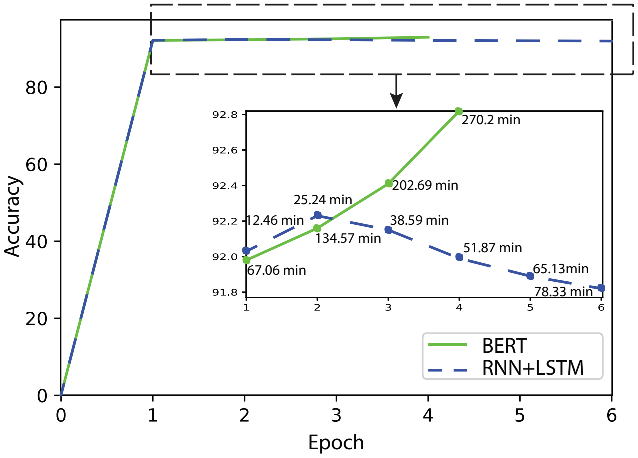

Fig. 3. Execution Time and Accuracy by Epoch.

stabilize at a lower value with respect to its starting point; being the number of Epochs chosen the appropriate one to achieve a stable value close to the one achieved by the BERT model.

On the other hand, BERT model shows an increasing behavior; suggesting that the model can achieve even higher levels of accuracy. However, a higher level of accuracy may represent a model´s overfitting.

At the same time, Fig. 3 evidences a convergence point with an accuracy´s approximate value between the two models at Epoch number 2.

However, this sample shows the difference in the execution time required by each model to achieve this accuracy value; where RNN+LSTM uses only 18% of time used by BERT.

C. Validation Analysis of Acertation

In order to evaluate the trained models, we have used data from two Twitter accounts whose users will be described as @USER_1 (Study Case 1) and @USER_2 (Study Case 2); in order to maintain these accounts anonymous. In turn, the accounts have different content approaches. The analysis are shown in Table VII. At the same time, we use the portion of the dataset intended for Test tasks.

TABLE VII. Study Cases Accounts Cyberbullying Average

Account	Account´s Harassment percentage [%]
@USER_1	86,3
@USER_2	19,27

The model´s results using non-explicitly offensive comments were satisfactory, especially with BERT model. With the dataset test portion, we generate a list of 20 random comments with alleged cyberbullying, BERT was able to identify 85% of them correctly. While, under the same parameters, RNN+LSTM identified 75% of comments correctly.

TABLE VIII. Tweet´s Fragments with Alleged Harassing Content According to the RNN+LSTM Model with the Study Case 1 Dataset

Commentary	Harassment	Real State
Punkeros marihuanos fracasados buenos para nada, y el otro defecto era que cuando ella vivía en Quito se metía hierba, mucha hierba...	✓	✓
A ver te explico. A ver te explico. A ver te explico. A ver te explico. Mira mamaverg*...	✓	✓
Poco a poco esas experiencias de hablar por horas se reducían a minutos, nos volvíamos extraños, como si nunca nos hubiéramos conocido...	✗	✗
Al rato le llega una llamada de cierto hijo de put* preguntándole qué hace y dice: ”nada aquí en el mall, cae”. O SEA BRAAADER, O SEA	✓	✓
Nos fuimos todos a casa y al rato me llama Kar– diciendo que la hermana le dijo que yo le gustaba a ella y se puso a llorar...	✗	✗
’Belleza mis negros de la mini tri, carajo. Put*s para todos!, yo invito.	✓	✗

TABLE IX. Tweet´s Fragments with Alleged Harassing Content According to the BERT Model with the Study case 1 Dataset

Commentary	Harassment	Real State
Hasta Cuando Nos Vengan a Ver- Peker la Maravilla Ft El Soldadito Broder, esa vaina te llega al soul	✗	✗
Yo andaba en bóxer y la gata hija de put* que estaba tranquila acurrucada, de la nada me mete sus garras en mi virilidad masculina varonil. . .	✓	✗
Admito a veces se me chispotea y escribo algo mal, pero braaaders, hay gente que no tiene esta condición, y escriben con el cul*. . .	✓	✓
Todo el mundo me quedó viendo con cara de: ”Vales verg* pendejo, la hiciste llorar, no se hace llorar a una mujer en público”...	✓	✓
Toda grilla es madrina de un equipo pelotero de la universidad o empresa, y las grillas de más caché son candidatas a reinas de lo que sea. . .	✓	✓
Nos llega la noticia sobre la universidad, a ella le dijeron que estudiaría en Cuenca y justo yo estaba tratando de estudiar otra carrera...	✗	✗

In Study Case 1, 20 random comments are evaluated by RNN+LSTM and BERT they were able to predict 85% of comments correctly; part of these results are shown in Table VIII and Table IX respectively.

However, for Study case 2, BERT demonstrated a higher level of prediction than RNN+LSTM. When generating a list of 20 comments with suspected cyberbullying, BERT got 30% of them right, as seen in Table XI; while RNN+LSTM barely managed to identify 10%, from Table X.

This is due to the comments in Study case 2, which have no insults or hurtful comments as such, unlike Study Case 1. Which don´t contain any insults or hurtful comments as such, unlike Study Case 1. This is where the advantages of using a pre-trained model over a relatively inexperienced one become evident.

In addition, upon visual inspection of the classification results, it can be determined that BERT suggests a better approach to harassment, since it does not take any sentence as harassment even if it shows a potential insult. Likewise, this phenomenon can be better appreciated in the comments of Study Case 2, where RNN+LSTM model generates more false-positive cases.

On the other hand, results show a 16% of comments whose terms suggest some kind of insult without actually being harassment.

This is where implemented models tend to show bad results with respect to a real state of interpretation; taking into account that it can be a subjective parameter depending on the user who interprets it.

TABLE X. TWEET´S FRAGMENTS WITH ALLEGED HARASSING CONTENT ACCORDING TO THE RNN+LSTM MODEL WITH THE STUDY CASE 2 DATASET

Commentary	Harassment	Real State
Ministra R— debe estar presa: Es una vulgar delincuente. Terremoto político: Difunden organigrama, elaborado por Dan— M—	✓	✓
Solo por tratarse de Cor— violan las reglas básicas del derecho procesal. Que vergenza!	✓	✓
Los torcidos renglones de este panfleto. Cuánta imaginación malsana!!.Que poca capacidad de análisis político!	✓	✓
PRONUNCIAMIENTO DE LOS EX MIEMBROS DE LA COMISIÓN DE AUDITORÍA DE LA DEUDA Cuestionan la renegociación.	✗	✗
And— Ar— aceptó la candidatura del progresismo: Representamos a todos los sectores del Ecuador.	✗	✗
Caso BOCHORNOS es la suma d todas las arbitrariedades	✓	✓

TABLE XI. TWEET´S FRAGMENTS WITH ALLEGED HARASSING CONTENT ACCORDING TO THE BERT MODEL WITH THE STUDY CASE 2 DATASET

Commentary	Harassment	Real State
La situación en Pichincha es grave. Los esfuerzos para evitar contagios necesitan la responsabilidad de todos los ciudadan	✗	✗
Ningún Asambleísta de está envuelto en sucios y corruptos repartos de hospitales. Eso no sale en la prensa...	✓	✓
Empezaron los recaderos de banqueros! Nos enseña algún sobreprecio? AHORA es que los hay. Ha de ser feo ver a...	✓	✓
Lo mismo se pasaron diciendo los 10 años de mi Gobierno. Recuerdan? La campaña sucia de siempre.	✓	✓
Cuando de pronto empezó a suceder algo raro...	✗	✗
Esta designación lejos de ser únicamente mérito indígena...	✓	✗

V. CONCLUSIONS

In this paper, we continue the work presented in [18], evaluating two models of deep learning in NLP tasks for the detection of cyberbullying in Twitter social network. The proposed models achieve at least a 5% improvement in accuracy over the best model of the previous work corresponding to SVM.

Our first model, RNN+LSTM, shows as the most balanced option between execution time with 78 [min] and an accuracy of 91.82%; which means a difference of 1% compared to 92.82% of BERT´s accuracy, but using 82% more run time.

However, RNN+LSTM doesn´t represent the most efficient cyberbulling classification option, due to BERT has a better criterion when detecting harassment, just like Study case 2, where BERT outperforms RNN+LSTM by 20%. Thanks to the fact that BERT is a pre-trained model and that the analyzed account does not contain explicitly offensive comments, which presents a challenge for both models in predicting harassment.

At the same time, RNN+LSTM model requires 33.33% more Epochs to approach the accuracy value of the BERT model. Nevertheless, RNN+LSTM is shown to be a robust option for presumptive detection tasks of cyberbullying in SSNs. For this reason, it is recommended that the model be redesigned to include a more robust Bidirectional layer with a larger number of neurons.

On the other hand, BERT model offers different model architectures with a greater or lesser number of parameters, depending on the desired implementation approach. Thus, there is potential improvement for accuracy if we evaluate the *bert-large-uncased* or *bert-base-multilingual-uncased* models. However, a different BERT model requires higher computational capabilities, which can be limiting factor for its implementation.

Finally, through the case studies, it was determined that the trained models are robust enough to predict whether an account includes cyberbullying comments in its content and at the same time its percentages.

REFERENCES

[1] E. Hafermalz and K. Riemer, "The question of materiality: Mattering in the network society," 2015.

[2] R. Ortega Ruiz, R. d. Rey Alamillo, and J. A. Casas Bolaños, "Redes sociales y cyberbullying: El proyecto conred," *Convives, 3, 34-44.*, 2013.

[3] G. A. Maldonado Berea, J. García González, and B. E. Sampedro Requena, "El efecto de las tic y redes sociales en estudiantes universitarios," *RIED Rev. Iberoam. Educ. Distancia*, vol. 22, pp. 153–176, 2019.

[4] L. E. Arab and G. A. Díaz, "Impacto de las redes sociales e internet en la adolescencia: aspectos positivos y negativos," *Revista Médica Clínica Las Condes*, vol. 26, no. 1, pp. 7–13, 2015.

[5] M. A. S. Ruiz and C. Inostroza, "Repercusiones sobre la salud del maltrato entre iguales: coso escolar y ciberacoso," *Revista de estudios de juventud*, no. 115, pp. 195–206, 2017.

[6] L. P. Bosque Vega, "Detección automática de ciber acoso en redes sociales." Ph.D. dissertation, Universidad Autónoma de Nuevo León, 2017.

[7] UNICEF *et al.*, "Violencia entre pares en el sistema educativo: Una mirada en profundidad al acoso escolar en el ecuador. unicef. 2017," *Revista virtual][Fecha de acceso: 08 de Agosto del 2019]. En: https://www. unicef. org/ecuador/Press_Kit__AbusoEscolar_Final. pdf.*

[8] Cnna-Cnii, *Agenda Nacional para la Igualdad Intergeneracional 2013-2017 del Ministerio de Inclusión Económica y Social de Ecuador.*, 2014. [Online]. Available: https://www.ohchr.org/Documents/Issues/OlderPersons/MIPAA/Ecuador_Annex1.pdf

[9] L. C. Díaz, "# nobullying: una acción integral contra el acoso escolar," *Revista de Estudios de Juventud*, no. 115, pp. 167–191, 2017.

[10] A. Allisiardi *et al.*, "Bullyng y ciberbullying en argentina: el rol de la comunicación en su prevención. guía con orientaciones para campañas de publicidad social." 2019.

[11] J. Díaz Ospina *et al.*, "Diseño de contenidos digitales para campañas publicitarias de bien social de sensibilización del ciberacoso o cyberbullying en niños de 8 a 12 años estratos 4-6 de manizales," Master's thesis, Escuela de Ciencias Sociales, 2018.

[12] N. Alswaidan and M. E. B. Menai, "A survey of state-of-the-art approaches for emotion recognition in text," *Knowledge and Information Systems*, pp. 1–51, 2020.

[13] N. Joselson and R. Hallén, "Emotion classification with natural language processing (comparing bert and bi-directional lstm models for use with twitter conversations)," 2019.

[14] M. Andriansyah, A. Akbar, A. Ahwan, N. A. Gilani, A. R. Nugraha, R. N. Sari, and R. Senjaya, "Cyberbullying comment classification on indonesian selebgram using support vector machine method," in *2017 Second International Conference on Informatics and Computing (ICIC)*, 2017, pp. 1–5.

[15] M. A. Al-Garadi, M. R. Hussain, N. Khan, G. Murtaza, H. F. Nweke, I. Ali, G. Mujtaba, H. Chiroma, H. A. Khattak, and A. Gani, "Predicting cyberbullying on social media in the big data era using machine learning algorithms: Review of literature and open challenges," *IEEE Access*, vol. 7, pp. 70 701–70 718, 2019.

[16] S. T. T and B. R. Jeetha, "Cyberbullying detection in twtter using language extraction based simplified support vector machine (ssvm) classifier," vol. 6, no. 3, 2017, pp. 21 – 30.

[17] G. A. Prieto Cruz and E. E. Montoya Vasquez, "Modelo de detección de violencia contra la mujer en redes sociales en español, utilizando opinion mining," 2020.

[18] G. A. León-Paredes, W. F. Palomeque-León, P. L. Gallegos-Segovia, P. E. Vintimilla-Tapia, J. F. Bravo-Torres, L. I. Barbosa-Santillán, and M. M. Paredes-Pinos, "Presumptive detection of cyberbullying on twitter through natural language processing and machine learning in the spanish language," in *2019 IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON)*, 2019, pp. 1–7.

[19] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1253, 2018.

[20] S. Paul and S. Saha, "Cyberbert: Bert for cyberbullying identification," *Multimedia Systems*, pp. 1–8, 2020.

[21] H. Sadr, M. M. Pedram, and M. Teshnehlab, "A robust sentiment analysis method based on sequential combination of convolutional and recursive neural networks," *Neural Processing Letters*, vol. 50, no. 3, pp. 2745–2761, 2019.

[22] G. A. León-Paredes, L. I. Barbosa-Santillán, and J. J. Sánchez-Escobar, "A Heterogeneous System Based on Latent Semantic Analysis Using GPU and Multi-CPU," *Scientific Programming*, vol. 2017, p. 19, 2017.

[23] G. A. León-Paredes, L. I. Barbosa-Santillán, J. J. Sánchez-Escobar, and A. Pareja-Lora, "Ship-sibiscas: A first step towards the identification of potential maritime law infringements by means of lsa-based image," *Scientific Programming*, vol. 2019, 2019.

[24] J. Chen, S. Yan, and K.-C. Wong, "Verbal aggression detection on twitter comments: Convolutional neural network for short-text sentiment analysis," *Neural Computing and Applications*, pp. 1–10, 2018.

[25] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *arXiv preprint arXiv:1310.4546*, 2013.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018. [Online]. Available: https://arxiv.org/abs/1810.04805

[27] J. Yadav, D. Kumar, and D. Chauhan, "Cyberbullying detection using pre-trained bert model," in *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*. IEEE, 2020, pp. 1096–1100.

[28] A. Huertas Mora *et al.*, "Algoritmos de aprendizaje supervisado utilizando datos de monitoreo de condiciones: Un estudio para el pronóstico de fallas en máquinas."

[29] M. Vallez and R. Pedraza, "El procesamiento del lenguaje natural en la recuperación de información textual y áreas afines," *Hipertext. net*, 2007.

[30] H. Saif, M. Fernandez, and H. Alani, "On stopwords, filtering and data sparsity for sentiment analysis of twitter," *Proceedings of the 9th International Language Resources and Evaluation Conference (LREC'14)*, pp. 810–817, 01 2014.