ISBN: 978-628-95207-4-3. ISSN: 2414-6390. Digital Object Identifier: 10.18687/LACCEI2023.1.1.261

Identification of a suitable NLP model for the detection of symptoms mentioned in textual conversations of Covid-19 infected persons.

Ivan Acosta-Guzmán, MSIG.[1], Eleanor Varela-Tapia, MSIG.[1], Lady Mariuxi Sangacha Tapia, MSIA.[2], Mirian Estefanía Solórzano-Monserrate, Ing.[1], Christopher Acosta-Varela, estudiante[3]

[1]Universidad de Guayaquil, Ecuador, ivan.acostag@ug.edu.ec, eleanor.varelat@ug.edu.ec, mirian.solorzanom@ug.edu.ec

[2]Instituto Superior Universitario Tecnológico del Azuay, Parque industrial, Ecuador. lady.sangacha@tecazuay.edu.ec

[3]Escuela Superior Politécnica del Litoral, Ecuador, chriacos@espol.edu.ec

Abstract – In March 2020, the World Health Organization (WHO) declared Covid-19 disease as a global pan-demic. Therefore, the need for reliable information arose, so several Virtual Health Assistants emerged to provide information to the public to teach the population how to prevent or cope with the Covid-19 Alpha variant infection, but progressively emerged the Beta, Delta, Omicron variants with different symptomatology which triggered new waves of infections and deaths in the world. For this reason, the present study promoted the creation of a NLP (Natural Language Processing) model to analyze the experiences of infected people in Guayaquil and to detect the predominant symptoms mentioned in their textual conversations. For this purpose, the Quantitative Methodology was followed by performing surveys through Google form, reaching 2873 people in the city of Guayaquil who had and overcame the Covid-19. Thus, the corpus of textual conversations was generated. On the other hand, the Qualitative Methodology was used by executing interviews to NLP experts, which allowed corroborating the classifiers suggested in this type of software. Finally, twenty-two different NLP models were built based on classifier algorithms like K-Nearest Neighbors, Random Forest, and Long-Short Term Memory (LSTM), and their quality was evaluated using metrics such as Accuracy, Precision, Recall, and AUC, finding that the model based on LSTM version 3 obtained the highest performance.

Keywords: Covid-19, NLP for Classification, Random Forest, Dense Neural Network, LSTM, Metrics.

I. INTRODUCTION

Due to the Covid-19 disease, the world has experienced an unexpected situation, since the Chinese health authorities notified the international community of the appearance of this disease in Wuhan on December 31, 2019, humanity recognized the existence of the anomaly, countries in their attempt to prevent the accelerated spread of this virus declared states of emergency. World Health Organization (WHO) declared Covid-19 as an epidemic on January 30, 2020, while on March 11, 2020, it was announced as a global pandemic. [1].

Digital Object Identifier:

ISBN: **ISSN:**

In Ecuador, through Ministerial Agreement No. 00126-2020 issued the same day by the Minister of Health, a State of Health Emergency was declared in the National Health System [2]. Guayaquil was the city most affected by the virus, accounting for 70% of deaths in Ecuador, while in other parts of the country the intensity of the epidemic was lower in the first months of the epidemic [3]. The measures taken by the Ecuadorian authorities were to apply a total containment [4], which generated a high economic and social impact [5], by June 30, 2020, the country had experienced the 7th highest number of cases, the 6th highest number of deaths in the region, and the highest fatality ratios [6] [7].

Due to the appearance of the Covid-19 disease, Ecuador was the country with the highest number of deaths in the region during March and April of that year, with an accelerated growth of new serious cases, leading to the collapse of the Ecuadorian health system, during the so-called "first wave", Brazil and Ecuador were the most affected countries in the region and the world [8].

The suspension of the right to freedom of movement, the right to freedom of association, restrictions on travel time, limited access to medical care for other diseases, and restrictions on access to food suppliers, were the strategy of authorities to avoid the increment of infections [4].

One of the technologies used in this type of emergency to obtain efficient monitoring of the Covid-19 virus was Artificial Intelligence (AI), helping to collect, analyze and find trends in

the behavior of the virus. For its part, the AI subbranch called Natural Language Processing (NLP) has been used in sentiment analysis, document classification, spam detection among other uses, resulting in NLP being very helpful in new research fields.

This paper is organized as follows:

Section II shows the works before this research are mentioned.
Section III describes the research and develop methodology.
Section IV detail of the phases of development.
Section V is about the validation metrics.
Section VI. describes the Discussion or Results.
Section VII contains the conclusions of this research.
Section VII presents references of this research.

II. Related literature

A paper conducted in 2021 related to health behavior theories (health belief model, social norm, and trust) employed machine learning to examine people's behaviors towards Covid-19. It used the comments about Covid-19 on Twitter and used candidate key phrases representing each health behavior construct to associate a label to the comments.

Then, that work has developed three models with Support Vector Machine (SVM), Decision Tree (DT) and Logistic Regression (LR) classifiers, finding that DT and SVM yielded an overall F1 score of 98% for the multi-class (single-label) classification, while DT outperforms the other classifiers with overall F1 score of up to 100% for the multi-class-multi-label classification [9].

Another study developed a solution in the field of digital health by taking information from popular DPP forums: What to Expect and BabyCenter. A machine learning model was developed to automatically classify the content of the posts. The results revealed the main topics discussed by users of these online forums: family and friends, medications, symptom disclosure, breastfeeding, and social support in the peripartum period. The results indicate that the Random Forest model outperformed the Support Vector Machine and Logistic Regression models. [10]

A third work developed a solution in a field close to the goal of this research, in that work the researchers promoted CovBERT algorithm and compare it results with another's BERT variants used in previous works, finding that CovBERT Algorithm, it reaches better values in Loss 18%,and I in metrics Precision 86%, Recall 88%, F1 86%, and AUC 86%.[11]

The present project has as general objective the identification of the most efficient NLP model for the identification of predominant symptoms presents in textual conversations of people infected by Covid-19. For this purpose, we proceeded with the review of the state of the art of NLP multi-label classifiers, symptoms presented by variants of Covid-19, generation of the Corpus of dialogues obtained from people from Guayaquil city infected by Covid-19, creation of several NLP models applying several NLP classifiers, verifying the quality of the models using quality metrics for NLP classification models.

For Random Forest Algorithm in a previous job [12] suggest using a combination of hyperparameters with the values:

```
ramdomforest_best_params =
{'bootstrap': True,
 'max_depth': 70,
 'max_features': 'auto',
 'min_samples_leaf': 4,
 'min_samples_split': 10,
 'n_estimators': 400}
```

Other job about Random Forest Algorithm mention that the best combination founded [13] was:

```
ramdomforest_best_params2 =
{'bootstrap': True,
 'criterion': 'entropy',
 'max_ depth': 12,
 'max_features': 'log2',
 'min_samples_leaf': 1,
 'min_samples_split': 5,
 'n_estimators': 90}
```

These options were tested in this work with another additional combinations of hyperparameters.

III. Methodology

A. *Research methodology*

In this work were applied the following methodologies: Documentary methodology based on the revision of scientific articles [14], Qualitative methodology to create instruments to conduct interviews with professionals in Artificial Intelligence to identify the suggested models for multi-label classification. Also, physicians were interviewed to identify their openness to the use of Artificial Intelligence technologies as a tool to address the highly changing symptomatology due to the emergence of Covid-19 variants. Finally, Quantitative methodology to analyze and to find possible correlations between the characteristics of the population studied [15]. In this research, an online survey was conducted to collect data of the habitants of Guayaquil city that was diagnostic as infected with Covid-19 between March-2020 to December-2022.

B. *Population and Sample*

According to the National Institute of Statistics and Census of Ecuador, for the year 2017 Guayaquil had 2,644,891 inhabitants [16].

Applying probabilistic method "Simple Random Sampling", with parameters Confidence Level at 95% corresponding to Z = 1.96, p = 50%, q = 50%, and margin error E = 1,831% it results on a sample size of 2,865 habitants.

$$n = \frac{Z^2 * p * (1-p)}{E^2} \quad (1)$$

$$n = 2{,}865\ habitants \quad (2)$$

C. *Interviews*

With the purpose of acquiring the criteria of each expert during the data collection stage, two interviews were conducted with specialists in Artificial Intelligence, the Eng. Karla Avilés Mendoza Machine Learning Researcher, and MS. Nayelhi de Anda mentioned that an appropriate NLP algorithm is the use of Neural Network when there is a high volume of data available for training the model.

D. *Surveys*

A survey consisting of 11 questions was developed, the survey was created using a form in Google Forms. A team of 160 pollsters from University of Guayaquil was promoted the execution of the survey between Dec-9-2021 to Feb-21-2022 through digital media making it reach 2,873 people who had Covid-19 between 2020 to 2021 in the canton of Guayaquil, 10 closed questions and one open question regarding symptoms suffered.

To validate the survey instrument, the Delphi Methodology was used [17] through the collaboration of NLP expert Eng. Nestor Montaño, an expert in Statistics and Artificial Intelligence.

A pilot of 500 surveys was done through a team of research assistants. After that, with recommendations Eng. Néstor Montaño, the pilot questions were improved, and was developed the surveys of 2873 people required for this project in Spanish language.

Table 1 shows the list of questions used in the research, with one example of the answers received:

TABLE 1.

QUESTIONS ASKED TO GUAYAQUIL RESIDENTS WHO HAD COVID-19 BETWEEN MAR-2020 TO DEC-2021.

Survey Questions	**Respondent 1**
1. City of residence?	Guayaquil
2. Have you had coronavirus?	Yes
3. Select your age	26 to 64 years old
4. Gender	M
5. Which variant of the virus infected you?	Alfa
6. Date of infection?	mar-30-2020
7. Symptom intensity?	Strong
8. Place where you were infected?	Workplace
9. How many doses did you have prior to becoming infected?	0
10. What vaccine did you receive, prior to becoming infected?	AstraZeneca
11. What symptoms did you suffer?	Shortness of breath, loss of taste, extreme tiredness, and body pain

E. *Corpus.*

A database was generated through the surveys to the inhabitants of the city of Guayaquil, storing the responses in the file CORPUS_SINTOMAS_GYE_COVID19_DE_2020_A_2021.xlsx which was uploaded to Google Colab for use in the following phases of the project.

F. *Data Exploration*

According to an analysis of the statistical distribution there are a balanced participation of men and women (fig. 1), majority participation of people between 18 and 64 years of age (fig. 2), and the 97.91% of people who responded the survey mentioned that suffered between 1 to 6 symptoms (fig. 3).

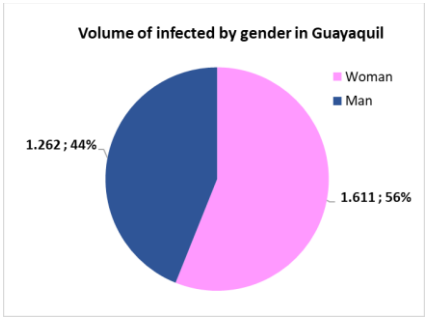

Fig. 1. Respondents by gender in the city of Guayaquil who had Covid-19 between March 2020 and December 2021 overcame it.

Among the most predominant symptoms reported were Fever, Headache, Loss of Smell, Loss of Taste, General Malaise, Cough, Difficulty Breathing, Fatigue, Muscle Pain, Sore Throat, and Influenza.

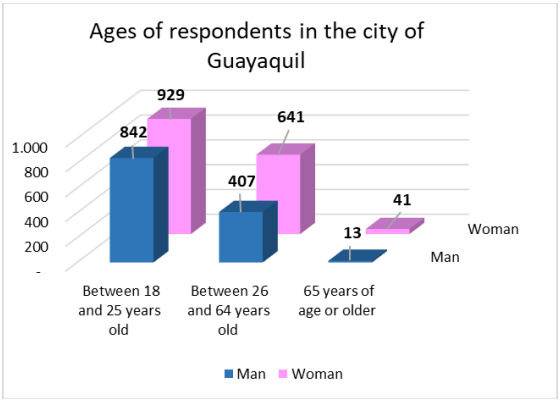

Fig. 2. Respondents by age and gender in the city of Guayaquil who had Covid-19 between March 2020 and December 2021 overcame it.

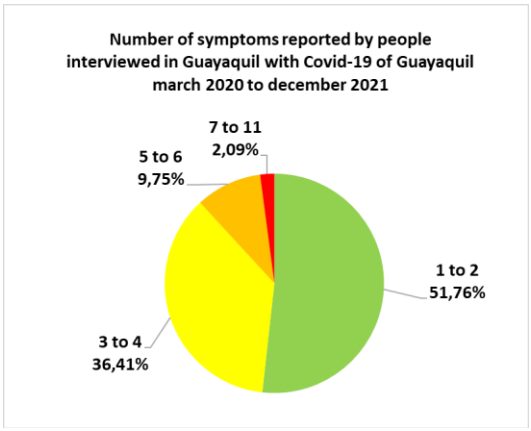

Fig. 3. Number of symptoms reported by people of Guayaquil who had Covid-19 between March 2020 and December 2020 and overcame it.

G. *Development Environment.*

To create the NLP modules for classification, the Python language was used, since it is a language that provides many libraries created for the processing and creation of Data Science and Artificial Intelligence software solutions.

The Google Collaboratory platform was chosen for this project because it is ease of use and let use a station with processing capabilities required to construction NLP solution.

H. *NLP Model*

For the NLP model building process, the following phases were applied:

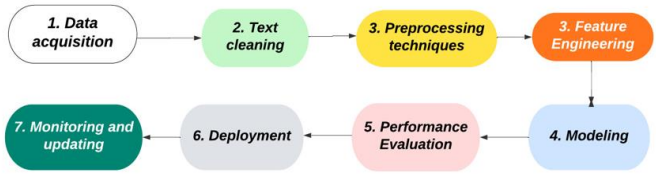

Fig. 4. Natural Language Processing (NLP) Model life cycle). [18] [19]

I. *Machine Learning Classification Algorithms*

NLP models were created for Multi-Label Classification based on Multi-Label Classification algorithms:

- K-Nearest Neighbors (KNN)
- Random Forest (RF)
- Long Short-Term Memory (LSTM),

A comparison of quality metrics was made to identify the most appropriate model for this study.

IV. IMPLEMENTATION

A. *Data Acquisition*

Symptom data were imported from the computer into the Google Colab work environment using Python pandas' libraries. The loaded data was stored in a Dataframe object named df_Síntomas.

B. *Text Cleaning and Preprocessing Techniques*

Was verified if there were empty or missing data for their respective treatment. The text was preprocessed to normalize the data by eliminating special characters, accents, and changing from uppercase to lowercase.

	11. ¿Qué síntomas tuvo?	Síntomas-Normalizado
0	FALTA DE AIRE, PERDIDA DEL GUSTO, CANSANCIO EX...	falta de aire perdida del gusto cansancio extr...
1	Falta de aire, fiebre alta, mucho dolor de cuerpo	falta de aire fiebre alta mucho dolor de cuerpo
2	Perdida del gusto y el olfato y tos por las no...	perdida del gusto y el olfato y tos por las no...
3	Problemas de respirar y malestar	problemas de respirar y malestar
4	Dolor de cabeza, malestar, tos, dolor de garga...	dolor de cabeza malestar tos dolor de garganta
5	Diarrea, dolor de cabeza, fiebre, tos, gripe, ...	diarrea dolor de cabeza fiebre tos gripe dolor...

Fig. 5. Comparison of sample of Spanish corpus with initial text and normalized text.

The information of the Dataframe df_Síntomas was separated into input variable X and output variables Ys. As output variables Ys, those columns corresponding to symptoms with 2% or more of occurrence in the data collected in the

dataset were established to reduce the level of slippage in the data. The variables chosen for the output Ys are:

list_col = ['Headache', 'Loss of Smell', 'Loss of Taste', 'General Malaise', 'Fever', 'Cough', 'Shortness of Breath', 'Fatigue', 'Muscle Pain', 'Asymptomatism', 'Sore Throat', 'Flu']

```
# Columna de Sintomas-Normalizado del dataset
df_X_inicial =df_Síntomas.iloc[ : ,  [12, 72]]

# Columna de Ys con Etiquetas de Sintomas
df_y_inicial  = df_Síntomas[lista_col]
```

Fig. 6. Division of data into input X and output Ys.

C. *Feature Engineering*

The StopWords technique was applied to the 'Symptoms-Normalized' column appling nltk.corpus and Spanish language.

```
from nltk.corpus import stopwords
txt_stop_words = set(stopwords.words("spanish"))
txt_stop_words
```

```
{'a',
 'al',
 'algo',
 'algunas',
 'algunos',
 'ante',
 [..]
 'vuestro',
 'vuestros',
 'y',
 'ya',
 'yo',
 'él',
 'éramos'}
```

Fig. 7. Extract of corpus of stop words available in NLK library for Spanish language.

Síntomas-Normalizado	Sintomas-SinStopWords
falta de aire perdida del gusto cansancio extr...	falta aire perdida gusto cansancio extremo dol...
falta de aire fiebre alta mucho dolor de cuerpo	falta aire fiebre alta dolor cuerpo
perdida del gusto y el olfato y tos por las no...	perdida gusto olfato tos noches
problemas de respirar y malestar	problemas respirar malestar
dolor de cabeza malestar tos dolor de garganta	dolor cabeza malestar tos dolor garganta
diarrea dolor de cabeza fiebre tos gripe dolor...	diarrea dolor cabeza fiebre tos gripe dolor hu...

Fig. 8. Tokenizer process applied to column Symptoms Normalized.

Using percentiles, was detected that the number of words used by 99% of the respondents was up to 14 words, so this was established as the parameter to be handled at the input of the models, for which the command was used:

```
df_X_sst['Sintomas-SinStopWords'].apply(lambda x: len(x.split(" "))).describe( percentiles= [0.25, 0.5, 0.75, 0.85, 0.99])
```

```
count    2811.000000
mean        4.379580
std         2.871506
min         1.000000
25%         2.000000
50%         4.000000
75%         6.000000
99%        14.000000
max        30.000000
```

Fig. 9. Identification of the number of words used by the respondents when expressing the symptoms, they had by Covid-19.

```
import nltk
nltk.download("punkt")
from nltk.tokenize import word_tokenize
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

```
df_X_sst['Sintomas-Tokenizado']=df_X_sst.apply(lambda row: nltk.word_tokenize(row['
df_X_sst.head(4)
```

	Sintomas-SinStopWords	Sintomas-Tokenizado
0	falta aire perdida gusto cansancio extremo dol...	[falta, aire, perdida, gusto, cansancio, extre...
1	falta aire fiebre alta dolor cuerpo	[falta, aire, fiebre, alta, dolor, cuerpo]
2	perdida gusto olfato tos noches	[perdida, gusto, olfato, tos, noches]
3	problemas respirar malestar	[problemas, respirar, malestar]

Fig. 10. Tokenizer process applied to column Symptoms Normalized.

For those sentences that exceed the number of typed words of 14, the excess words will be truncated, an action required in the neural network models for its future operation.

The next step was to apply the tokenization technique to convert each word into a number assigned by the tokenizer.

D. *NLP Modeling*

The following partitions were applied to the dataset: 80% of the dataset data was used for training, 20% for testing, and for internal validation of the model 20% of the training was chosen.

TABLE 2
DATA PARTITIONS USED IN THE NLP MODELING PROCESS.

Fase	Partition	Sub-fase	Partition	Dataset Fraction
Train	80%	Train	80%	64%
		Validation	20%	16%
Test	20%	Test	30%	20%

V. MODELS EVALUATION

According to the data exploration it determines that many symptoms less than 200 responds, because that to avoid excess

of unbalances data, there was stablished to maintain the symptoms that has at least 200 responds, obtaining 12 symptoms for outputs with enough processable data for this project.

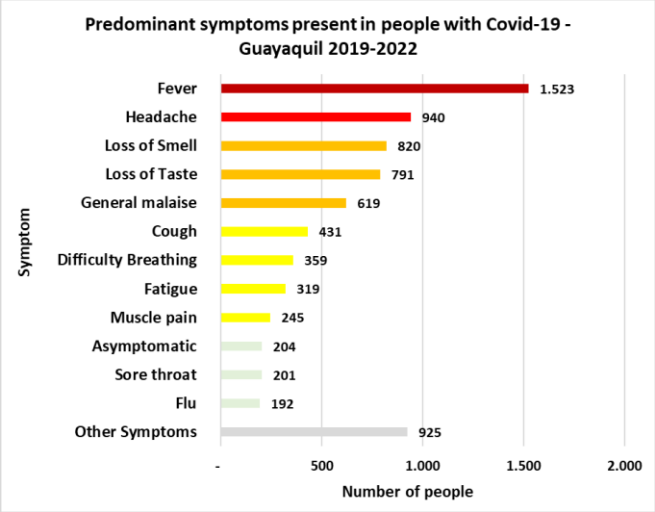

Fig. 11. List of twelve symptoms most mentioned by Guayaquil inhabitants with Covid-19.

Researchers of previous works [9][20] [21b], agree that for the adequate evaluation of the quality of classification algorithms with multiple outputs or with unbalanced data, the use of accuracy is not enough; instead, a set of metrics should be used, including Accuracy, but also the metrics of Precision, Recall, F1, AUC should be added. The various models NLP were executed with 14 inputs and 12 outputs.

ML Algorithm	Version	Metric	K-Neighbors
kNN	1	minkowski	5.0
kNN	2	minkowski	7.0
kNN	3	minkowski	10.0
kNN	4	minkowski	15.0
kNN	5	minkowski	14.0

Fig. 12. Hyperparameters applied, in NLP models with K-Nearest Neighbor Classification algorithm.

In this project there were used ten variants of LSTM, seven variants of Random Forest and five variants of K-Nearest Neighbors, each of one has different hyperparameters that can be adjusted to test the quality of classification results.

In LSTM model there was adjusted parameters like epoch, validation_split and batch_size, at the Random Forest there were ajusted hyperparametes like bootstrap, criterion, n_estimators, max_depth, min_samples_split, max_features, min_samples_leaf, class_weight and oob_score, finally K-Nearest Neighbors were ajusted hyperparametes like n_neighbors, metric and p.

In the case of the NLP models based on the Random Forests classification algorithm, seven parameter combinations were used to find the best results. In these tests, versions 6 and version 7 were considered, using the recommended combinations as the best results. According to two previous works [12] [13] related to the optimization of hyperparameters of the Random Forest Algorithm, version 1 yielded the highest values in metrics of F1(0.7674), Accuracy (0.8795), Recall (0.6806) and AUC. (0.8292.

Version	N_Estim	Max_Feat	Bootstrap	Criterion	Max_Depth1	Min_Samp_Split	Min_Samp_Leaf	class_Weight	OOB_Score
1	200.0	sqrt	True	entropy	None	2.0	1.0	None	True
2	300.0	sqrt	True	entropy	None	2.0	1.0	al_sub	True
3	400.0	sqrt	True	entropy	None	2.0	1.0	al_sub	True
4	100.0	log2	True	entropy	None	2.0	1.0	al_sub	True
5	300.0	log2	True	entropy	None	2.0	1.0	al_sub	True
6	400.0	sqrt	True	gini	70	10.0	4.0	None	False
7	90.0	log2	True	entropy	12	5.0	1.0	None	False

Fig. 13. Hyperparameters applied, in NLP models with Random Forest Classification algorithm.

In the case of the NLP models based on the Random Forests classification algorithm, seven parameter combinations were used to find the best results. In these tests, versions 6 and version 7 were considered, using the recommended combinations as the best results. According to two previous works [12] [13] related to the optimization of hyperparameters of the Random Forest Algorithm, version 1 yielded the highest values in metrics of F1(0.7674), Accuracy (0.8795), Recall (0.6806) and AUC. (0.8292).

ML Algorithm	Version	Bach_size	Epoch	Validation_split
LSTM	1	6.0	100.0	0.2
LSTM	2	12.0	50.0	0.2
LSTM	3	8.0	70.0	0.2
LSTM	4	8.0	50.0	0.3
LSTM	5	12.0	60.0	0.3
LSTM	6	12.0	50.0	0.3
LSTM	7	8.0	100.0	0.2
LSTM	8	6.0	50.0	0.2
LSTM	9	12.0	50.0	0.2
LSTM	10	6.0	50.0	0.3

Fig. 14. Hyperparameters applied in NLP models with LSTM Classification algorithm.

ML Algorithm	Version	Accuracy	F1	Precision	Recall	AUC
Random Forest	1	0.504488	0.767411	0.879518	0.680653	0.829209
Random Forest	2	0.481149	0.761448	0.867793	0.678322	0.826839
Random Forest	3	0.482944	0.762238	0.871129	0.677545	0.826821
Random Forest	4	0.484740	0.763493	0.877016	0.675991	0.826693
Random Forest	5	0.481149	0.761448	0.867793	0.678322	0.826839
Random Forest	6	0.423698	0.728102	0.881768	0.620047	0.800110
Random Forest	7	0.463196	0.745748	0.879620	0.647242	0.813059

Fig. 15. Evaluation Metrics to NLP models with Random Forest Classification algorithm

According to the results obtained related to the optimization of hyperparameters of the K-Nearest Neighbors Algorithm, the version 1 the metrics reach heist values of this team, with the following values Accuracy (0.3393), F1(07503), Precision (0.8698), Recall (0.6596) and AUC (0.8180), but those values were lower than the others reached by Random Forest variants.

		Score				
ML Algorithm	Version	Accuracy	F1	Precision	Recall	AUC
kNN	1	0.339318	0.750331	0.869877	0.659674	0.818071
kNN	2	0.314183	0.750331	0.869877	0.659674	0.818071
kNN	3	0.308797	0.750331	0.869877	0.659674	0.818071
kNN	4	0.303411	0.750331	0.869877	0.659674	0.818071
kNN	5	0.303411	0.750331	0.869877	0.659674	0.818071

Fig. 16. Evaluation Metrics to NLP models with K-Nearest Neighbor Classification algorithm.

		Metric values with test data					
Ver.	Algorithm	Accuracy	Loss	Precision	Recall	f1	AUC
1	LSTM	0.5673	0.1406	0.9156	0.9106	0.9131	0.9778
2	LSTM	0.6355	0.1313	0.9112	0.8695	0.8899	0.9766
3	LSTM	0.7343	0.1185	0.9218	0.9068	0.9142	0.9793
4	LSTM	0.6715	0.1389	0.9056	0.8500	0.8770	0.9736
6	LSTM	0.4650	0.3028	0.7950	0.4942	0.6095	0.8966
7	LSTM	0.5583	0.2322	0.8412	0.6255	0.7175	0.9387
8	LSTM	0.3501	0.2421	0.8558	0.5765	0.6890	0.9278
9	LSTM	0.3232	0.2855	0.8595	0.1616	0.2721	0.8883
10	LSTM	0.5135	0.2842	0.8598	0.1810	0.2991	0.8943

Fig. 17. Evaluation Metrics from NLP models based on LSTM Classification algorithm.

VI. Discussion of Results

After comparing ten variants of NLP multi-label classifier the Model based in LSTM Classification Algorithm, the combination of hyperparameters with the best results was the model3 that included activation of sigmoid function, optimizer Adam, loss binary_crossentropy function, batch_size in 8, epochs 70, and validation split 0.2. It shown the best quality metrics in the twenty-two models tested for training without overfitting and test, the values of metrics reached for testing were Precision 92,18%, Recall 90,68%, F1 91.42% and AUC 97,93%.

TABLE 3
METRICS ACHIEVED IN THE CURRENT WORK VERSUS PREVIOUS NLP WORK FOR MULTI-LABEL CLASSIFIER.

Research	Classifier	Metrics in the test data				
		Loss	Precision	Recall	F1	AUC
Metrics of present work	LSTM (N.3)	12%	92.2%	90.7%	91.4%	97.9%
	R.F. (N.1)	-	88.0%	68.1%	76.7%	82.9%
	KNN (N.1)	-	87.0%	66.0%	75.0%	81.8%
Metrics of previous works [11]	roberta-base	29%	68%	51%	57%	57%
	albert-base-v1	39%	56%	41%	47%	47%
	emilyalsentzer/ Bio_ClinicalBERT	25%	82%	83%	82%	82%
	CovBERT	18%	86%	88%	86%	86%

To finalize, the results of the present work were compared with results of previous works related to text classification about Covid-19. Then was found that the values of metrics of LSTM model version 3 exceeded previous works results. (Table 3)

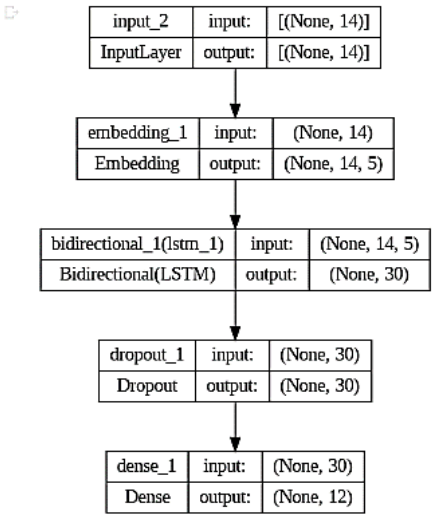

Fig. 18. NLP model based on LSTM version 3 classifier algorithm.

VII. Conclusions

Analyzing the metrics of accuracy, recall, precision and f1, obtained in NLP multi-label models with LSTM, Random Forest and KNN algorithms, these showed that KNN got the lowest values of classification quality, Random Forest got the intermedia quality and LSTM reached the highest classification quality in training and testing processes.

The general objective proposed in this work was achieved, which consisted of identifying the most efficient NLP model in text processing related to conversations of Covid19 symptoms, finding that the model with the LSTM classifier obtains the highest results among the models tested.

On the other hand, it was identified that the ten symptoms with the highest incidence in people infected by Covid-19 were Back pain, Dizziness, Joint pain, Bone pain, Diarrhea, Loss of appetite, Flu, Sore throat, asymptomatic, and muscle pain.

In future works on NLP models, other classification algorithms such as Support Vector Machine (SVC), Multinomial Logistic Regression (MLR) could be applied to complement the present work and improve the results obtained with the models already tested.

VI. Acknowledgment

I express my gratitude to Facultad de Ingeniería Industrial y Facultad de Ciencias Matemáticas y Físicas of Universidad de Guayaquil, for supporting this research in the field of Artificial Intelligence and Natural Processing Language.

VIII. References

[1] Ducharme, J. "The WHO Just Declared Coronavirus COVID-19 a Pandemic | Time". 2020. https://time.com/5791661/who-coronavirus-pandemic-declaration/

[2] COE Nacional, "Informe-de-Situación-No008-Casos-Coronavirus-Ecuador-16032020-20h00", Accessed: Sep. 12, 2022. https://www.gestionderiesgos.gob.ec/?s=Informe+de+Situaci%C3%B3n+COVID19+#search

[3] S. Labarthe. "¿Qué pasa en Ecuador? Covid-19, crisis sanitaria y conflictividad política | Nueva Sociedad.". 2020. https://nuso.org/articulo/que-pasa-en-ecuador/.

[4] Coronel & Perez. "La crisis ocasionada por el Covid-19 y sus implicaciones legales, en el Ecuador". 2020. https://www.coronelyperez.com/2020/04/23/la-crisis-ocasionada-por-el-covid-19-y-sus-implicaciones-legales-en-el-ecuador/.

[5] G. K., Cevallos, et al. "Impacto social causado por la COVID-19 en Ecuador". 3C Empresa. Investigación y pensamiento crítico. Edición Especial COVID-19: Empresa, China y Geopolítica, 115-127. (2020). https://doi.org/10.17993/3cemp.2020.edicionespecial1.115-127

[6] Organizacion Mundial de la Salud. "Coronavirus disease (COVID-19): situation report, 172.". 2020. https://apps.who.int/iris/handle/10665/333297

[7] El Comercio. "La letalidad por el nuevo coronavirus de Ecuador es la más alta de Sudamérica". 2020. https://www.elcomercio.com/actualidad/peru-chile-brasil-covid-19.html.

[8] K. H. Briones-Claudett, R.A. Murillo Vásconez, CDR Rivera Salas, et al. Management of COVID-19 at the pandemic's first wave in Ecuador. SAGE Open Med Case Rep. 2021;9:2050313X211045232. 2021. https://doi.org/10.1177/2050313X211045232

[9] B. Graham-Kalio, O. Oyebode, N. Zincir-Heywood, and R. Orji, "Analyzing COVID-19 Tweets using Health Behaviour Theories and Machine Learning," SeGAH 2021 - 2021 IEEE 9th Inter-national Conference on Serious Games and Applications for Health, Aug. 2021, doi: https://doi.org/10.1109/SEGAH52098.2021.9551908

[10] A. Zingg, T. Singh, and S. Myneni, "Analysis of Online Peripartum Depression Communities: Application of Multilabel Text Classification Techniques to Inform Digitally-Mediated Preven-tion and Management," Front Digit Health, vol. 3, 2021, doi: https://doi.org/10.3389/fdgth.2021.653769

[11] M. Khadhraoui, H. Bellaaj, M. Ammar, H. Hamam, M. Jmaiel. Survey of BERT-Base Models for Scientific Text Classification: COVID-19 Case Study. *Appl. Sci.* 2022, *12*, 2891. https://doi.org/10.3390/app12062891

[12] W. Koehrsen, Hyperparameter Tuning the Random Forest in Python. 2018. https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74

[13] Chauhan, A.(2021). Random Forest Classifier and its Hyperparameters. https://medium.com/analytics-vidhya/random-forest-classifier-and-its-hyperparameters-8467bec755f6

[14] J. C. O. Alvarado and A. A. D. Pérez, "¿Cómo redactar los antecedentes de una investigación cualitativa?," Revista Electrónica de Conocimientos, Saberes y Prácticas, vol. 1, no. 2, pp. 66–82, Oct. 2018, doi: https://doi.org/10.30698/recsp.v1i2.13

[15] J. Cárdenas, "Investigación cuantitativa," 2018, doi: 10.17169/REFUBIUM-216. https://www.researchgate.net/publication/337826972

[16] INEC. "Guayaquil en cifras". 2017. https://www.ecuadorencifras.gob.ec/guayaquil-en-cifras/

[17] M. E. García-Ruiz and F. J. Lena-Acebo, "Application of the delphi method in the design of a quantitative investigation on the FABLABS," Empiria, no. 40, pp. 129–166, May 2018, doi: https://doi.org/10.5944/empiria.40.2018.22014

[18] P. A. Muñoz Cáceres, "Diseño y construcción de modelo de clasificación de incidentes de se-guridad usando NLP en los registros de texto escrito para automatizar etiquetación," 2020, Ac-cessed: Sep. 15, 2022. https://repositorio.uchile.cl/handle/2250/176979

[19] InterviewBit, "Introduction to Natural Language Processing (NLP)," 2021. https://www.interviewbit.com/nlp-interview-questions/

[20] Li, R. et al. Multilevel Risk Prediction of Cardiovascular Disease based on Adaboost+RF Ensemble Learning. In IOP Conference Series: Materials Science and Engineering (Vol. 533). Institute of Physics Publishing. 2019. https://doi.org/10.1088/1757-899X/533/1/012050

[21] G . Nikhila and A.C. Meghashree. "Machine Learning Framework to Predict Chronic Kidney Disease Using Ensemble Algorithm." International Journal of Engineering and Advanced Technology 9 (5). Blue Eyes Intelligence Engineering and Sciences Engineering and Sciences Publication - BEIESP: 1–6. 2020. doi: https://doi.org/10.5585/gep.v12i2.18942