e-ISSN: 1390-5902
CEDAMAZ Revista del Centro de Estudio y Desarrollo de la Amazonia , Vol. 11, No. 1, pp. 70–77, enero–junio 2021

Análisis de opinión sobre tuits del COVID-19 generados por usuarios ecuatorianos

Opinion analysis of COVID-19 tweets generated by Ecuadorian users

John A. Torres A.[1,*]

[1] *Carrera de Ingeniería en Sistemas/Computación,Universidad Nacional de Loja, Loja, Ecuador*
[*] *Autor para correspondencia: jatorresa@unl.edu.ec*

Fecha de recepción del manuscrito: 19/05/2021 Fecha de aceptación del manuscrito: 23/06/2021 Fecha de publicación: 15/07/2021

Resumen—Actualmente, se generan millones de datos por medio de la red social Twitter. El análisis de estos datos es fundamental e importante para examinar e investigar el conocimiento que se encuentra oculto entre estos. En este trabajo de investigación se realiza un análisis de opinión de tuits generados en Ecuador que tienen relación con el COVID-19 en el año 2020. Para ello, se utilizó la metodología Knowledge Discovery in Databases (KDD) para la gestión de los datos y para el descubrimiento de patrones ocultos en el conjunto de datos que tiene un total de 149.077 tuits. Se utilizaron varias herramientas para el Procesamiento del Lenguaje Natural, tales como: MeaningCloud, TextBlob, IBM Watson, Algoritmo Bayesiano (Creación Propia), Google Cloud Natural Language. Los clústeres generados presentaron la siguiente polaridad: 84.044 tuits positivos, 52.451 tuits negativos y 12.582 tuits neutros.

Palabras clave—Análisis de opinión, KDD, Minería de texto, Procesamiento del Lenguaje Natural.

Abstract—Currently, millions of data are generated through the social network Twitter. The analysis of this data is fundamental and important to examine and investigate the knowledge that is hidden among them. In this research work, an opinion analysis of tweets generated in Ecuador related to COVID-19 in the year 2020 is carried out. For this purpose, the Knowledge Discovery in Databases (KDD) methodology was used for data management and for the discovery of hidden patterns in the dataset that has a total of 149,077 tweets. Several tools were used for Natural Language Processing, such as MeaningCloud, TextBlob, IBM Watson, Bayesian Algorithm (Own Creation), Google Cloud Natural Language. The generated clusters presented the following polarity: 84,044 positive tweets, 52,451 negative tweets and 12,582 neutral tweets.

Keywords—Opinion Analysis, KDD, Text mining, Processing Natural Language.

INTRODUCCIÓN

La cantidad de datos producidos en la actualidad a nivel global es muy alta. Empresas, gobiernos, universidades y en general todas las organizaciones las producen a gran escala relacionados con sus actividades (Romero-Vega et al. 2021). Dichos datos se recopilan en grandes repositorios, principalmente en bases relacionales que permiten el almacenamiento de la información (Méndez et al., 2018).

Los datos se encuentran almacenados en diferentes repositorios o en la nube, es decir, se encuentran a la espera de ser analizados por la comunidad científica o por personas de la academia que tengan como necesidad el descubrimiento del conocimiento oculto entre los datos (Aldana et al., 2018). La red social Twitter cuenta con una gran cantidad de usuarios y durante la pandemia del COVID-19 se han generado un gran número de tuits. Sin embargo, esta situación ha generado la aparición de comentarios con información de contenido dudoso, lo cual provoca diferentes emociones en la audiencia. En cuanto a la propagación de este contenido se puede evidenciar que los textos contienen opiniones de diferente polaridad. Particularmente en Ecuador, los tuits relacionados con el COVID-19 generan controversia entre los usuarios y las opiniones emitidas causan diferentes reacciones en la red (Álvarez et al., 2020).

Este articulo presenta un análisis de opinión con respecto a tuits generados por causa del COVID-19. Los datos analizados fueron datos no estructurados (texto) y se delimitan a presentarse en idioma español. El objetivo de este trabajo fue descubrir patrones en los tuits utilizando técnicas de procesamiento de lenguaje natural y obtener resultados que permitan realizar un análisis de lo que opinan las personas sobre este tema (Alonso et al., 2018). Este análisis de datos fue realizado únicamente para tuits que tenían como origen el país de Ecuador en los meses comprendidos entre abril del 2020 y noviembre del 2020.

Errecalde y otros (Errecalde et al., 2017), recalcan la necesidad de analizar el material disponible en estos medios sociales con el propósito de conocer la personalidad, relaciones interpersonales y la situación social de las personas. En concordancia con la investigación de (Morales et al., 2016), existe la necesidad de hacer minería de opinión en Twitter porque permite cuantificar el interés y la opinión de los usuarios sobre un tema o caso de estudio específico. El resto del artículo está organizado de la siguiente manera. La Sección II describe el equipo y el software utilizados en la investigación y la metodología aplicada en cada etapa del proceso KDD. La Sección III presenta hallazgos basados en la metodología utilizada. La discusión y la importancia de las contribuciones a la ciencia se presentan en la Sección IV. En la Sección V, se presentan las conclusiones obtenidas del trabajo realizado. La Sección VI, presenta el debido reconocimiento a las personas que aportaron significativamente al desarrollo de la investigación. El financiamiento del trabajo realizado se menciona en la Sección VII. Finalmente, la Sección VIII muestra las fuentes bibliográficas que sustentan académicamente el trabajo realizado.

Materiales y métodos

Para realizar el trabajo de análisis de opinión se tomó como referencia la Metodología KDD que plantea las siguientes etapas: selección de los datos, preprocesamiento de datos, transformación de datos, minería de datos e interpretación de datos. Los lenguajes de programación seleccionados fueron R (IDE RStudio) y Python (IDE Jupyter) para cumplir con las etapas de minería e interpretación de los datos. También, se escogió la herramienta OpenRefine para llevar a cabo las etapas de preprocesamiento y transformación de los datos.

En la etapa de selección de los datos se estableció los criterios de búsqueda y selección para obtener un Dataset acorde al objeto de estudio. En primera instancia, se usó el Api Rest gratuita de Twitter por medio del Lenguaje R, en donde, se logró obtener 3.479 tuits. Posteriormente, se utilizó el motor de búsqueda de Google (Data Search) y se logró obtener un Dataset proporcionado por: Digital Narratives of COVID-19 de la Universidad Miami (EEUU) en colaboración con el Consejo Nacional de Investigaciones Científicas y Técnicas de Argentina (CONICET); este Dataset contiene 206.174 tuits el cual fue seleccionado debido al volumen y calidad que presenta.

Para las etapas de procesamiento y transformación de los datos se utilizó la herramienta OpenRefine, en donde, se aplicaron criterios de limpieza, reducción y clustering de datos. En los criterios de limpieza se eliminó símbolos, números y patrones (URL) además, se realizó conversiones básicas como: mayúsculas a minúsculas, palabras con tilde a sin tilde y unificación de espacios. En la reducción de los datos de eliminó los tuits que presentaban duplicidad. Finalmente, se aplicó facetas propias de la herramienta para obtener los múltiples clustering con la finalidad de unificar los datos en base a sus lexemas.

En la etapa de minería de datos se realizó la codificación de un algoritmo bayesiano, por ello, se utilizó el IDE RStudio para detallar el funcionamiento del algoritmo y presentar la configuración utilizada. Se utilizó como fuente de conocimiento la unificación de los lexicones iSol, ML-SentiCon y Tass con la finalidad de optimizar los resultados. Como paso final de esta etapa el algoritmo fue validado por medio de una comparativa de herramientas similares, asimismo, validado por medio de la Matriz de Confusión y la curva ROC.

Finalmente, en la etapa de interpretación de datos se utilizó el IDE Jupyter para visualizar los clustering generados en la etapa anterior. Los gráficos estadísticos generados se relacionan con la Frecuencia de palabras, Nube de palabras, Dispersión y Regresión Lineal, Parcela de cajas y Reglas de asociación.

A continuación, la Figura 1 presenta el trabajo realizado en cada etapa de la metodología KDD y la secuencia ingenieril que se utilizó para cumplir con el objetivo planteado en la investigación.

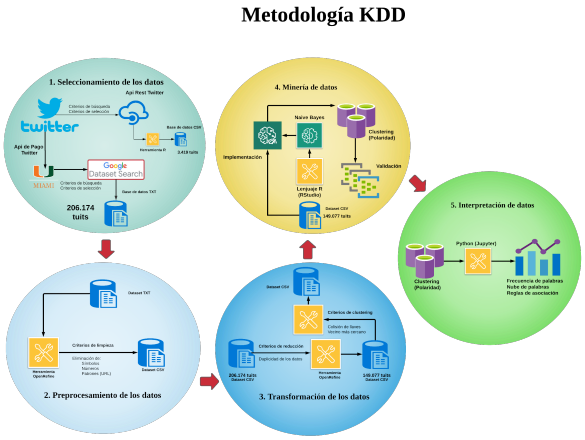

Fig. 1: Metodología utilizada en la investigación para encontrar el conocimiento oculto (Dataset).

Resultados

Fase 1: Selección de los datos

En esta fase se hizo uso del API REST TWITTER, una API de consumo gratuito que permite obtener tuits de acuerdo a etiquetas de hashtag “COVID19 and Ecuador”. Realizada la consulta y obtenido el Dataset se observó que la información se limita a presentar opiniones generadas hace ocho días; es decir, no es permitido acceder a tuits que fueron generados en fechas anteirores. Por ello, el Dataset obtenido por medio de este proceso, no fue utilizado en el desarrollo de la investigación. Debido al objeto de estudio que se da en meses anteriores, en donde, se generó abundantes tuits con los primeros casos de COVID-19 en Ecuador.

Se hizo uso del motor de búsqueda proporcionado por Google (Data Search) que se dedica específicamente a proveer Dataset de diversos campos de investigación, en efecto, se encontró un repositorio digital denominado “Digital

Narratives of COVID-19[1]" que cumple con los criterios de búsqueda (Idioma: español, Localización: Ecuador) y criterios de selección (Objeto de estudio: COVID-19, Tipo de dato: texto).

Digital Narratives of COVID-19[2] es un proyecto planteado por la Universidad Miami (EEUU) en colaboración con el Consejo Nacional de Investigaciones Científicas y Técnicas de Argentina (CONICET). A continuación, la Figura 2 presenta un Dashboard en donde se observa de forma general el volumen de los datos generados en relación con los meses.

El Dataset obtenido presenta el registro histórico de los datos que equivale a los meses de: abril - noviembre del año 2020. Es decir, los Dataset descargados son idóneos para el trabajo de investigación.

Situándonos específicamente en Ecuador, la cuarentena inició el 12 de marzo del 2020 y es conveniente analizar los datos que se aproximen lo máximo posible a los meses en donde se evidenció mayor volumen. Por otra parte, de acuerdo a la Figura 2 el Dataset creado evidencia un total de 206.174 tuits con un tamaño de 32.8 MB, por esta razón, fue seleccionado para este trabajo de investigación.

Fase 2: Preprocesamiento de los datos

En esta etapa se realizó la limpieza de los datos por medio de operaciones de eliminación y conversión. El objetivo de esta etapa es tratar de limpiar los datos que contienen ruido. Es decir, trata de suprimir aquellos datos que se encuentran distantes del rango de valores esperados, contienen errores humanos o son irrelevantes de acuerdo con el objeto de estudio. En el siguiente apartado, se presenta los criterios de limpieza que se aplicó al Dataset.

Para emplear los criterios de limpieza se utilizó la herramienta OpenRefine y se hizo uso de la función Transformaciones Comunes y Transformar para eliminar símbolos, números y patrones que no eran significantes en el estudio. Además, se realizó sustituciones básicas a nivel léxico, por ejemplo: conversiones de mayúsculas a minúsculas, exclusión de tildes y unificación de espacios consecutivos.

Fase 3: Transformación de los datos

En esta etapa se realizó la reducción de los datos por medio de operaciones de duplicidad de los datos y faceta de texto (clustering). El objetivo de esta etapa fue de reducir los datos que contienen duplicidad de acuerdo a las filas. Asimismo, se trató de consolidar los datos de acuerdo a las columnas.

La herramienta seleccionada (OpenRefine) sirvió para limpiar o reducir el Dataset, en este caso, se utilizó el menú de faceta personalizada para aplicar la faceta por duplicados. Esta faceta permitió eliminar aquellos registros duplicados (retuits) de acuerdo a las filas. De esta manera, se pretende omitir un análisis de datos de tuits generados por posibles BOTS.

Tabla 1: Método y algoritmos utilizados para la lematización.

Método	Función	Configuración
Colisión de llaves	Huella digital	Patrón llave
Colisión de llaves	Huella del n-grama	Tamaño del n-grama 2-6
Colisión de llaves	Metaphone3	Patrón llave
Colisión de llaves	Cologne-phonetic	Patrón llave
Colisión de llaves	Daitch-Mokotoff	Patrón llave
Colisión de llaves	Beider-Morse	Patrón llave
Vecino más cercano	Levenshtein	Radio (1.0) Bloque (1-6)
Vecino más cercano	Ppm	Radio (1.0) Bloque (1-6)

El resultado que se obtuvo fue de 149.077 filas únicas y de 57.097 filas duplicadas, es decir, el Dataset se redujo en un 28 % de acuerdo a las filas. Sin embargo, fué necesario consolidar la base de datos de acuerdo a las columnas por lo que se utilizó la faceta de texto para aplicar los criterios de clustering.

Es importante mencionar que después de haber aplicado la limpieza anterior al Dataset se procedió a dividir en varias columnas el área de trabajo. Con la finalidad de utilizar la función Faceta de texto que permite agrupar las palabras de acuerdo a su frecuencia. Por ello, es posible consolidar los datos con un enfoque agrupación por lexemas y de corrección ortográfica conforme con los clustering encontrados. A continuación, la Tabla 1 presenta los métodos y algoritmos propios de la herramienta OpenRefine que se utilizaron para realizar el trabajo de lematización.

Fase 4: Minería de datos

En esta etapa se codificó un Algoritmo Bayesiano y algunos recursos que se encuentran disponibles para clasificar opiniones en el idioma español. En consecuencia, se realizó un diseño muestral del Dataset previamente procesado con el objetivo de realizar el estudio en el menor tiempo posible. La muestra estadística calculada es una parte o una porción del Dataset que permitió conocer la calidad de la clasificación realizada; la mencionada muestra fue contrastada con los resultados obtenidos de recursos o herramientas similares que permiten clasificar las opiniones.

Para calcular el tamaño de la muestra, se utilizó la siguiente fórmula[3] :

[1]DIGITAL NARRATIVES
[2]COVID-19
[3]Minería de Datos

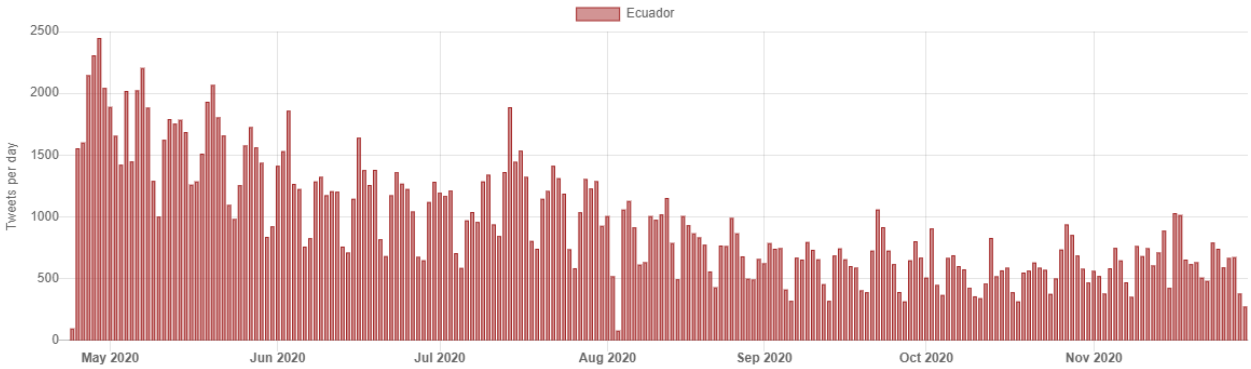

Fig. 2: Dashboard (Volumen de datos vs Meses).

$$n = \frac{(Z^2 pqN)}{(Ze^2 + Z^2 pq)}$$

Es importante resaltar que una vez que se obtenida la muestra (138 objetos) se contó con la ayuda de un Especialista en Lengua y Literatura para una clasificación manual de las opiniones. Esto se estableció como primera instancia para realizar un análisis comparativo con los resultados obtenidos del Algoritmo Bayesiano y demás recursos utilizados. La muestra en este caso se la obtuvo del resultado obtenido en la implementación del algoritmo, por esta razón, la muestra utilizada se encuentra previamente etiquetada y balanceada según la polaridad del sentimiento.

La Tabla 2 detalla los resultados cuantitativos que se obtuvo de las diferentes herramientas que permitieron clasificar las opiniones con respecto a su polaridad sentimental.

De acuerdo a la ISO 19157[4] , una de las medidas para asegurar la calidad de los datos es la conocida matriz de confusión y la curva ROC; son tablas de contingencia que sirven como herramientas de estadística para el análisis de observaciones emparejadas; adicionalmente ofrece una visión completa de la distribución de los aciertos y errores entre clases (Cumbicus et al., 2019).

La Figura 3 presenta la matriz de confusión que se obtuvo de acuerdo al contraste realizado entre la clasificación manual (human evaluation) realizada por el Especialista (columnas) y el desempeño del Algoritmo Bayesiano (filas). Los resultados obtenidos muestran lo siguiente: la métrica Recall o sensibilidad, evidencia que aproximadamente el 80 % de los casos positivos fueron correctos. La métrica Precisión, muestra de forma abstracta un bajo coeficiente de dispersión de los datos, en consecuencia, presenta un porcentaje alto de 94 % de precisión. La métrica Accuracy o exactitud, señala que en un 70 % de los casos probables es posible que se obtenga una clasificación correcta. Por ello, se logró determinar que el Algoritmo Bayesiano (Creación Propia) es un modelo confiable.

La Figura 4 muestra la curva ROC que se basa en la comparación de los dos modelos de clasificación: Algoritmo

[4]ISO 19157

Fig. 3: Matriz de confusión – Validación del algoritmo

Bayesiano y Especialista (human evaluation).

Esta Figura en particular, especifica el intervalo de confianza que se promedia en 77,5 % con un rango de alcance entre: 67,8 % y 87,2 %, lo cual, determina que es un modelo aceptable.

Fase 5: Interpretación de datos

El algoritmo de acuerdo a la validación realizada se sitúa como mejor tercero en comparativa con otras herramientas para el PLN como son: Google Cloud Natural Language, IBM Watson, MeaningCloud, Textblob entre otros. Adicionalmente, conforme a la matriz de confusión presenta un 94 % de Precisión, 80 % de Recall o Sensitivity y demás métricas estadísticas, asimismo, acorde con la curva ROC muestra un intervalo de confianza que se promedia en 77,5 %. Por estas razones, se argumenta que el modelo es confiable y los resultados evidenciados de igual forma.

Se seleccionó la herramienta IDE Jupyter (Python) y RStudio (R), con el objetivo de visualizar de mejor manera los clústeres y descubrir patrones que a simple vista no se

e-ISSN: 1390-5902
CEDAMAZ Revista del Centro de Estudio y Desarrollo de la Amazonia , Vol. 11, No. 1, pp. 70–77, enero–junio 2021

Tabla 2: Contraste de los resultados obtenidos con respecto a los múltiples recursos utilizados.

Herramienta	Descripción	Pos	Neg	Neu	Cal
Human Evaluation	Profesional en Lengua y Literatura	52	54	32	10
MeaningCloud	Aplicación web con API REST	46	60	32	9,63
Algoritmo Bayesiano	Diseño y codificación propios de autor	46	46	46	9,12
TextBlob	Librería	48	37	53	8,71
IBM Watson	API REST	20	61	57	7,96
LexalTics	Aplicación web con API REST	23	46	69	7,63
Orange-multilingual	Software libre	28	39	71	7,53
Google Cloud Natural Language	API REST	46	13	79	7,06
ParallelDots	API REST	22	11	105	5,41
TheySay	API REST	19	13	106	5,36
NLTK	Librería	4	25	109	5,16
CoreNLP	API REST	1	18	119	4,52
Sentiment140	Aplicación web con API REST	1	5	132	3,72

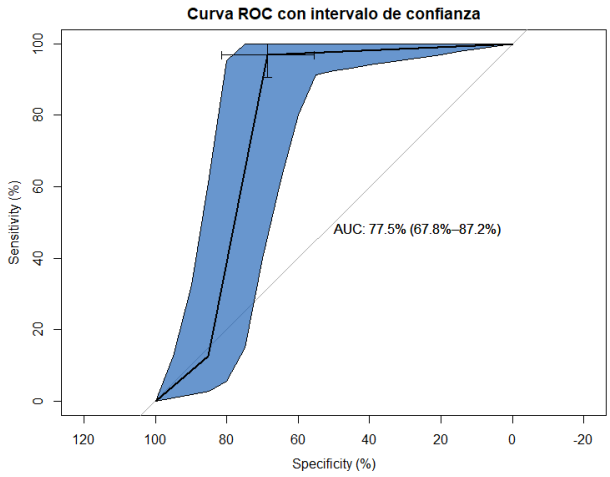

Fig. 4: Curva ROC – Validación del algoritmo

pueden apreciar. De esta manera, se descubrió el conocimiento que se encuentra inmerso en el conglomerado de datos.

La Figura 5 presenta el Histograma de acuerdo a la categoría Mejor Ajuste (Resultado Final) del Algoritmo Bayesiano. En la barra verde (Positivo): 84.044 tuits, barra naranja (Negativo): 52.451 tuits y en barra azul (Neutro): 12.582 tuits.

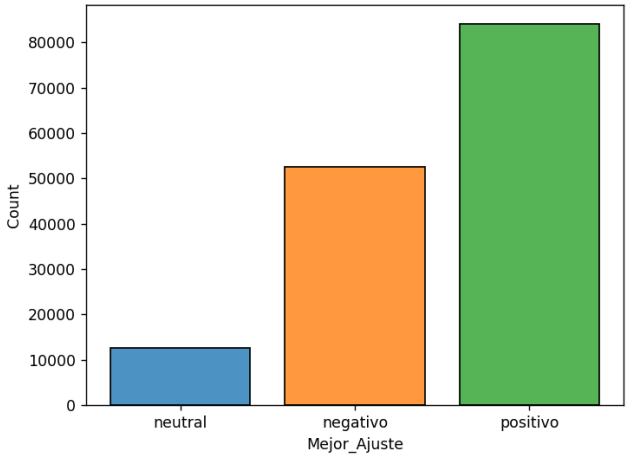

Fig. 5: Histograma de los clústeres generados

Análisis clúster positivo

La Figura 6 presenta la reconocida nube de palabras y la frecuencia de palabras (Clúster Positivo). Estas graficas se encuentran estrechamente relacionadas y exponen de forma simple los patrones en cuanto a las palabras más usadas para referirse al COVID-19. Además, se observa la exclusión de la palabra “COVID” porque es obvio en el campo de estudio analizado. Sin embargo, para asegurar la relación de las palabras e inferir en el conocimiento de forma técnica, más adelante, se presenta la Figura de Reglas de Asociación con sus respectivas premisas.

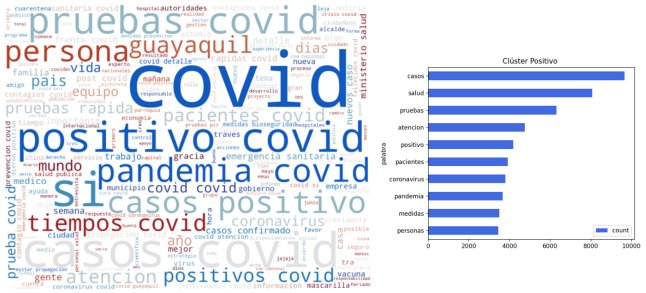

Fig. 6: Nube de palabras y frecuencia de palabras (Clúster Positivo)

La Figura 7 presenta las reglas de asociación (Apriori) generadas con una confianza del 90 % y un soporte de 0.009.

Finalmente, de acuerdo con la Figura anterior se puede inferir las premisas más relevantes que son las siguientes: Pruebas Rápidas para el COVID, Personal de Salud por COVID, Pruebas de Salud por COVID, Atención a Casos por COVID, Frente al COVID, Prevención para COVID, Propagación del COVID, Contra COVID, Pacientes por COVID, Gracias COVID, Evitar el COVID, Pcr COVID y Quédate en casa por COVID.

Análisis clúster negativo

La Figura 8 presenta la conocida nube de palabras y la frecuencia de palabras con respecto al clúster Negativo.

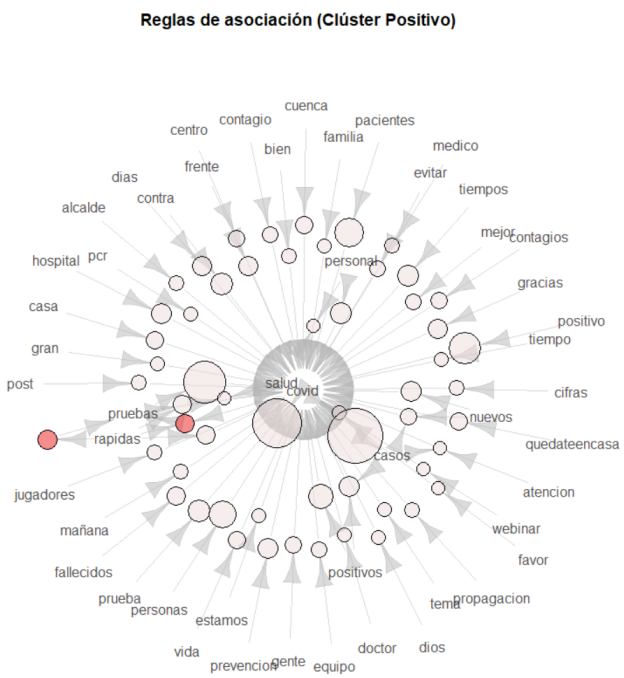

Fig. 7: Reglas de asociación (Clúster Positivo)

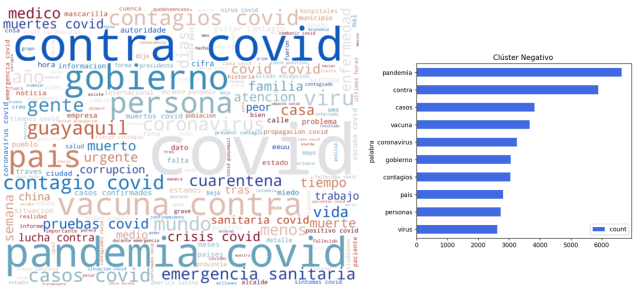

Fig. 8: Reglas de asociación (Clúster Negativo)

Posteriormente, la Figura 9 presenta las reglas de asociación que se obtuvo con una confianza del 90 % y un soporte de 0.009.

Para finalizar, de acuerdo con la Figura anterior se puede inferir las premisas más importantes conforme a la relación apriori que existe entre las palabras, a continuación, se presentan las siguientes: Positivo para COVID, Caso confirmado por COVID, Corrupción por COVID, Miedo del COVID, Síntomas de COVID, Vacuna Contra el COVID, Muerte por COVID, Enfermedad por COVID, Lucha Contra el COVID, Cuarentena por COVID, Crisis por COVID y Gobierno del COVID.

Análisis clúster neutro

En este clúster se evidenció opiniones tanto positivas como negativas, es decir, son opiniones que el Algoritmo Bayesiano matemáticamente clasificó como neutrales debido al balance realizado con respecto a los diccionarios utilizados (iSOL[5] , ML-SentiCon[6] , TASS[7]). Además, se evidenció patrones de sarcasmo y doble negatividad por

[5]iSOL
[6]ML-SentiCon
[7]TASS

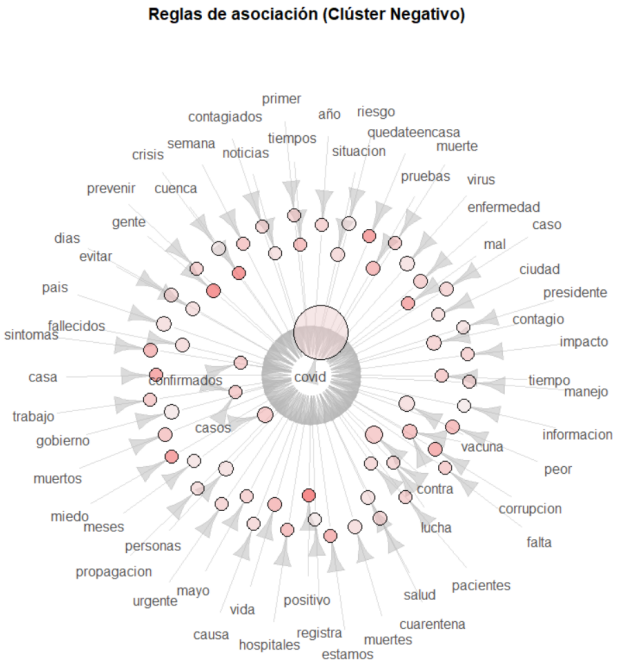

Fig. 9: Reglas de asociación (Clúster Negativo)

lo que indica que el algoritmo en este sentido presentó dificultades para clasificar en su polaridad respectiva. Por ello, se considera que un análisis con respecto a la nube de palabras, frecuencia de palabras y reglas de asociación es irrelevante.

Finalmente, en el siguiente enlace electrónico [8] se comparte el código fuente desarrollado en lenguaje R y Python. Además, se encuentran los Dataset referentes al objeto de estudio (Tweets) y el tratamiento de dichos datos con respecto a la metodología KDD (resultados obtenidos

DISCUSIÓN

De acuerdo con la etapa de selección de datos, los datos que se obtuvieron fueron proporcionados por la Universidad de Miami en participación con el Consejo Nacional de Investigadores Científicos y Técnicos de Argentina (CONICET) lo que otorga confianza y calidad al Dataset. Por otra parte, hay algunas alternativas para obtener tuits de la red social Twitter, por ejemplo, la librería gratuita Textblob proporcionada por Python; permite acceder a este tipo de datos y realizar estos tipos de trabajos relacionadas a la minería de opinión. Sin embargo, se seleccionó este último Dataset debido a la confiabilidad académica de su procedencia y al acceso gratuito de los datos.

En la etapa de preprocesamiento de los datos se aplicó criterios de limpieza, por ejemplo, eliminación de números, símbolos y patrones que no aportan de forma significativa en materia de analítica de opinión. Este procedimiento no presentó ninguna dificultad, puesto que, los Dataset que se descargó en la etapa anterior exponen una pre-limpieza. Es decir, se observó que los tuits contenían caracteres alfanuméricos y que la mayoría de emoticones fueron omitidos

[8]Repositorio Web

previamente, no obstante, aquellos emoticones restantes fueron reemplazados por su respectivo identificador.

En la etapa de transformación de los datos se aplicó criterios de reducción y de clustering. En relación con la reducción, se suprimió aquellos datos que se encontraban duplicados y se conservó una sola muestra, por lo cual, se redujo la base de datos a 149.077 tuits. Adicionalmente, con respecto al clustering se aplicó homogeneidad a los datos por medio de facetas de texto; dichas facetas utilizan técnicas propias de Inteligencia Artificial, por ejemplo, colisión de llaves y vecino más cercano. En consecuencia, se logró lematizar la base de datos y mejorar su calidad.

No obstante, en la actualidad existen diversas herramientas que permiten realizar el trabajo de la limpieza de los datos, tales como: Python, R, Orange, Tanagra, Rapid Minner, ROctave, Scavis entre otros. Sin embargo, se seleccionó OpenRefine debido a la interfaz gráfica (sencillez) que presenta para manipular los datos y permite utilizar el lenguaje de programación Python para la limpieza, además, existe una extensa documentación para utilizar la herramienta OpenRefine y Python. En la etapa de minería de datos se extrajo una muestra del conglomerado de datos con el objetivo de contrastar los resultados con otras herramientas que realizan el trabajo de análisis de opinión y con una clasificación manual realizada por un experto. En efecto, el algoritmo propuesto se ubicó en tercer lugar después de la revisión manual y de la herramienta MeneaningCloud.

El modelo propuesto fue validado por medio de la matriz de confusión y por la curva ROC por este motivo, las métricas estadísticas evidencian una Precisión del 94 %, Recall de 80 % y una media de intervalo de confianza calculada en 77.5 %. Se puede concluir que el modelo es confiable de acuerdo con los resultados analizados, pero carece de memoria con respecto a la técnica utilizada (Naive Bayes), por tal motivo, técnicas como Aprendizaje Automático o Redes Neuronales pueden presentar a futuro mejores resultados.

Finalmente, en la etapa de interpretación de los datos se utilizó la herramienta Jupyter para analizar los clústeres generados (positivo, negativo y neutro) y se realizó gráficos con el objetivo de inferir en el conocimiento oculto, por ejemplo, Correlación y dispersión lineal, Frecuencia de palabras, Nube de palabras, Parcela de cajas y Reglas de asociación. Sin embargo, cuestionando el avance de la ciencia de los datos en materia de análisis de sentimiento, se observa que el progreso es limitado en el manejo del idioma español debido a su complejidad. Puesto que, a nivel sintáctico la mayoría de herramientas que existen actualmente no presentan resultados similares, por otra parte, a nivel semántico y pragmático se espera proporcionar posibles soluciones a futuro.

Conclusiones

La herramienta Google Data Search, permitió obtener un conjunto de datos de excelente calidad para llevar a cabo el análisis de opinión. Gracias a él se logró encontrar de forma organizada y gratuita diversos conjuntos de datos sobre diversos temas de interés colectivo (COVID-19). Entre estos se pueden mencionar datos gubernamentales, datos de organizaciones de noticias o instituciones universitarias como Harvard, Miami y el MIT, entre otros.

De acuerdo con las etapas de preprocesamiento de los datos y transformación de los datos (KDD). La herramienta OpenRefine ayudó a limpiar la base de datos en forma sencilla y visual, además, permitió aplicar los criterios de duplicidad de los datos disminuyendo el tamaño de la misma en un 27,69 %. Así mismo, ayudó con la aplicación de los criterios de clustering para unificar la base de datos en conformidad con la lematización.

El Algoritmo Naive Bayes propuesto presentó los mejores resultados con la siguiente configuración: probabilidad previa (1.0), probabilidad débil (1.0) y probabilidad fuerte (0.5), la cual, clasificó los tuits de la siguiente manera: 84.044(positivo), 52.451(negativo) y 12.582 (neutro). El algoritmo se delimita a su capacidad de clasificar debido a su simplicidad, por lo que, carece de recuerdo. En consecuencia, para inferir de mejor manera la polaridad de las opiniones se utilizó como fuente de conocimiento los lexicones Isol, ML-SentiCon y Tass.

El modelo bayesiano expuesto fue validado por medio de la Matriz de Confusión y por la curva ROC por esta razón, las métricas estadísticas evidencian una Precisión del 94 %, Recall de 80 % y un intervalo de confianza comprendido entre 67.8 % y 87.2 %. Adicionalmente, se obtuvo una muestra del Dataset en general con el objetivo de comparar los resultados con otras herramientas similares, por ejemplo, clasificación manual (human evaluation), MeneaningCloud, Textblob, Google Cloud Natural Language, IBM Watson, CoreNLP, NLTK entre otras más. Por esta razón, el algoritmo se posesionó en tercer lugar con una calificación equivalente a 9,12/10 puntos.

Agradecimientos

Agradezco al Licdo. José Manuel Padilla Puchaciela por su guía y ayuda durante esta investigación en calidad de experto (Profesional en Lengua y Literatura – Postgrado en Edición de medio impresos).

Contribuciones de los Autores

Conceptualización JT, Curación de datos JT, Análisis formal JT, Investigación JT, Metodología JT, Redacción - borrador original JT, Redacción - revisión y edición JT.

Financiamiento

El presente estudio fue financiado por el autor y la Universidad Nacional de Loja, bajo resolución 250-2021-DI-UNL.

Referencias

Aldana, H. S. M., Rivas, J. D. C., Hidalgo, J. M. V. (2018). Big Data, el futuro de las predicciones certeras. Revista

Avenir, 2(2), 10-16.

Alonso-Arévalo, J., Vázquez Vázquez, M. (2016). Big Data: la próxima gran cosa en la gestión de la información.

Álvarez Sarmiento, K. L. (2020). Investigación y análisis de herramientas para extracción de Tweets sobre COVID19 focalizadas en RStudio y Python que permitan crear una base de datos relacional (Doctoral dissertation, Universidad de Guayaquil. Facultad de Ciencias Matemáticas y Físicas. Carrera de Ingeniería en Networking y Telecomunicaciones).

Anual P. Aplicación del proceso de descubrimiento del conocimiento para la detección de diabetes (2020). In: 10° Congr Int Comput México - Colomb.;10(ISSN 2462-9588):234.

Arroyo Laimito, K. F. (2020). Desarrollo de un sistema de análisis de datos mediante la metodología Knowledge Discover Database para el procesamiento de información en la determinación de estrategias de salud pública nutricional. Univ Nac del Cent del Perú;(064). http://repositorio.uncp.edu.pe/handle/UNCP/5781

Cortez Reyes, R. A. (2018). Extracción de conocimiento a partir de textos obtenidos de Twitter.(65):30-41.

Cumbicus-Pineda O.M., Ordoñez-Ordoñez P.F., Neyra-Romero L.A., Figueroa-Diaz R. (2019). Automatic Categorization of Tweets on the Political Electoral Theme Using Supervised Classification Algorithms. In: Botto-Tobar M., Pizarro G., Zúñiga-Prieto M., D'Armas M., Zúñiga Sánchez M. (eds) Technology Trends. CITT 2018. Communications in Computer and Information Science, vol 895. Springer, Cham. https://doi.org/10.1007/978-3-030-05532-$5_5$1.

Del Alcazar Ponce JP. (2019). Consultoría de marketing, clientes, innovación y planificación. Published 2019. Accessed July 2, https://www.formaciongerencial.com/

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., Seligman, M. E. (2015). Psychological language on Twitter predicts county-level heart disease mortality. Psychological science, 26(2), 159-169.

Jiménez-Zafra SM. (2017). Detección de la negación en textos en español y aplicación al Análisis de Sentimientos. CEUR Workshop Proc.,1961.

Lakshmi, P. V., Shwetha, G., Raja, N. S. M. (2017, March). Preliminary big data analytics of hepatitis disease by random forest and SVM using r-tool. In 2017 Third International Conference on Biosignals, Images and Instrumentation (ICBSII) (pp. 1-5). IEEE.

López Pedraza FJ, González Macías M del C, Sandoval García Edgar R. (2019). Minería de Datos: Identificando causas de deserción en las Instituciones Públicas de Educación Superior de México. TiES, Rev Tecnol e Innovación en Educ Super.;1(2):1-12. http://www.ties.unam.mx/

Martín Morales, S. (2016). Análisis de información proveniente de redes sociales como Twitter (Bachelor's thesis).

Méndez, N. P., Rubier, J. P. (2018). Ciencia de datos: una revisión del estado del arte. UCE Ciencia. Revista de postgrado, 6(3).

Olarte, E., Panizzi, M. D., Bertone, R. A. (2018). Segmentación de mercado usando técnicas de minería de datos en redes sociales. In XXIV Congreso Argentino de Ciencias de la Computación (La Plata, 2018).

Romero-Vega R.R., Cumbicus-Pineda O.M., López-Lapo R.A., Neyra-Romero L.A. (2021). Detecting Xenophobic Hate Speech in Spanish Tweets Against Venezuelan Immigrants in Ecuador Using Natural Language Processing. In: Botto-Tobar M., Montes León S., Camacho O., Chávez D., Torres-Carrión P., Zambrano Vizuete M. (eds) Applied Technologies. ICAT 2020. Communications in Computer and Information Science, vol 1388. Springer, Cham. https://doi.org/10.1007/978-3-030-71503-824

Sharmin, S., Zaman, Z. (2017, December). Spam detection in social media employing machine learning tool for text mining. In 2017 13th International Conference on Signal-Image Technology Internet-Based Systems (SITIS) (pp. 137-142). IEEE.

Symeonidis S, Effrosynidis D, Arampatzis A. (2018). A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. Expert Syst Appl. 110:298-310. doi:10.1016/j.eswa.2018.06.022

Viegas, F., Rocha, L., Resende, E., Salles, T., Martins, W., e Freitas, M. F., Gonçalves, M. A. (2018). Exploiting efficient and effective lazy Semi-Bayesian strategies for text classification. Neurocomputing, 307, 153-171.