Initial Progress of Identification of the Appropriate NLP Technique for Content Evaluation in Textual Conversations of People Infected by Sars-Cov-2

Iván L. Acosta-Guzmán[a], Eleanor A. Varela-Tapia [b], Alexandra E. Piza-Guale[a], Nory X. Acosta-Guzmán [c], Christopher I. Acosta-Varela[d]

[a] Facultad de Ingeniería Industrial, Universidad de Guayaquil. Guayaquil, Ecuador, 090112
[b] Facultad de Ciencias Matemáticas y Físicas. Universidad de Guayaquil. Guayaquil, Ecuador, 090112
[c] Facultad de Educación, Universidad Estatal de Milagro. Milagro, Ecuador, 091050
[d] Facultad de Ingeniería en Electricidad y Computación, Escuela Superior Politécnica del Litoral, Guayaquil-Ecuador, 090903.
Corresponding author: *ivan.acostag@ug.edu.ec*

Vol. 02, Issue 03 (2023): December
DOI: https://doi.org/.
10.53591/easi.v2i3.2488
ISSN: 2953-6634
Submitted: October 23, 2023
Revised: December 20, 2023
Accepted: December 20, 2023

Engineering and Applied
Sciences in Industry
University of Guayaquil. Ecuador
Frequency/Year: 2
Web:
revistas.ug.edu.ec/index.php/easi
Email:
easi-publication.industrial@ug.edu.ec

How to cite this article:
Acosta-Guzmán, I. et al. (2023). Initial Progress of Identification of the Appropriate NLP Technique for Content Evaluation in Textual Conversations of People Infected by Sars-Cov-2. EASI: Engineering and Applied Sciences in Industry, 2(3), 5-18.
https://doi.org/10.53591/easi.v2i3.2488

Resumen. El Covid-19 se convirtió en pandemia en el 2020, generando la necesidad urgente de información fiable, se crearon Asistentes Virtuales que enseñasen al público cómo evitarlos en la variante Alfa. Pero nuevas variantes Beta, Delta y Ómicron surgieron con síntomas diferentes, provocando nuevas oleadas de infecciones y muertes. Para hacer frente a esto, se creó un prototipo de Procesamiento del Lenguaje Natural (NLP) que permita analizar las experiencias de 4.422 personas que se infectaron en Ecuador, detectando los síntomas más comunes mencionados en sus conversaciones. Este estudio impulsó la creación del prototipo NLP, empleando lenguaje Python, la plataforma Google Collab, se consideraron dos combinaciones de técnicas NLP, se realizó la medición de resultados mediante métricas de calidad, precisión, Recall, F1, encontrando que la combinación más adecuada de técnicas de las dos probadas la que dio más alta efectividad para un modelo clasificador Multietiqueta, incluyo Stop Word, Tokenización, Stemming con clasificador LSTM (Long Short- Term Memory), como primer avance del estudio.

Palabras claves: Python, Google Collab, NLP, LSTM.

Abstract. When Covid-19 became a pandemic on March 2020, an urgent need arose for reliable info and advice, so Virtual Assistants were created to help teach the public how to avoid the Alpha variant. But when new variants like Beta, Delta, and Omicron appeared with different symptoms, they caused new waves of infections and deaths. To tackle this, a Natural Language Processing prototype was created to analyze experiences of 4422 people, who had been infected in Ecuador, and to detect which symptoms were most common in their conversations. This study prompted the creation of the NLP prototype, using Python language, the Google Collab platform, two combinations of NLP techniques were considered, measuring results through quality metrics, accuracy, Recall, F1, finding that the most appropriate combination of techniques of the two tested the one that gave the highest effectiveness for a Multi-Label classifier model, including Stop Word, Tokenization, Stemming with LSTM (Long Short-Term Memory) classifier, as a first advance of the study.

Keywords: Python, Google Collab, NLP, LSTM.

1. INTRODUCTION

The world has been through into an unexpected situation due to the SARS-COV2 virus and COVID-19, which the Chinese health authorities revealed to the world on December 31st, 2019. The WHO declared it an epidemic on January 30th, 2020, and then a pandemic on March 11th, 2020. In Ecuador, a State of Health Emergency was declared on the same day by the Minister of Health, and Guayaquil was the most impacted city by the virus. To fight the virus, the authorities in Ecuador implemented total confinement, although it had a significant economic impact. People had to use digital means to communicate and stay informed. During this health situation, the Ecuadorian society found in ICT the alternative to continue with the processes of work, education, and training from home, likewise the need for communication allowed the implementation of AI technologies. To enable communication channels to provide information related to the Covid-19 pandemic through Virtual Assistants, available on various platforms. Additionally, Artificial Intelligence can help formulate appropriate treatment plans, preventive strategies and the development of drugs and vaccines (Haleem, 2020). The present research aims to create an NLP base architecture, with Machine Learning algorithms that allow detecting the most predominant symptoms that are being mentioned in the conversations of people affected by COVID19. For this purpose, several combinations of NLP data processing techniques were performed to identify the combination that yields the best results.

Due to the appearance of the Covid- 19 disease, the attention at the level of medical centers was not supplied to be able to attend a number of affected patients that was growing day by day, showing progressively that the installed capacity and the available resources were not enough to face the spread of the virus, which generated the increase of anguish in the relatives and in the infected people when facing the situation that medical centers did not receive any more patients because they had reached their maximum capacity of attention (Bonilla, 2020). Ecuador was one of the several countries affected by Covid-19, on February 29th, 2020, the first case was identified, and the expansion began to increase very rapidly, a high number of health care personnel in the exercise of their profession tested positive for Covid-19, which complicated the capacity of patient care, so tools were needed to reduce the time required for the diagnosis, monitoring, and treatment phases of Covid-19. That situation arose in environments where resources were relatively insufficient. As of March 17th, 2020, face-to-face days in the public and private sectors were suspended, granting the National Emergency Operations Committee (EOC) the right to extend the suspension. (Coronel y Perez, 2020). On the other hand, to guarantee the provision of basic public services, health, security, fire protection, risk, transportation centers, banks, food, strategic departments, and other services necessary to combat the pandemic, the government authorized that companies in these sectors would be authorized to maintain mobility and maintain working days with certain level of presence. The authorities tried to keep the bare minimum of conditions for cities to continue functioning and giving the public access to food, healthcare, basic services, export chains, agriculture and livestock, and telecommunications. This led to vehicle and personnel curfews, limits on movement time, reduced access to medical care and restrictions on accessing food sources. As a result, freedom of movement and freedom of association and assembly were suspended.

AI was brought in to watch over the Covid-19 virus and it helped to collect, analyze, and find patterns in the data.

In the current landscape, Artificial Intelligence (AI), and more specifically Natural Language Processing (NLP), has emerged as an essential pillar in understanding and addressing critical situations, such as the COVID-19 pandemic. This subfield of AI, NLP, focuses on the interaction between machines and human language, and its application has proven invaluable in analyzing conversations and extracting meaningful information.

In exploring NLP research, we delve into the ability of this technology to process human-generated text. The inherent advantages of NLP techniques become evident, as they allow us not only to understand textual information efficiently, but also to identify patterns, trends and, in our specific case, to detect predominant symptoms in conversations related to COVID-19.

Data collection for this study is based on the integration of social networks and web scraping, strategies that offer unique advantages and ethical considerations. Using social networks as data sources provides immediate and rich insights, while web scraping broadens the scope by accessing relevant information across diverse online platforms. However, these practices also raise ethical challenges, highlighting the importance of a careful approach to data collection and management.

A survey was conducted in region 8 of the province of Guayas, which includes the cantons of Guayaquil, Samborondón and Durán, to evaluate the acceptance of the final product to be developed. In this process, essential data were collected for the research, generating a DataSet with input and output variables. However, the text does not detail the process followed to create and validate the quality of the survey and interview questions.

2. MATERIALS AND METHODS

Language models play a pivotal role in natural language processing (NLP), performing tasks like text classification and generation. They learn word and context representations from extensive datasets, enhancing language understanding. Various prototypes in NLP include:

1. Word Embeddings: Represent words in a vector space (e.g., Word2Vec, Glove).
2. Contextualized Word Representations: Consider context, reducing ambiguity (e.g., BERT, GPT).
3. Sequence-to-Sequence Models: Translate sequences between domains (e.g., LSTM, Transformer).
4. Neural Networks for NLP: Learn and generalize text patterns (e.g., RNN, CNN).
5. Language Models: Generate word sequences (e.g., n-grams, Transformer).
6. Sentiment Analysis Models: Analyze sentiment in text (e.g., word vectors, neural networks).
7. Topic Models: Identify topics in documents (e.g., LDA, NMF).

In essence, these models, by learning from large datasets, contribute to highly functional and accurate language processing in NLP. Prototypes, essential in artificial intelligence, are abstract models aiding in understanding, learning, and adapting to new situations. Various types include:

1. Instance Prototypes: Concrete examples derived from real or simulated instances.
2. Nominal Prototypes: Represent categories without specific instances.
3. Abstract Prototypes: General representations applicable to different contexts.
4. Functional Prototypes: Represent the behavior an object or event can exhibit.
5. Generic Prototypes: Intersection of multiple categories, applicable to all specified categories.

Choosing the right prototype type is crucial, as each has its advantages and disadvantages. In artificial intelligence, prototypes play a role in inference and classification, forming the basis for categorizing and assigning new objects or events. In software design and development, prototypes serve as basic representations of products or systems in progress, facilitating issue identification, feedback collection, and idea refinement before full-scale development. The main types of software prototypes include:

1. Low Fidelity Prototype: Validates ideas using basic design elements, often simulated through tools like PowerPoint.
2. High Fidelity Prototype: Closer-to-final user experience with actual system code, excluding performance optimizations.
3. Graffiti Prototype: User code prototype adapted from an existing solution to incorporate desired features in a new system.
4. Visual Prototype: Graphical representation of user interface design for visualizing system appearance and function.
5. 3D Prototype: Three-dimensional representation used for system simulation and performance testing.

These prototype types empower developers and designers to explore and enhance ideas before committing to full system development.

The most suitable NLP model for the created solution is BertForSequenceClassification. Built on the Transformer algorithm, it excels in text classification tasks, making it an ideal choice for multi-label classification. Bert handles NLP tasks like sentiment analysis and entity extraction, offering high configurability for different tasks and data. Here's a Python code example using 'BertForSequenceClassification':

```
from transformers import BertTokenizer, BertForSequenceClassificationimport torch.
# Load the tokenizer and the pre-trained model
tokenizer = BertTokenizer.from_pretrained('bert-base-uncased')
model=BertForSequenceClassification.from_pretrained('bert-base-uncased')
# Process the text and convert it into a form the model can understand
text = "Here goes the text you wish to parse."
encoded_input = tokenizer (text, truncation=True, padding=True, return_tensors='pt')
# Perform the prediction and obtain the probable labelsoutput = model(**encoded_input)
predicted_labels = torch.argmax(output.logits, dim=1)
```

While powerful, note that Bert models demand substantial resources, requiring careful consideration of available resources and time constraints. In conclusion, BertForSequenceClassification is highly recommended for NLP prototypes with multi-label classification needs within the context of the created solution. The research methodologies applied are developed using the following methods:

2.1. Exploratory research

Exploratory research was used to collect preliminary information and to identify the case study to be carried out in this project. Different bibliographic sources and research using texts as primary sources were used to obtain data that made it possible to establish the input and output variables of this object of study (Solis, 2020).

2.2. Qualitative research

To carry out this type of research study, a form of scientific questions was used. The purpose of which was to obtain analysis information determined by the content of each question established in the form. Interviews were conducted with technology specialists with knowledge in AI, general practitioners or nutritionists and ordinary people who have interacted with a virtual assistant. Among them, a random sample of corresponding experts was selected from zone 8 of the province of Guayas. It is defined as a conversation with several questions with the purpose of acquiring the personal criteria of each one of them with a specific purpose of collecting data. (QuestionPro, 2019)

Interviews

An interview was conducted with three technology specialists with knowledge in AI, the engineer specialist 1, where her point of view was addressed, concluding that the most appropriate algorithms are; Decision Tree, since currently there are libraries that contain mathematical models that make it much easier to use it. Specialist 2 indicated that it is crucial to use technology related to Covid-19, such as artificial intelligence, which will greatly benefit the health department. According to specialist 3 expertise, neural networks-which mimic the workings of the human brain and allow computer programs to identify patterns and provide solutions for common AI problems-are the best algorithm for using NLP architecture.

Specialist 1	The use of NLP in the businesssector is a growing trend.	Article published by the interviewee in the academic journal "Nature".
Specialist 2	NLP techniques have the potential to revolutionize the way we interactwith machines.	Book written by the interviewee entitled "The Future of Human-Machine Interaction".
Specialist 3	The most suitable algorithm for the use of NLP architecture is Neural Networks.	The natural language processing techniques with the most functionality is Stop Word, Named Entity Recognition (NER) and Tokenization.

2.3. Quantitative research

A survey was conducted to build a form of questions where people presented their opinions in a structured way to collect information and obtain reliable statistical results in order to measure the problem and understand its scope, by seeking predictable results to a number of 384 surveys to identify the feasibility of implementing the solution and its acceptance by the residents of zone 8 of the province of Guayas, which is distributed by the cantons of Guayaquil, Samborondón and Durán, also intended for technology specialists with knowledge in AI, general practitioners or nutritionists (Campos, 2020).

Surveys

To gather data for training NLP techniques, Zone 8 in the Guayas province, including Guayaquil, Samborondón,and Durán, is selected as the study population. When using the INEC database, it was observed that Guayas canton has 2,350,915 inhabitants. A sample of 384, closely matching the surveyed count of 387, was chosen.

2.4. NLP Processing algorithms

For the creation of an NLP algorithm should be considered the next steps (Attal, 2021):

- Cleaning: It includes tasks of removing URLs, emojis, among others.
- Data normalization:
    - Tokenization, or division of the text into several parts called tokens.
    - Stemming: The same word can be found in different forms; derivation usually specifies a simple heuristic process of cutting the endings to keep only the root.
- Lemmatization: Includes performing a similar task to stemming but using vocabulary and a thorough analysis of word structure. Derivation allows removing only inflexible endings, thus isolating the canonical form of the word, called lemma.
- Other operations: Removing numbers, punctuation, symbols, and empty words and changing them to lowercase.

2.5. Common NLP Techniques

The most common techniques that have been used in the field of NLP are the following:

Tagging parts of speech (PoS)

POS tagging labels parts of speech (nouns, verbs, adverbs, adjectives, pronouns, conjunctions). There are two types of POS tags: rule-based and stochastic. Many NLP applications prefer the stochastic technique (Pham, 2020).

Shallow parsing/ Chunks

It is used to understand the grammar in a sentence. The tokens are parsed, and a structure tree is built from their PoS. Meaning: Semantic.

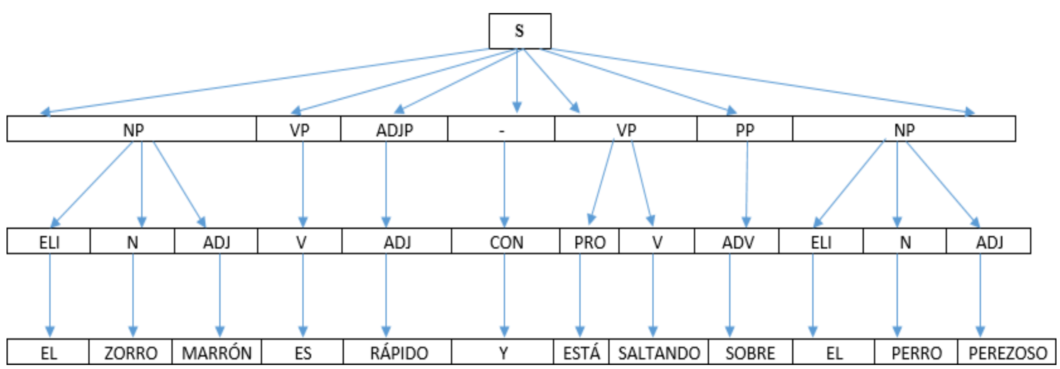


Figure 1. Example of surface parsing.

Tokenization

The concept of a "common natural language processing task (LIMA, 2021)" is introduced, which plays a fundamental role in both advanced structures such as transformers and traditional PLN techniques such as the count vectorizer. An integral step in this process is tokenization, the practice of decomposing a text into smaller pieces or tokens. These tokens can consist of words, subwords or characters.

Pragmatic analysis

Pragmatic analysis involves extracting the underlying meaning from text to make accurate inferences.

Bag of words

It consists in determining, given a set of documents or corpus, the frequency of occurrence of each word of the set in each document.

Word2vec

Word2vec generate word embeddings, uses two main word contextualization methods: the Continuous Bag-of-Words (CBOW) model and the Skip-Gram model (Sitiobigdata, 2018).

CBOW

Use the words surrounding a certain word to guess what it would be.

Skip-Gram

This is a learning technique that consists of reading large amounts of text and remembering words that appear similar in different contexts.

Stop word

Stop words, such as "an," "a," and "the," carry little meaningful information and are common in various documents. Removing these empty words, facilitated by natural language processing (NLP), streamlines tasks like sentiment analysis and document classification (Chen, 2019; Microsoft, 2021), improving efficiency in data storage and retrieval.

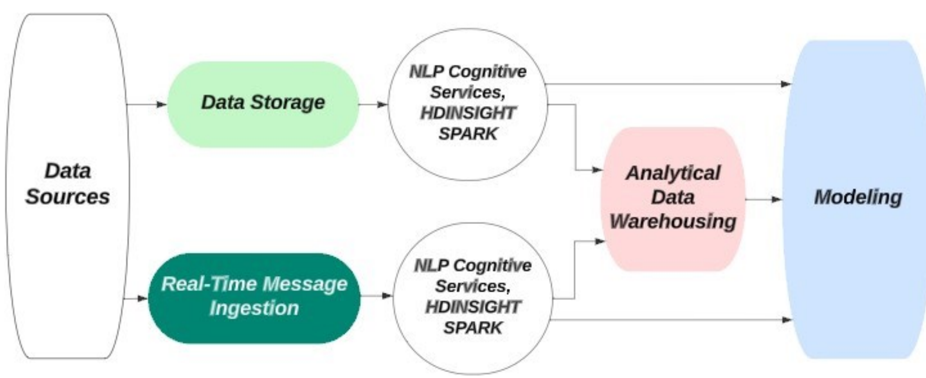


Figure 2. Orchestration of a System with NLP module.

3. IMPLEMENTATION

3.1. Analysis of survey results

The interview was conducted for the purpose of collecting detailed information about the symptoms that affected the inhabitants of Guayaquil during the COVID-19 pandemic. We inquired about people's experiences regarding the specific symptoms they experienced, the duration of symptoms, and any relevant information related to the manifestation of the disease.

We also aimed to obtain data on the healthy habits that were most effective in helping people overcome COVID-19. This approach made it possible to identify behavioral patterns that could have played a crucial role in the health and recovery of those affected, thus providing valuable information to better understand how to cope with thedisease in specific contexts, such as Guayaquil.

Once the survey of Ecuadorians who were infected and managed to overcome the disease was carried out, and the following findings were found:

EL 92.25% of the population agrees with the importance of this information of how Coronavirus affects health, and it was public immediately, and only 7, 24% partially agrees and 0, 52% disagrees.

Table 1. Importance for the villagers on having knowledge of how Coronavirus affects Health.

Options	Respondents	Percentage
Agree	357	92,25%
Neither agree nor disagree	28	7,24%
Disagree	2	0,52%
Total	387	100,00%

The 64, 59% of people trust in information received from the Ministry of Health or health sub-center related to coronavirus (covid-19), and 35, 41 has few or nothing confidence in that type of information.

Table 2. Level of confidence of information received from the Ministry of Health or health sub-center related tocoronavirus (covid-19).

Options	Respondents	Percentage
Totally agree	122	31,52%
Partially agree	128	33,07%
Slightly agree	95	24,55%
Partially disagree	31	8,01%
Strongly disagree	11	2,84%
Total	**387**	**100,00%**

49.35% of the respondents expressed a high level of knowledge regarding the application of healthy methods to reduce the risk of Covid-19 infection or death. In contrast, 50.65% indicated a lack of knowledge on this matter.

Table 3. Knowledge of the population regarding the impact of the management of healthy habits in people infectedwith coronavirus (covid-19) as a strategy to reduce the risk of serious illnesses and even death.

Options	Respondents	Percentage
Possesses a high level of knowledge of the subject	191	49,35%
Possesses low knowledge of the subject	166	42,89%
I had no knowledge of	30	7,75%
Total	**387**	**100,00%**

El 93, 54% confirmed who had a phone with internet, and only 6.46% did not have this means to receive newsrelated to Covid-19.

Table 4. Do you have Smartphone with internet access?

Options	Respondents	Percentage
I own a mid-range or high-end cell phone	233	60,21%
If I own a low range cell phone with internet	129	33,33%
I do not have a cell phone with internet	25	6,46%
Total	**387**	**100,00%**

Only 39.79% of the respondents knew that Artificial Intelligence technologies can be used to keep them informed of announcements related to mechanisms to prevent or combat Covid-19; 60.21% said they did not know about this facility that technology can provide.

Table 5. Did you know that artificial intelligence technology can develop mobile applications that allow interacting and maintaining updated information to combat the coronavirus (covid-19) anywhere in the world, at any time, including holidays?

Options	Respondents	Percentage
Possesses a high level of knowledge of the subject	154	39,79%
Possesses low knowledge of the subject	183	47,29%
I had no knowledge of	50	12,92%
Total	**387**	**100,00%**

3.2. Construction of the Machine Learning Techniques

For the development of the NLP techniques, the database generated by the surveys conducted in zone 8 of the province of Guayas is used. Google Collab is used as the main tool.

Data importation

The database is imported from the computer into the Google Collab portal.

Data processing

It consists of the preprocessing of data from 4422 infected Ecuadorians, which facilitated the refinement of raw data, obtaining a more polished dataset essential for the subsequent stage:

- I replace the tildes by vowels without tildes.
- Replace capital letters with lower case letters.
- Removal of numeric characters.
- Removal of punctuation marks.
- Removal of non-alphabetic characters.

	10. Describa lo más detallado ¿Qué síntomas ha tenido?-SIN DEPURAR	Sintomas_PROCESADA
0	Perdida de gusto y olfato	perdida de gusto y olfato
1	Dolor leve de espalda y fiebre	dolor leve de espalda y fiebre
2	Dolor al momento de respirar por algunos días	dolor al momento de respirar por algunos dias
3	Dolor de cabeza tos seca y mucha fiebre no ten...	dolor de cabeza tos seca y mucha fiebre no ten...
4	Sin sabor	sin sabor

Figure 3. Preprocessing of raw data and obtaining cleaned data.

3.3. Choice of Technique

Two combinations of textual data processing techniques were tested:

Techniques combination #1 - Tokenization with NLTK, Stop word and SpaCy lemmatization.

The application of Tokenization technique with NLTK, allows to divide texts into sentences and then divide them into words. The application of Stop Word technique allows to extract meaningless words such as articles, pronouns and prepositions that are stored in the Spanish dictionary. The application of the Lemmatization techniquewith SpaCy allows to reduce and eliminate the derivation of inflectional words and return them to the base form.

	Sintomas_PROCESADA	Tokenizado NLTK	SinStopwords
279	fiebre y dolor de cabeza	[fiebre, y, dolor, de, cabeza]	fiebre dolor cabeza
3925	malestar general	[malestar, general]	malestar general
4086	estuvo con febricula fatiga y dolor de cabeza	[estuvo, con, febricula, fatiga, y, dolor, de,...	febricula fatiga dolor cabeza
4031	me agito rapido y me duelen los huesos del cor...	[me, agito, rapido, y, me, duelen, los, huesos...	agito rapido duelen huesos corazon recien vino...

Figure 4. Techniques Combination #1 - Tokenization with NLTK, Stop Word and SpaCy Lemmatization.

Techniques Combination #2 - Tokenization with SpaCy, Stop Word and SpaCy Lemmatization.

The application of the SpaCy tokenization technique allows to divide texts into sentences and then divide them into words.

The Stop Word technique allows to extract meaningless words such as articles, pronouns and prepositions that are stored in the Spanish dictionary. Application of SpaCy Lemmatization technique allows retrieving information, reducing the redundancy of words, and generating them to their root form, as well as finding relations between words that do not exist.

Sintomas_PROCESADA	Tokenizado SpaCy	SinStopwords
perdida de gusto y olfato	[perdida, de, gusto, y, olfato]	perdida gusto olfato
dolor leve de espalda y fiebre	[dolor, leve, de, espalda, y, fiebre]	dolor leve espalda fiebre
dolor al momento de respirar por algunos dias	[dolor, al, momento, de, respirar, por, alguno...	dolor momento respirar dias
dolor de cabeza tos seca y mucha fiebre no ten...	[dolor, de, cabeza, tos, seca, y, mucha, fiebr...	dolor cabeza tos seca mucha fiebre tenia gusto...

Figure 5. Techniques Combination #2 - Tokenization with SpaCy, Stop Word and SpaCy Lemmatization.

Both combinations seek to improve data quality and relevance, but the choice between them will depend on the preference for specific tools and the emphasis on lemmatization. Combination 1 might be preferred if the diversity of tools is valued, while Combination 2 stands out for its more comprehensive approach using mainly SpaCy.

Combination 1 proved to be more effective during LSTM model training. It achieved a lower loss in unknown data compared to Combination 2, which showed a less pronounced decrease and stabilized at a higher level. This suggests that, although both strategies aim to improve data quality, Combination 1 excelled in terms of performance during the LSTM model training process, showing a superior ability to reduce the loss in unknown data.

3.4. Model Selection

The review of previous work to identify the NLP models used recommended for solving problems involving multiple outputs, also called multi-label output sorting algorithms, was carried out and it was identified that among the recommended ones are the following:

- **Combination #1:** Model LTMS, Tokenization, Stop Words and Lemmatization
- **Combination #2:** Model LTMS, Tokenization, Stop Words and Stemming

The LSTM model was chosen based on findings from previous studies that demonstrated better results forthe NLP model based on the LSTM algorithm compared to other models based on different kernels. This choice is further supported by considering a detailed evaluation of techniques, which included specifying details such as the code used to define the technique. Additionally, a sequential model was opted for in the construction of the neuralnetwork, adding a layer of complexity and optimization to the process.

For this reason, for the present study we chose to apply the LSTM Neural Network model. With this model, the two combinations of sets of techniques mentioned above were tested to identify which combination of techniques achieved the highest values in the quality metrics of the model. As part of the neural network construction process, the sequential model was chosen, with a layer of embeddings, a bidirectional layer, and two dense layers, indicating each of its indicators with limits, the activation of the mathematical logistic function is performed.

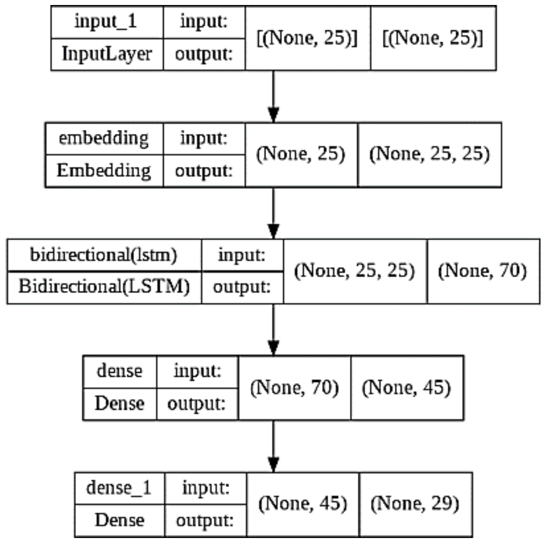
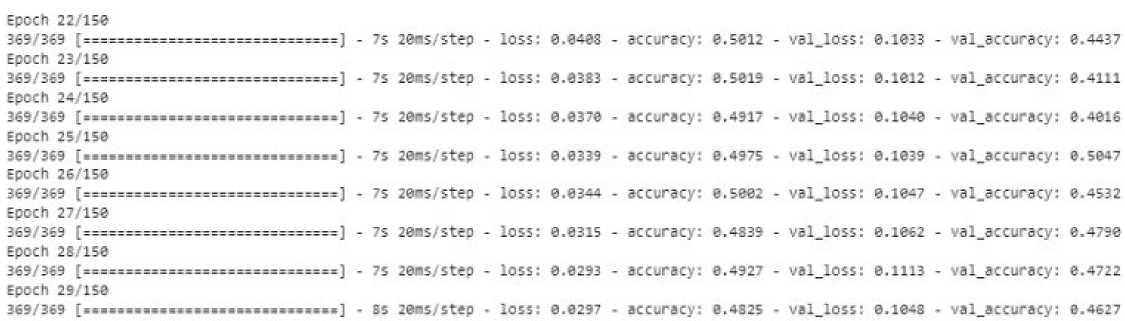


Figure 6. Choice of the type of model and activation of Softmax function.

3.5. Algorithm Training

The Figure 7 shows the training of the LSTM model, based on the predictions made with the combination 1 of techniques (Tokenization, Stop Words and Lemmatization) using the neural network model with sequential algorithm.



```
Epoch 22/150
369/369 [==============================] - 7s 20ms/step - loss: 0.0408 - accuracy: 0.5012 - val_loss: 0.1033 - val_accuracy: 0.4437
Epoch 23/150
369/369 [==============================] - 7s 20ms/step - loss: 0.0383 - accuracy: 0.5019 - val_loss: 0.1012 - val_accuracy: 0.4111
Epoch 24/150
369/369 [==============================] - 7s 20ms/step - loss: 0.0370 - accuracy: 0.4917 - val_loss: 0.1040 - val_accuracy: 0.4016
Epoch 25/150
369/369 [==============================] - 7s 20ms/step - loss: 0.0339 - accuracy: 0.4975 - val_loss: 0.1039 - val_accuracy: 0.5047
Epoch 26/150
369/369 [==============================] - 7s 20ms/step - loss: 0.0344 - accuracy: 0.5002 - val_loss: 0.1047 - val_accuracy: 0.4532
Epoch 27/150
369/369 [==============================] - 7s 20ms/step - loss: 0.0315 - accuracy: 0.4839 - val_loss: 0.1062 - val_accuracy: 0.4790
Epoch 28/150
369/369 [==============================] - 7s 20ms/step - loss: 0.0293 - accuracy: 0.4927 - val_loss: 0.1113 - val_accuracy: 0.4722
Epoch 29/150
369/369 [==============================] - 8s 20ms/step - loss: 0.0297 - accuracy: 0.4825 - val_loss: 0.1048 - val_accuracy: 0.4627
```

Figure 7. Training of the sequential algorithm with combination #1.

In the following image we can observe the training of the LSTM model, based on the predictions made with the combination 2 of techniques (Tokenization, Stop Word and Stemming) using the neural network model with sequential algorithm.

```
Epoch 23/150
369/369 [==============================] - 7s 19ms/step - loss: 0.1513 - accuracy: 0.4523 - val_loss: 0.1683 - val_accuracy: 0.4478
Epoch 24/150
369/369 [==============================] - 7s 19ms/step - loss: 0.1485 - accuracy: 0.4710 - val_loss: 0.1676 - val_accuracy: 0.4261
Epoch 25/150
369/369 [==============================] - 7s 19ms/step - loss: 0.1468 - accuracy: 0.4737 - val_loss: 0.1657 - val_accuracy: 0.4423
Epoch 26/150
369/369 [==============================] - 7s 19ms/step - loss: 0.1433 - accuracy: 0.4693 - val_loss: 0.1682 - val_accuracy: 0.4668
Epoch 27/150
369/369 [==============================] - 7s 19ms/step - loss: 0.1415 - accuracy: 0.4886 - val_loss: 0.1663 - val_accuracy: 0.4545
Epoch 28/150
369/369 [==============================] - 7s 19ms/step - loss: 0.1387 - accuracy: 0.4869 - val_loss: 0.1645 - val_accuracy: 0.4627
Epoch 29/150
369/369 [==============================] - 7s 19ms/step - loss: 0.1365 - accuracy: 0.4832 - val_loss: 0.1650 - val_accuracy: 0.4654
Epoch 30/150
```

Figure 8. Training of the sequential algorithm with combination #2.

3.6. Algorithm Evaluation

In figure 10, the evolution of the prediction quality is represented, the quality curves obtained with training data and validation data are presented, the Y value is the accuracy, and the X value corresponds to the training epochs (iterations).

Combination metrics #1

As a result of the first prediction of combinations of techniques, the training of the first combination gave a better result because the loss curve in unknown data decreased to 0.10.

After analyzing the initial predictions involving various combinations of techniques, it was observed that the first combination outperformed others during training. This superiority was particularly evident in the reduction of the loss curve concerning previously unseen or unknown data, reaching an impressively low value of 0.10. This indicates that the model's performance, especially in handling unfamiliar information, was notably effective when employing the specific set of techniques in the first combination.

Combination metrics #2

As a result of the second prediction of combinations of techniques, the training of the second combination was able to obtain the following results, with less effectiveness because the loss curve in unknown data dropped to 0.17 and at that point it stopped dropping.

Following the second round of predictions involving different combinations of techniques, the training of the second combination produced less favorable outcomes. Specifically, the loss curve associated with unfamiliar data experienced a decline but only reached 0.17, after which it ceased to decrease further. This indicates a comparatively lower level of effectiveness in handling unknown information when employing the set of techniques in the second combination, as opposed to the more successful first combination.

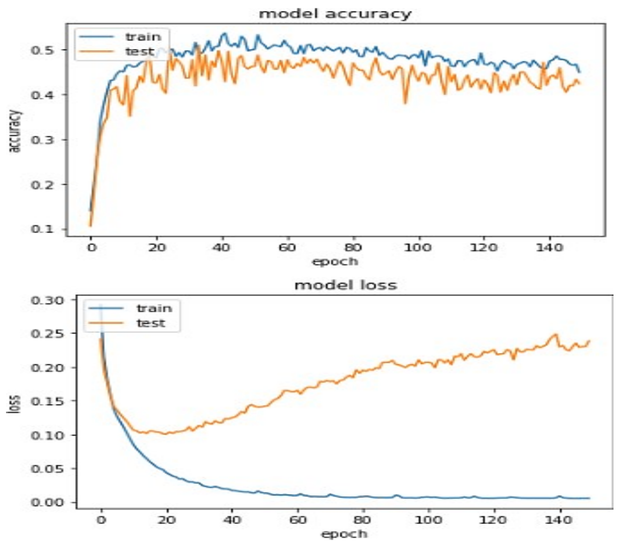


Figure 9. Metrics visualization of combination #1.

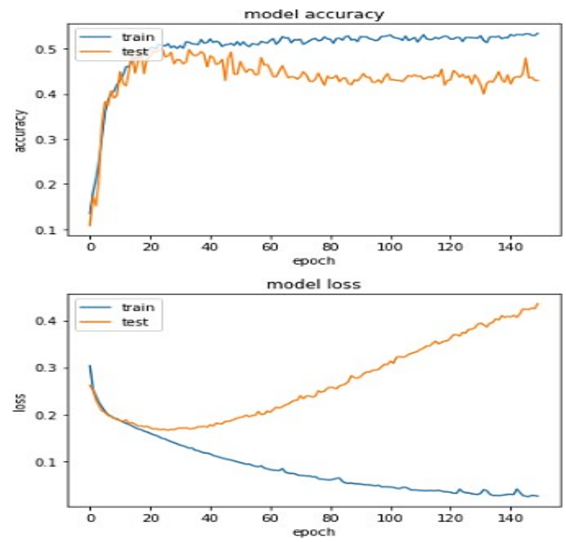


Figure 10. Metrics visualization of combination #2.

4. RESULTS AND DISCUSSION

The results obtained in this study demonstrate the successful development of basic natural language processing techniques applied to automatic information processing systems, with a particular focus on Spanish information retrieval. The proposed classification model, which is based on semantic labeling of information, has proven to be effective in reusing algorithms and methods used in character extraction to classify information.

The achievement of the main research objective is reflected in the ability of the proposed method to process Spanish texts and classify relevant information, especially in the context of the symptom column of a database. The use of natural language processing and machine learning techniques stands out as fundamental in this study, as it allows determining the text processing before and after the system training.

This approach provides a process that facilitates the extraction of the most relevant information from the symptom database, which is essential for automatic information processing systems in medical settings or other contexts where accurate information retrieval is crucial.

CONCLUSIONS

The combination techniques 1: Tokenization, Stop Word, and SpaCy Lemmatization for the Spanish language, were the techniques that achieved the best results with known data (train data) and unknown data (test data).

Using a sequential LSTM Neural Network model for human text classification, it was possible to identify one combination of preprocessing techniques yielded the good results in predictions with multi-labels outputs.

Natural language processing and machine learning research is crucial for establishing how text is processed before and after system training.

It is recommended to continue promoting research related to use another combination of NLP techniques that let develop better tools that support the human being before the emergence of future viruses that affect humanity.

Acknowledgements

The authors would like to express their sincere gratitude to the following institutions for their invaluable contributions to this research:

- Facultad de Ingeniería Industrial, Universidad de Guayaquil.
- Facultad de Ciencias Matemáticas y Físicas, Universidad de Guayaquil.
- Facultad de Educación, Universidad Estatal de Milagro
- Facultad de Ingeniería en Electricidad y Computación, Escuela Superior Politécnica del Litoral at Ecuador.

Their support and collaboration have been essential in advancing the work in the areas of Artificial Intelligenceand Natural Language Processing (NLP). The authors extend their heartfelt thanks to these institutions for their significant contributions to the research efforts.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest within this research, authorship, and/or publication of this article.

REFERENCES

Attal, M. (Diciembre de 2021). Mapeo de incrustaciones de Word con Word2vec. https://datascientest.com/es/nlp-natural-language-processing-introduccion

Bonilla, G. J. (2020). Las dos caras de la educación en el Covid-19. Cuestiones de Administración, 9(2). https://doi.org/10.33210/ca.v9i2.294

Brownlee, J. (Octubre de 2017). Machine Learning Mastery. https://machinelearningmastery.com/gentle-introduction-bag-words-model/

Campos, C. (Febrero de 2020). Fases del proceso de investigación científica y elementos de la investigación cuantitativa y cualitativa. https://www.scribd.com/document/447304281/Actividad-N-02-Fases-del-proyecto-de-Investigacion-Cientifica-inv-cualitativa-y-cuantitativa

Chen, P. H. (2019). Essential Elements of Natural Language Processing: What the Radiologist Should Know. Academic Radiology. https://doi.org/10.1016/j.acra.2019.08.010

Coronel y Pérez (Abril de 2020). Covid-19 y efectos. https://www.coronelyperez.com/2020/04/23/la-crisis-ocasionada-por-el-covid-19-y-sus-implicaciones-legales-en-el-ecuador/

Haleem, RV. (2020). Artificial Intelligence (AI) applications for COVID-19 pandemic. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 14(4), 337-339. https://doi.org/10.1016/j.dsx.2020.04.012

Instituto Nacional de Estadística y Censos (INEC). (2023). Encuesta de salud y nutrición (ENSANUT). Recuperado de https://www.ecuadorencifras.gob.ec/encuesta-de-salud-y-nutricion-ensanut/

Johnson, D. (Enero de 2022). What is Natural Language Processing? https://www.guru99.com/nlp-tutorial.html

Kohlbacher, F. (2006). The use of qualitative content analysis in case study research. Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, 7(1), Art. 21. Recuperado de http://www.qualitative-research.net/index.php/fqs/article/view/75/153

Labarthe, S. (2020). ¿Qué pasa en Ecuador? https://www.nuso.org/articulo/que-pasa-en-ecuador/

LIMA, A. (2021). PNL CÓMO FUNCIONA LA TOKENIZACIÓN DE TEXTO, ORACIONES Y PALABRAS. https://es.acervolima.com/pnl-como-funciona-la-tokenizacion-de-texto-oraciones-y-palabras/

León, E. (Diciembre de 2020). Procesamiento del lenguaje natural (PLN) con Python. Baoss Analytics Everywhere. https://www.baoss.es/procesamiento-del-lenguaje-natural-pln-con-python

López, I. P. (2018). Análisis comparativo de algoritmos de Deep Learning para la clasificación de textos. https://e-archivo.uc3m.es/bitstream/handle/10016/29209/TFG_Ivan_Lopez_Pacheco_2018.pdf?sequence=1

Microsoft. (2021). Tecnología de procesamiento de lenguaje natural. https://docs.microsoft.com/es- es/azure/architecture/data-guide/technology-choices/natural-language-processing

Ministerio De Salud Pública. (Marzo de 2020). Informe De Situación Covid-19 Ecuador. https://www.gestionderiesgos.gob.ec/wp-content/uploads/2020/03/informe-de-situaci%c3%b3n-no008-casos-coronavirus-ecuador-16032020-20h00.pdf

OMS. (Marzo de 2020). La OMS caracteriza a COVID-19 como una pandemia. Recuperado de https://www.paho.org/es/noticias/11-3-2020-oms-caracteriza-covid-19-como-pandemia

Pedregosa, F., Varoquaux, G. & Gramfort, et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825—2830. https://scikit-learn.org/stable/modules/multiclass.html

Pham, B. (Febrero de 2020). Parts of Speech Tagging: Rule-Based. Computer and Information Sciences

Undergraduate. https://digitalcommons.harrisburgu.edu/cgi/viewcontent.cgi?article=1001&context=cisc_student- coursework

QuestionPro. (Noviembre de 2019). ¿Qué es la investigación cualitativa?https://www.questionpro.com/es/investigacion-cualitativa.html

Sitiobigdata. (Agosto de 2018). Mejora de incrustaciones de Word con Word2Vec. https://sitiobigdata.com/2018/08/24/mapeo-de-incrustaciones-de-word-con-word2vec/#

Solís, L. D. (Febrero de 2020). La entrevista en la investigación cualitativa. https://investigaliacr.com/investigacion/la-entrevista-en-la-investigacion-cualitativa/