



Modelos probabilísticos IA del procesamiento de lenguaje natural en conversaciones de personas contagiadas con Covid-19

Probabilistic AI models of natural language processing in conversations of people infected with Covid-19

Modelos probabilísticos de IA de procesamiento de linguagem natural em conversas de pessoas infectadas com Covid-19

Lady M. Sangacha-Tapia I
Lady_tapia@hotmail.com
<https://orcid.org/0000-0002-5169-8918>

Ricardo Javier Celi-Párraga II
ricardo.celi@utelvt.edu.ec
<https://orcid.org/0000-0002-8525-5744>

Eleanor A. Varela-Tapia III
eleanor.varelat@ug.edu.ec
<https://orcid.org/0000-0002-5357-4046>

Ivan Leonel Acosta-Guzmán IV
ivan.acostag@ug.edu.ec
<https://orcid.org/0000-0002-1589-1825>

Correspondencia: Lady_tapia@hotmail.com

Ciencias de la Salud
Artículos de investigación

***Recibido:** 16 de julio de 2021 ***Aceptado:** 30 de agosto de 2021 * **Publicado:** 08 de septiembre de 2021

- I. Magíster en Sistemas de Información Aplicada; Magíster Universitario en ingeniería de software y sistemas informático, Guayaquil, Ecuador.
- II. Ingeniero en Sistemas Informáticos; Magister Universitario en ingeniería de software y sistemas informático; Docente Universidad Técnica Luis Vargas Torres de Esmeraldas.
- III. Magíster en Sistemas de Información Gerencial; Magister en Docencia y Gerencia en Educación Superior; Diploma Superior en Diseño Curricular por Competencias; Ingeniera en Computación; Docente Universidad de Guayaquil, Guayaquil, Ecuador.
- IV. Magíster en Sistemas de Información Gerencial; Magister en Administración de Empresas; Ingeniero en Computación; Docente Universidad de Guayaquil, Guayaquil, Ecuador.

Resumen

El Procesamiento del Lenguaje Natural que en inglés se lo identifica como NLP es una de las ramas de la Inteligencia Artificial, donde los ordenadores son programables para simular la comunicación entre la máquina y la lengua humana, en nuestro caso se pretendió ofrecer un aplicativo que realice una conversación textual móvil dirigido a un sector vulnerable de la zona 8 en vista que existe la necesidad de otorgar información de buenos hábitos de higiene cuando se encuentran con covid-19, sobre el dimensionado o sub dimensionado con la ambigüedad del lenguaje aún es un gran problema de tratar sin embargo en el planteamiento actual se enfatiza el avance de resultados de la primera fase de la investigación, teniendo como objetivo la identificación de los modelos para lenguaje de procesamiento natural en las conversaciones textuales de personas contagiadas de covid-19, en ello se presenta tablas comparativas, donde se aplicó una metodología de un método de la recopilación documental, una revisión de análisis de los diferentes modelos para ese procesamiento de lenguaje natural, conociendo donde el lenguaje puede ser tratada por el dispositivo móvil, de esta manera el resultado de reconocer fue la combinación de varios modelos del TRIGRAM y el perplejidad, que permiten la probabilidad estimada de una secuencia de palabras, la cual nos optimiza el tiempo de respuesta del mismo ya que calcula la probabilidad condicionada, disminuyendo la confusión del texto, con la aplicación de un corpus de entrenamiento.

Palabras claves: Procesamiento de lenguaje natural; corpus; modelo para nlp; trigram; perplejidad.

Abstract

The Natural Language Processing that in English is identified as NLP is one of the branches of Artificial Intelligence, where computers are programmable to simulate the communication between the machine and the human language, in our case it was intended to offer you an application that performs a mobile text conversation directed to a vulnerable sector of zone 8 in view of the need to provide information on good hygiene habits when they encounter covid-19, on the dimensioning or undersizing with the ambiguity of the language is still a great However, the current approach emphasizes the advancement of the results of the first phase of the research, aiming at the identification of models for natural processing language in the textual conversations of people

infected with covid-19. Comparative tables are presented, where a methodology of a document collection method was applied. Thus, a review of the analysis of the different models for that natural language processing, knowing where the language can be treated by the mobile device, in this way the result of recognizing was the combination of several models of TRIGRAM and perplexity, which allow the estimated probability of a sequence of words, which optimizes its response time since it calculates the conditional probability, reducing the confusion of the text, with the application of a training corpus.

Keywords: Natural language processing; corpus; model for nlp; trigram; perplexity.

Resumo

O Processamento de Linguagem Natural que em inglês é identificado como PNL é um dos ramos da Inteligência Artificial, onde os computadores são programáveis para simular a comunicação entre a máquina e a linguagem humana, no nosso caso pretendia-se oferecer um aplicativo que realizasse um mobile conversa de texto dirigida a um setor vulnerável da zona 8 tendo em vista a necessidade de fornecer informações sobre bons hábitos de higiene quando se deparam com covid-19, sobre o dimensionamento ou subdimensionamento com a ambigüidade da língua ainda é um grande. No entanto, a abordagem atual enfatiza o avanço dos resultados da primeira fase da pesquisa, visando a identificação de modelos de processamento natural da linguagem nas conversas textuais de pessoas infectadas com covid-19. São apresentadas tabelas comparativas, onde foi aplicada uma metodologia de método de coleta de documentos . 1, uma revisão da análise dos diferentes modelos para aquele processamento de linguagem natural, sabendo onde a linguagem pode ser tratada pelo dispositivo móvel, desta forma o resultado do reconhecimento foi a combinação de vários modelos de TRIGRAM e perplexidade, que permitem a probabilidade estimada de uma sequência de palavras, o que otimiza seu tempo de resposta, pois calcula a probabilidade condicional, reduzindo a confusão do texto, com a aplicação de um corpus de treinamento.

Keywords: Processamento de linguagem natural; corpus; modelo para nlp; trigrama; perplexidade.

Introducción

En 1940 inició su funcionamiento la primera bomba criptológica inglesa en Bletchley Park, inventada por Allan Turing, llamada el Proyecto Ultra o Bomba de Turing, con la cual se logró interpretar los mensajes cifrados por la marina alemana quienes aplicaban el esquema de cifrado Enigma con la variante delfín, la cual fue de mucha ayuda en la segunda guerra mundial (Morales-Luna, 2013), luego de varios acontecimientos referentes a la relación del lenguaje y el pensamiento, la teoría de Chomsky emitida en 1957 según el enfoque lingüístico, permanece vigente a pesar de sus detractores, ya que en los últimos 60 años ha sido uno de los fundamentos académicos más importantes para comprender la mente, y es actualmente continúa siendo una de las explicaciones más influyentes sobre cómo los seres humanos adquirimos, y producimos el lenguaje. (Birchenall & Müller, 2014)

El entendimiento del lenguaje, y su correspondiente procesamiento de textos de forma conversacional ha sido un problema de investigación desde los años 50s hasta la actualidad, sin embargo hoy en día se cuenta con la rama de la Inteligencia Artificial (IA), denominada Procesamiento de Lenguaje Natural (PLN, o NLP en inglés), la cual se espera permita impulsar la lingüística computacional aplicada en lenguaje verbal y/o en lenguaje textual.

En la actualidad existen dos grandes desafíos al momento de procesar textos, los cuales son:

- La ambigüedad y
- La dimensionalidad del lenguaje

Estos aspectos hacen que el proceso de NLP se transforme en un problema complejo, puesto que la combinación entre Ciencia de Datos y Lingüística Computacional.

Mientras que la Ciencia de Datos permite recopilar, analizar, deducir información, apoyada en el aprendizaje automático, la lingüística computacional permite la comprensión y la elaboración de contenidos en lenguajes nativos. Ambos componentes tornan compleja la búsqueda de un modelo que sea el más adecuado para el procesamiento, interpretación y utilización automatizada del Lenguaje. (Egea, 2019)

El presente trabajo, busca realizar la revisión de modelos disponibles en PLN, con la finalidad de identificar el algoritmo, técnica y entrenamiento adecuados, disponibles en las técnicas de machine learning, enfocados al análisis del lenguaje textual.

Por otra parte, entre los niveles del procesamiento de Lenguaje Natural se tiene el morfológico, sintáctico, semántico, pragmático. Se revisó el desarrollo de cada uno así como la clasificación de NLP, dos modelos son esenciales en el procesamiento de lenguaje natural, el lógico (basado en gramática) y el probabilístico (basado en los datos) (Egea, 2019).

En este documento se desarrolla en el estudio de los datos en el análisis de un corpus, es decir en el modelo para NLP probabilístico, para ello es fundamental tomar el lenguaje como objeto de estudio y la comprensión del mismo como un proceso complejo en que intervienen gigantes cantidades de datos, en los que se descubrirán contenidos mediante mecanismos de tratamientos variados. Se procesarán conversaciones textuales relacionadas a la pandemia covid-19 en Ecuador, con el modelo identificado aplicable a NLP. Esta primera etapa de la investigación busca establecer el punto de partida para la fase 2, en la que se propone implementar un algoritmo mejorado con agentes virtuales optimizados utilizables en dispositivos móviles (Guaman et al., 2020).

Como parte de la primeras actividades se encuentra la de generar, conseguir o seleccionar el Dataset adecuado basado en conversaciones textuales obtenidas de ecuatorianos que fueron contagiados por Covid-19, por tal razón se ha considerado varias estrategias para la obtención de esa data, entre ellas se ha optado por la utilización de las bondades disponibles en redes sociales, en este estudio se eligió twitter, empleada en otras investigaciones por la disponibilidad la API de Twitter utilizable desde programas creados en Python o R, que permiten la contar con una interfaz que apoya al proceso ETL (Twitter, 2017). al permitir contar con la etapa de Extracción (E) o toma de conversaciones, viabilizando así la posterior etapa que consiste en el tratamiento o transformación (T) de la data y carga (L de Load) al repositorio de destino (Mutlu et al., 2020).

Fundamentos de la Inteligencia Artificial

Entre los acontecimiento más relevantes de la evolución de la IA se identifican los siguientes:

Orígen IA en la Filosofía (428 a.C)

La formulación del conjunto de precisión de leyes que explicaban la parte racional de la inteligencia fue de Aristóteles entre los 384-322 a.C.

La formulación informal para razonar acertadamente con silogismo comprende en la argumentación estructurada, puesto que antes del siglo 19 era por conclusiones, por premisas. Por lo que Ramón Lull dio a conocer que ese razonamiento (Suárez, 2014) funcionaba mejor a través del uso de medios artificiales.

Según Orozco-Echeverri(2009), indicó que Thomas Hobbes señaló como propuesta que el razonamiento era la computación numérica. Incluso Leonardo da Vinci (Valencia Giraldo, 2000) solo diseñó una calculadora mecánica donde en la actualidad se ha demostrado que era funcional ya que alrededor de 1623 se construyó la máquina calculadora por Wilhelm Schickard pero la pascalina fue creada en 1642 por Blaise Pascal (Reckoner, 1930), en esa época Gottfried Wilhelm **Libniz creó el dispositivo mecánico la cual poseía limitaciones.**

Entre los años 1596 a 1650 surge la primera discusión clara acerca de la diferencia entre la mente y la materia (Russell, 2004), además de los problemas que suscitan, en aquella época se decía que los pensamientos eran gobernados, por lo cual en el transcurso del tiempo se presentan discusiones como:

- El Dualismo
- El Materialismo
- El Movimiento Empírico
- La Inducción
- El Positivismo Lógico
- Las Sentencias de Observación
- La Teoría de Conformación

Luego del concepto de la lógica aparece el pensamiento del primer algoritmo, aunque en la lógica de primer orden no era posible capturar principios de inducción matemática de la caracterización de los números naturales debido a que el teorema de la incompletitud demostró las propiedades de los números naturales (Da Silva, 2014).

Por otra parte, aparecen las “Aseveraciones Verdaderas” las cuales sostienen que no era posible decidir la validez de algunas funciones con números enteros indicando que no es posible ser representada por algoritmos, por tal razón esas funciones que son realizadas con números enteros no permitían calcular por lo que esta situación llevó a Alan Turing (1912-1954) a tomar en cuenta características que conlleven a las funciones a ser caracterizadas, a raíz de esta idea surge la máquina de Turing en 1936, la cual era capaz de realizar los cálculos de funciones computables. El creador de la máquina indicó con frontalidad que la máquina no podía decidir por lo que surgió

la noción de intratabilidad, en 1971 surge la teoría de la NP-completitud generada por Steven Cook (Vanoye, 2008).

En el año 1929 aparece la tercera aportación de las matemáticas a la IA, la teoría de la probabilidad por Geolamo Cardano, siendo el primero en proponer la idea en el mundo de las matemáticas (Pellicer, 2007).

En la economía el pensamiento surgió (Smith, 1776) en la que se combina la teoría de la decisión con la teoría de la probabilidad, en aquella época era tan emocionante realizarla para luego adecuarla con la teoría de los juegos.

La aparición de la neurociencia en el año 1861, enfocado o apoyado el estudio del sistema neurológico en especial del cerebro.

En cambio en el año de 1879 aparece la psicología científica por el alemán Hermann Von (Abrego Almazán et al., 2017)

En los años de 1940 aparece la ingeniería computacional, donde se conoce que para ello se necesita combinar el artefacto y la inteligencia, lo que sería llamado el ordenador, el cual fue empleado en la segunda guerra mundial constituyéndose en el primer computador operacional de los 40s con la finalidad de descifrar mensajes alemanes.

Claro está que antes de la aparición de computadoras ya habían dispositivos creados para realizar cálculos, la aparición de máquinas automáticas, programables puede remontarse a 1805 donde se crea una máquina para la elaboración de un telar, en 1822 tenemos la máquina diferencias y 1837 la máquina analítica ambas propuestas de Babbage (Giudice, 2001) (Merodio, 2020), en los recientes años por su parte la Inteligencia Artificial ha proporcionado ayuda a los seres humanos en múltiples frentes o disciplinas, una de ellas es su presencia en los sistemas operativos con IA, los lenguajes de programación con IA, Asistentes Virtuales, Seguridad, Reconocimiento de imágenes en fotos, videos, entre otras numerosas aplicaciones (Libertad Digital, 2018) (ENAE, 2021).

En el año 1948 donde Norbert Wiener planteó la teoría de control y retroalimentación, dando paso a la ciencia de la cibernética, plasmando sus ideas en la creación de la primera máquina autocontrolada.

En el año 1957 algunos autores demostraron interés por el estudio del Aprendizaje del Lenguaje, entre ellos Chomsky quien propone un estudio a través del enfoque lingüístico, cuyos inicios se remontan a los modelos sintácticos de la lingüística de Panini en el año 350 a.C.

Origen del Procesamiento de Lenguaje Natural (PLN o NLP)

En el año de 1957 se presentó la complejidad del entendimiento del lenguaje, se encontró lo complicado que era comprender el contexto con el contenido de la materia. Por tal razón el nacimiento del Procesamiento de Lenguaje Natural está bajo la lingüística computacional o procesamiento del lenguaje natural (Gomez, 2019).

Está en el área multidisciplinaria de la lingüística computacional contiene clasificaciones como son:

- La Comprensión del Lenguaje Natural con sus siglas en inglés NLU
- La Generación del Lenguaje Natural con sus siglas en inglés es NLG

Definición Procesamiento de Lenguaje Natural (PLN o NPL)

En los años 60 revoluciono la curiosidad del contenido llamado lingüística de corpus (Domínguez Burgos, 2002). Recordando que Eliza fue uno de los programas con la capacidad de realizar conversaciones de forma muy limitada con personas.

Existen algunas definiciones de procesamiento de Lenguaje Natural, que se presentaron en los años 2000, una de ellas indica que el NPL es parte de la Inteligencia Artificial, y un área de la Lingüística Computacional, la cual permite el estudio de la comunicación entre el humano y la máquina (Borja, 2020)

Otra definición indica que el lenguaje Natural está conformado por palabras, reglas sintácticas, semánticas por el sonido de la voz, gestos que permite la comunicación de los humanos (Borja, 2020)

Modelos para PLN

Existe mucha información de la Inteligencia Artificial, hoy en día aún se está investigando distintas ramas de la misma, entre ellas el Procesamiento de Lenguaje Natural, sobre todo por los dos grandes problemas que este procedimiento conlleva, los cuales son la ambigüedad y la dimensionalidad de los textos. Es por ello que para lograr llegar a un entrenamiento es importante

conocer bajo qué modelo se realizaría el procesamiento, en este trabajo se reviso 2 modelos muy utilizados (Gomez, 2019) en diversas aplicaciones mismos que se menciona a continuación:

- Modelos Lógicos
- Modelos Probabilísticos

Estos modelos son tomados desde el conocimiento matemático, esto implica que es necesario tener claro los beneficios de cada modelo para realizar el análisis y por supuesto el tratamiento del lenguaje. En particular el desarrollo de este artículo se enfoco en el estudio del modelo probabilístico.

Modelo lógicos

En los años 70 se presentaron varias propuestas de arquitecturas, entre ellas el modelo de MARKov Ocultos (MMO) el cual ha tomado importancia por su interesante teoría matemática, manteniendo la significancia por varias décadas (Angela A, 2018)

Por su parte surgen los modelos gramaticales, propuestos por lingüistas expertos quienes fueron los creadores con base a determinadas formas gramaticales, extrayendo características del mundo en modelo conceptual a un lenguaje lógico (Contreras, 2001).

Modelos probabilístico

Son mencionados como basados en corpus, basándose en datos, donde su esencia es el análisis de la información recopilada lingüísticamente, se pueden tomar de muchas formas como las grabaciones, registros, en la actualidad son tomados desde las conversaciones de las redes sociales, un conjunto infinitos de enunciados gramaticalmente, donde sus términos, reglas gramaticales, fonemas, palabras, frases etc, donde ese lenguaje es tomado para encontrar un algoritmo que determine a qué grupo pertenece el lenguaje, es decir que reconozca el lenguaje.

Para ello se es empleada con la utilización de estructura gramaticales, que pueden ser de forma secuencial como de pares o de tríos, aquí también se incluyen la estadística del uso de la regla de gramática. A continuación se presenta el enfoque de los diferentes algoritmos probabilísticos

- N-gram
- Unigram

- Bigram
- Trigram

El modelo N-gram formado por n palabras seguidas integradas en el corpus

Conversaciones de Covid-19 con personas contagiadas

El gobierno ecuatoriano declaró el estado de excepción en el mes de marzo del año 2020, con la finalidad de confinar a los ciudadanos o no ciudadano del Ecuador, debido al virus declarado en pandemia porque la contaminación que era a nivel mundial, para nadie ha sido sorpresa los momentos difíciles que vivió cada familia, cada clase social, y de los diferentes sectores, aunque en la economía del sector productivo de alimentos se benefició por el momento presentado.

Las conversaciones se dieron por mensajerías, por llamadas a voz, incluso resaltó el uso de las redes sociales como el tweeter.

Método

Para lograr identificar el modelo para el procesamiento de datos como fase 1 de la investigación se tuvo la necesidad de la recopilación de fuentes bibliográficas, artículos, revistas, realizando una serie de actividades para la obtención de un cuadro comparativo. Dentro de ellas se considera una 3 momentos de la FASE 1 que se detalla a continuación:

Desarrollo de la FASE 1

Momento 1: Creación de espacio cluster que se podrá reconectar todo el equipo investigador, con la ayuda de los contenido de la revista, ayudará alimentar el espacio de artículos, revistas, académicas, la cual son revisadas, evaluadas y seleccionadas todos aquellos contenidos de mayor relevancias, esa preparación de vital importancia para la consecución de la investigación, obteniendo la tabla comparativa de los dos diferente modelos para PLN.

Momento 2: Creación del pre-procesamiento de datos, antes de la realización del estudio de conversaciones textuales de personas contagiadas de covid-19, se creó el OBJETIVO de proyecto, muy vital para este momento. Una vez teniendo la información textual de los ecuatorianos desde las redes sociales, eso permitió la creación de una mini data set que permitirá realizar el entrenamiento con el modelo escogido para PLN.

Momento 3: La obtención de resultados, se crean las tablas comparativas en base a los momentos anteriores pero de los diferentes algoritmos.

El desarrollo del momento 1 no presentó problema porque contamos con un equipo de buen conocimiento a la informática, la creación de un espacio que se pueda abarcar, crear una base donde

está la recolección de diferentes fuentes de información, desde los años iniciales del procesamiento de lenguaje natural hasta en los años actuales, la herramienta permite la interacción con cada uno de los investigadores.

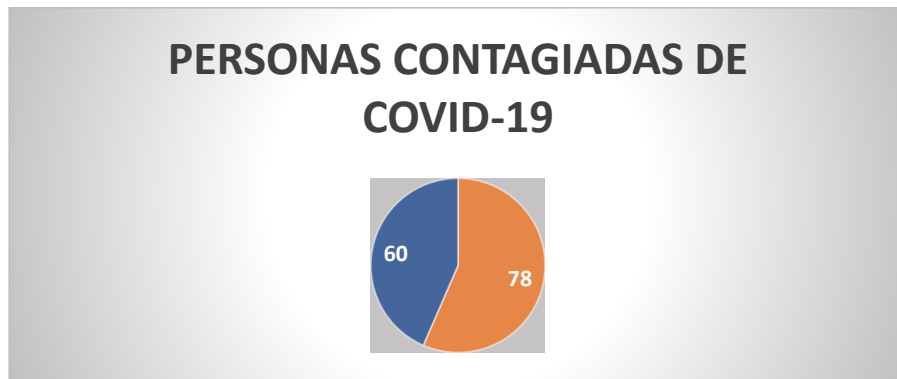
Se presenta una tabla comparativa entre modelos (Ochoa, 2009) de los dos modelos para PLN que en la sección superior se transcribió con (Dahl, n.d. 2010) y (Cerezo, 2020).

Figura 1: Conceptos comparativos de los diferentes modelos para PLN

Lógicos (gramaticales)	Probabilísticos (basados en corpus)
Se basan en determinadas formas gramaticales.	Se basan en los datos como eje principal del análisis.
Son creados por lingüistas especializados.	Los lingüistas recopilan información para analizar y establecer la frecuencia de aparición de todas las unidades que forman la lengua.
Las máquinas pueden reconocer diferentes patrones de estructuras lingüísticas.	Los algoritmos pueden establecer la probabilidad de que aparezcan en un contexto.
Identifica acrónimos existentes, seleccionando los tokens de oraciones o palabras que deseen.	Recopilan una cantidad de información para analizar y establecer la frecuencia de aparición de todas las unidades.
Se combina con la información que ya existe en los diccionarios de computación para poder definir los modelos de resolución de una tarea concreta a través del proceso de automatización de los procesos conversacionales.	Se realiza sin tener que valorar los requisitos establecidos en las reglas gramaticales en ningún momento.
Recogen su esencia para aplicarla a la comunicación entre máquina y persona.	Esencia, de lo que se denomina aprendizaje automático en Inteligencia Artificial.

En el desarrollo del momento 2, se consideró un orden a tomar para llegar a lo deseado, por lo que es importante respetar el análisis de lo que se va a requerir, en este caso lo primero es tener claro el objetivo del proyecto para llegar a un pre procesamiento de datos, como segundo orden es la elaboración de preguntas que en este caso sería utilizar la herramienta de investigación de formularios como tercer caso es la recolección de información para que finalmente los datos sean tratados y pre procesados como primera toma que fue aproximadamente más de 138 personas encuestadas.

Figura 2: Datos para ser pre procesados



Desarrollo del momento 3, una vez revisado, analizado en un nivel de exhaustividad intermedia se obtiene una tabla de los diferentes modelos para PLN entre (Alberich, 2007), (Cortez, n.d, 2012.), y (Velázquez, 2001) que se presenta a continuación:

MODELO	NOMBRE	DESCRIPCION
PROBABILISTICO	Modelos n-gram	El modelo de lenguaje para la predicción de palabras usa las N-1 palabras anteriores para predecir la siguiente. Donde la memoria N define el número de probabilidades a ser estimadas ($= V N$), valores bajos de N son requeridos para obtener suficiente precisión dado un conjunto limitado de entrenamiento.

Tabla 2 Continuidad comparativa de los algoritmos del modelo probabilístico

MODELO	NOMBRE	DESCRIPCION
PROBABILISTICO	Modelo unigram	Se coloca una probabilidad $P(w)$ a cada palabra del léxico donde el modelo aduce que las palabras están elegidas independientemente, así que la probabilidad de una secuencia es el producto de la probabilidad de sus palabras dada por $\prod P(w_i)$. Se utiliza de manera aislada con poca frecuencia, debido a que la restricción de la probabilidad de una palabra no es muy útil por si sola.
PROBABILISTICO	Modelo bigram	Se asigna una probabilidad $\pi P(w_i/w_{i-1})$ a cada palabra dependiendo de la otra palabra anterior. Un modelo n-gram condiciona una palabra a las $N - 1$ anteriores, asignando una probabilidad: $\pi P(w_i/w_{i-1} \dots w_{i-N})$, está basado en la aproximación de que una palabra es estadísticamente dependiente de la palabra temporal anterior. Además se los utiliza para reconocer el habla.
PROBABILISTICO	Modelo trigram	Llamados trigramas, son un caso especial del N-gram, donde N es 3. Usualmente se usan en el procesamiento del lenguaje natural para hacer el análisis estadístico de los textos, son modelos utilizados para el reconocimiento del habla
PROBABILISTICO	Perplejidad	Es la probabilidad que tiene el conjunto de prueba, normalizada por la cantidad de palabras. Disminuir la perplejidad de un modelo es igual a maximizar su probabilidad total.

En vista que se tomará una dataset con los datos recopilados de conversaciones textuales, el proceso indica que luego esa información son clasificados para el modelamiento, ya que serán analizados estadísticamente en el habla escrito, o sea por el pensamiento, se sugiere tomar el modelo trigram y perplejidad en las pruebas de investigación de los diferentes entrenamiento.

Resultados

Después de las revisiones medianamente exhaustivas de las diferentes fuentes de información ya sea de hace más de 10 años atrás como de años recientes, se encontró que aún existe la problemática del manejo de texto por su ambigüedad y su dimensionalidad. Por tal razón tener claro la funcionalidad probabilística de los diferentes modelos para PLN fue esencial, dando a conocer que existe modelo n-gram, unigram, bigram, trigram y el de perplejidad. Esto se aplicaría en un análisis estadístico, en datos que fueron tomados de la data set de más de 138 personas que compartieron sus experiencias cuando fueron contagiados con covid-19 .

Referencias

1. Alberich, M. (2007). Procesamiento del Lenguaje Natural - Guía Introductoria. Guía Introductoria, 27. <https://www.cs.us.es/cursos/ia2/temas/tema-06.pdf>
2. Velázquez, J, S (2001, agosto 13). Análisis Robusto en Sistemas Conversacionales con Iniciativa Mixta. Disponible en http://catarina.udlap.mx/u_dl_a/tales/documentos/msp/velazquez_s_j/
3. Borja, M, Q. (2020). Análisis de las Herramientas de Procesamiento de Lenguaje Natural para estructurar textos médicos. <https://dadun.unav.edu/handle/10171/60003>
4. Contreras, H. Z. (2001). Procesamiento del Lenguaje Natural basado en una “gramática de estilos” para el idioma español. Facultad de Ingeniería, Colombia, Universidad de Los Andes., 54. http://www.saber.ula.ve/bitstream/123456789/13157/1/hc_propuestatesis.pdf
5. Cortez, A. (2012). 555-Texto del artículo-1119-1-10-20161126.pdf.
6. Dahl, V. (2010). Vista de An Introduction to Natural Language Processing_ the Main Problems-tabala comparativa.pdf.
7. Cerezo, H, C. (2020). Contribución a la Investigación de Nuevos Modelos de Extracción de Conocimiento sobre Sistemas Big Data. Disponible en <http://www.investigacion.biblioteca.uvigo.es/xmlui/handle/11093/1655>
8. Gomez, H, L (2019, septiembre 18). Agente conversacional para búsqueda y recomendación de ofertas laborales. 177. Disponible en <https://reunir.unir.net/handle/123456789/9564>

9. Guaman, W. E., Pi, P. F. M.-, Sof, P., & Iv, G. (2020). Agentes Virtuales en Dispositivos Móviles Virtual Agents on Mobile Devices Agentes virtuais em dispositivos móveis. 5(08), 544–560. Disponible en <https://dialnet.unirioja.es/servlet/articulo?codigo=7554393>
10. Egea, J. A. L. (2019). Guía para la publicación. Los estados de la Inteligencia Artificial. Capítulo 7-2º. Procesamiento del Lenguaje Natural (PLN). https://www.researchgate.net/publication/337172055_Procesamiento_del_Lenguaje_Natural_PLN_Natural_Language_Processing_NLP
11. GRAÑA GIL, Jorge. “Técnicas de análisis sintáctico robusto para la etiquetación del lenguaje natural”. Procesamiento del lenguaje natural. N° 28 (mayo 2002), pp. 117-118. Disponible en <http://rua.ua.es/dspace/handle/10045/1843>
12. Mutlu, E. C., Oghaz, T., Jasser, J., Tutunculer, E., Rajabi, A., Tayebi, A., Ozmen, O., & Garibay, I. (2020). A stance data set on polarized conversations on Twitter about the efficacy of hydroxychloroquine as a treatment for COVID-19. Data in Brief, 33. <https://doi.org/10.1016/j.dib.2020.106401>
13. Morales-Luna, G. (2013, mayo 7). El héroe Alan Turing: Aportaciones de la Criptología a la victoria aliada en la Segunda Guerra Mundial. Disponible en <http://cs.cinvestav.mx/~gmorales/12Enigma/enigma.pdf>
14. Birchenall, L. B., & Müller, O. (2014). La teoría lingüística de Noam Chomsky: del inicio a la actualidad. Lenguaje, 42(2), 417-442. Disponible en <http://www.scielo.org.co/pdf/leng/v42n2/v42n2a08.pdf>
15. Twitter (2017). Reglas de automatización. Disponible en <https://help.twitter.com/es/rules-and-policies/twitter-automation>
16. Suárez, A., G. (2014). La inteligencia artificial a través de sus científicos. Encuentros Multidisciplinares. Disponible en <https://repositorio.uam.es/handle/10486/678701>
17. Alberich, M. (2007). Procesamiento del Lenguaje Natural - Guía Introductoria. Guía Introductoria, 27. http://www.sopadebits.com/content/file_download/cms_file/file/58/pln-1.0-20070630.pdf

18. Da Silva, R. (2014). Los teoremas de incompletitud de Gödel, teoría de conjuntos y el programa de David Hilbert. *Episteme*, 34(1), 19–40. http://ve.scielo.org/scielo.php?script=sci_arttext&pid=S0798-43242014000100002
19. Orozco, S, O. (2009). Thomas Hobbes y la geometría del estado: anotaciones sobre el estado de naturaleza desde la historia de la ciencia. *Estudios de Filosofía (Medellin)*, 39, 153–175. Disponible en https://revistas.udea.edu.co/index.php/estudios_de_filosofia/article/view/12680
20. Reckoner, S. (1930). *Calculadoras mecánicas*. 1897. Valencia Giraldo, A. (2000). *Revista Facultad de Ingeniería. Leonardo Da Vinci*. *Revista Facultad de Ingeniería Universidad de Antioquia*, 7(32), 13–19. <https://www.redalyc.org/html/430/43003211/%0Ahttp://redalyc.uaemex.mx/src/inicio/ArtPdfRed.jsp?iCve=11400702>
21. Giudice, J. P. (2001). Complejidad y dimensiones en los estudios sobre Babbage: la máquina analítica. Un análisis del fracaso cultural del primer proyecto de calculadora digital programable secuencialmente. *Argumentos de Razón Técnica*, 2001,(4): 13-56.
22. Merodio, J. (2020, Junio 17). El matemático que soñaba con ordenadores en 1830. Disponible en <https://elpais.com/ciencia/2020-06-17/el-matematico-que-sonaba-con-ordenadores-en-1830.html>
23. Libertad Digital (2018, octubre 17). La inteligencia artificial toma el control de tu móvil. Disponible en <https://www.libertaddigital.com/ciencia-tecnologia/tecnologia/2018-10-17/la-inteligencia-artificial-toma-el-control-de-tu-movil-1276626583/>