



UNIVERSIDAD POLITÉCNICA SALESIANA
SEDE GUAYAQUIL
CARRERA DE INGENIERÍA DE SISTEMAS

TEMA:

ANÁLISIS Y EVALUACIÓN DE LA TÉCNICA DE PROCESAMIENTO DE LENGUAJE
NATURAL AUTOMÁTICO SUPERVISADO PARA DETERMINAR LA POLARIDAD DE
UN TEXTO NO ESTRUCTURADO EN REDES SOCIALES

Trabajo de titulación previo a la obtención del
Título de Ingeniero de Sistemas

AUTOR:

Jeampier Alexander Carriel Roca

TUTOR:

Miguel Ángel Quiroz Martínez

Guayaquil – Ecuador

2022

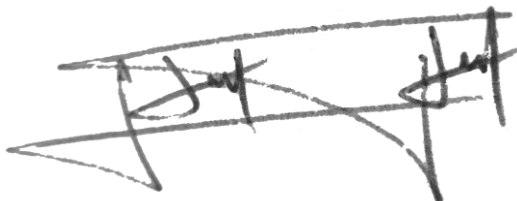
**CERTIFICADO DE RESPONSABILIDAD Y AUTORÍA DEL
TRABAJO DE TITULACIÓN**

Yo, Jeampier Alexander Carriel Roca documento de identificación N°
1250025879 manifiesto que:

Soy el autor y responsable del presente trabajo; y, autorizo a que sin fines
de lucro la Universidad Politécnica Salesiana pueda usar, difundir,
reproducir o publicar de manera total o parcial el presente trabajo de
titulación.

Guayaquil, 07 de febrero del año 2022

Atentamente,



Jeampier Alexander Carriel Roca

C.I. 1250025879

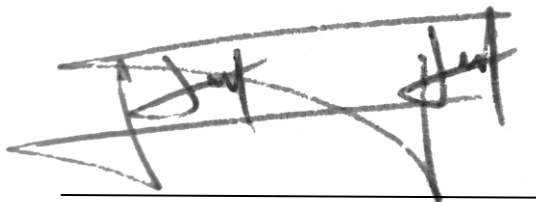
**CERTIFICADO DE CESIÓN DE DERECHOS DE AUTOR DEL
TRABAJO DE TITULACIÓN A LA UNIVERSIDAD
POLITÉCNICA SALESIANA**

Yo, Jeampier Alexander Carriel Roca con documento de identificación No. 1250025879, expreso mi voluntad y por medio del presente documento cedo a la Universidad Politécnica Salesiana la titularidad sobre los derechos patrimoniales en virtud de que soy autor del Artículo académico: “ANÁLISIS Y EVALUACIÓN DE LA TÉCNICA DE PROCESAMIENTO DE LENGUAJE NATURAL AUTOMÁTICO SUPERVISADO PARA DETERMINAR LA POLARIDAD DE UN TEXTO NO ESTRUCTURADO EN REDES SOCIALES”, el cual ha sido desarrollado para optar por el título de: Ingeniero de Sistemas, en la Universidad Politécnica Salesiana, quedando la Universidad facultada para ejercer plenamente los derechos cedidos anteriormente.

En concordancia con lo manifestado, suscribo este documento en el momento que hago la entrega del trabajo final en formato digital a la Biblioteca de la Universidad Politécnica Salesiana.

Guayaquil, 07 de febrero del año 2022

Atentamente



Jeampier Alexander Carriel Roca

C.I. 1250025879

**CERTIFICADO DE DIRECCIÓN DEL TRABAJO DE
TITULACIÓN**

Yo, Miguel Ángel Quiroz Martínez con documento de identificación N° 0922799655, docente de la Universidad Politécnica Salesiana, declaro que bajo mi tutoría fue desarrollado el trabajo de titulación: ANÁLISIS Y EVALUACIÓN DE LA TÉCNICA DE PROCESAMIENTO DE LENGUAJE NATURAL AUTOMÁTICO SUPERVISADO PARA DETERMINAR LA POLARIDAD DE UN TEXTO NO ESTRUCTURADO EN REDES SOCIALES, realizado por Jeampier Alexander Carriel Roca con documento de identificación N° 1250025879, obteniendo como resultado final el trabajo de titulación bajo la opción Artículo Académico que cumple con todos los requisitos determinados por la Universidad Politécnica Salesiana.

Guayaquil, 07 de febrero del año 2022

Atentamente,

Miguel Quiroz Martínez

Docente Miguel Ángel Quiroz Martínez
0922799655

DEDICATORIA

Dedico este trabajo a mis padres Daniel Carriel y Elizabeth Roca quienes con su amor paciencia y arduo trabajo me permitieron hoy cumplir otro sueño, gracias por inculcar un ejemplo de esfuerzo y valentía sin miedo a la adversidad Porque Dios siempre está conmigo.

A toda mi familia porque con sus oraciones consejos y ánimos me han hecho mejor persona y de una forma u otra forma me acompañan en todos mis sueños y metas. Finalmente quisiera dedicar esta tesis a todos mis compañeros por apoyarme cuando lo necesitaba y por el amor que se da cada día, de verdad muchas gracias a todos.

AGRADECIMIENTO

Agradezco a la Universidad Politécnica Salesiana y a los docentes que conforman la carrera de Ingeniería De Sistemas los cuales formaron parte de mi formación académica dentro de esta institución. Al profesor y tutor el Ingeniero Miguel Ángel Quiroz Martínez por sus conocimientos brindados, su acompañamiento y consejo durante la carrera y elaboración de este trabajo.

Jeampier Carriel.

Análisis y evaluación de la técnica de procesamiento de lenguaje natural automático supervisado para determinar la polaridad de un texto no estructurado en redes sociales

Jeampier Alexander Carriel Roca¹

Universidad Politécnica Salesiana,
Guayaquil, Ecuador

¹{jcarrielr1}@ups.edu.ec

Resumen. La clasificación de documentos basada en opiniones se convirtió en un tema interesante para la sociedad de investigación del procesamiento del lenguaje natural. El interés en el procesamiento automatizado es resultado del crecimiento del contenido creado por el usuario. El presente trabajo tiene como fin desarrollar un modelo basado en la técnica de procesamiento de lenguaje natural automático, la metodología es la utilizada en el análisis de algoritmos de aprendizaje, un conjunto de entrenamiento, validación y prueba. El conjunto de entrenamiento y validación se usan para encontrar el algoritmo público, además se usó el método experimental con datos de casos reales. El uso de herramientas y software para la visualización y evaluación de datos sirvió para evaluar el modelo, usándose el algoritmo de regresión logística, los resultados logrados han demostrado ser bastantes satisfactorios lográndose buenos resultados, con una precisión del 98% cumpliendo con las expectativas.

Palabras clave: Lenguaje Natural, polaridad, algoritmo.

1 Introducción

El procesamiento del lenguaje natural (PLN) es una ciencia definida de manera oficial como una rama de estudio que combina la inteligencia artificial, la informática, la lingüística y los procesos del lenguaje natural para generar inteligencia y conocimiento. La importancia de esta ciencia radica en la comprensión y procesamiento de información que posteriormente puede ser utilizada en diferentes campos como: búsqueda, máquinas traductoras, reconocimiento de entidad nombrada (NER), agrupación de información, clasificación, análisis de sentimientos [1].

El PLN cubre los pasos y técnicas aplicables a campos específicos como clasificación de texto, NER y análisis de sentimientos, siendo uno de los más

importantes el preprocesamiento [2].

Algunas técnicas dependen del contenido del texto con el que se está trabajando. Por ejemplo, al usar Twitter se destacan las técnicas de limpieza como eliminación de enlaces web, símbolos de hashtag, nombres de usuario, nombres propios, espacios en blanco y signos de puntuación [3].

La importancia de aplicar estas técnicas en NLP se debe a la frecuencia con la que se generan datos de mala calidad, por lo que estas pueden usarse para procesar datos de una manera eficiente y lograr buenos resultados.

2 Preliminares

Esta sección proporciona una breve revisión de los denominados Aprendizaje automático y Procesamiento del lenguaje Natural.

2.1 Aprendizaje Automático

El aprendizaje automático es una especialidad de la IA (inteligencia artificial) dedicada al análisis de programas o agentes que se preparan o crecen a partir de la experiencia para realizar cierta labor cada vez mejor [3].

Particularmente, en el aprendizaje supervisado se tiene conocimiento de los datos con anterioridad.

Los siguientes algoritmos están disponibles para determinar la clasificación de polaridad de un texto no estructurado [4]:

- Árboles de decisión (Random Forest)
- Máquina de soporte vectorial (SVM)
- K-vecinos más cercanos (Knn)
- Redes bayesianas (Naive Bayes).

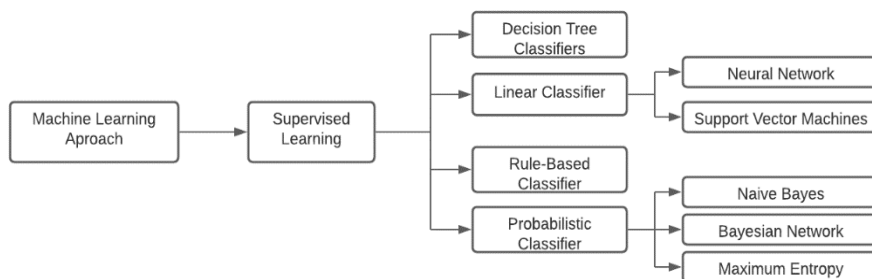


Fig. 1. Algoritmos del aprendizaje supervisado.

2.2 Procesamiento del lenguaje natural

El procesamiento del lenguaje natural es la capacidad que tiene una máquina para procesar información transmitida en lenguaje natural [4]. Crean modelos computacionales en lenguajes lo suficientemente detallados como para permitir la creación de información que ejecutan diversos comandos y solicitudes con las que interfiere el lenguaje natural. El NLP se basa en utilizar una expresión natural que logre comunicarse de manera directa con el ordenador, como escribir o dar comandos de voz, facilitando dar una orden o solicitud con el lenguaje y continuar desarrollando habilidades[5]. Dejar que el lenguaje del mecanismo se conecte con los patrones ayuda a las personas a comprenderlo [6].

Durante la programación de software de ordenadores, se simulan las habilidades del lenguaje humano, con esto la gramática computacional es la ciencia que estudia el procesamiento del lenguaje natural [7]. Esta aclaración distingue LC-NLP de la IA y la gramática computacional. Siendo la IA la responsable de codificar programas cognitivos con la capacidad para tomar decisiones, adquirir habilidades y sacar conclusiones a partir de los hechos. La gramática computacional es parte integral de la IA, y de la misma forma que la definen los gramáticos, hablamos de una parte de la psicología por tratarse de habilidades cognitivas por excelencia, el lenguaje [8].

2.3 Procesamiento informático en el lenguaje natural

El atributo más atractivo del procesamiento informático es que se pueden analizar a partir de un criterio plenamente sin dependencia del mecanismo que debería implementarlos[9]. El NLP se fundamenta en procesar información caracterizada como proyecciones a partir de una entrada y una salida, de modo que sea capaz de simbolizarse como un algoritmo definido [10]. Conforme pasa el tiempo la relación entre los humanos y máquinas se hace más frecuente, cualquier avance en el intento de entendernos mejor será también un gran paso hacia delante para la evolución de este nuevo mundo en la sociedad de la información [9]. La primordial labor de la IA es trabajar con lenguajes naturales utilizando instrumentos informáticos, los lenguajes de programación son muy importantes, ya que gracias a ellos podemos desarrollar la interacción del lenguaje natural y máquina [10]. El NLP se fundamenta en la relación entre el lenguaje natural y el ordenador, teniendo que procesar oraciones y textos que se encuentren otorgando, se desarrollan prototipos que ofrecen ayuda para comprender cómo interactúan los artilugios humanos conectados al lenguaje[11]. El procesamiento del lenguaje natural se aborda desde la gramática informática, un campo de la gramática aplicada a la IA cuyo objetivo primordial es la realización de estudios informáticos que simulen las capacidades humanas, el procesamiento es un campo fundamental de la IA y es

considerado uno de los más maduros [8].

3 Metodología

El presente artículo se centra en el área de aprendizaje supervisado, en particular de tipo exploratorio diagnóstico de carácter correlacional [12], su alcance principal se basa en las variables que interceden en la predicción de la demanda. Basándonos en esto, la metodología que se seguirá es la usada tradicionalmente en el análisis de cualquier algoritmo de aprendizaje, es decir, se tiene un conjunto de entrenamiento X, un conjunto de validación V y un conjunto de prueba o público P. El conjunto de entrenamiento y validación se usan para encontrar el algoritmo público o de prueba [12, 13]. Adicional a esto, se utilizará el método experimental con datos provenientes de casos reales.

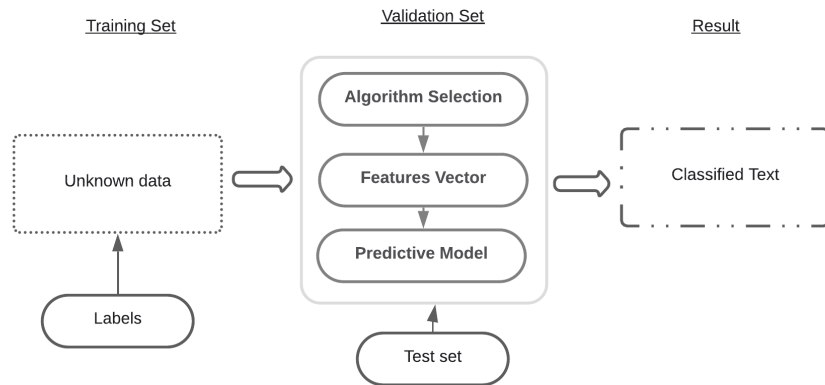


Fig. 2. Metodología propuesta.

Para modelar la metodología se usó el software de visualización de datos y aprendizaje automático de código abierto [Orange](#) y el dataset de prueba fue tomado de la red social Twitter los mismos que se encuentran en el repositorio (ver [Repositorio](#)).

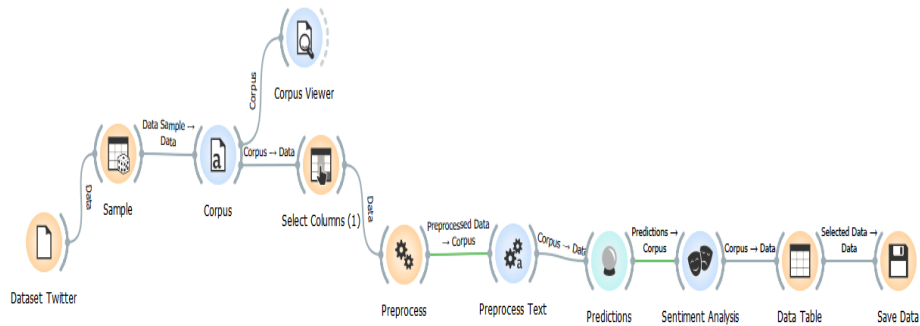


Fig. 3. Metodología propuesta en el software Orange.

El dataset descargado contiene información no estructurada y sin procesar, es decir, no se le ha aplicado el proceso de clasificación. Además, los tuits contienen atributos no relevantes, en la tabla 1 se muestran todos sus atributos [15].

Tabla 1. Atributos de los tweets.

Atributos
id
text
favorited
favoriteCount
<i>statusSource</i>
<i>truncated</i>
<i>replyToSID</i>
<i>replyToUID</i>
<i>replyToSN</i>
<i>screenName</i>
<i>retweetCount</i>
<i>isRetweet</i>
<i>retweeted</i>
<i>longitude</i>
<i>latitude</i>

De estos, solo se considera importante el atributo **text** el cual será nuestra clase, ya que este determinará la polaridad del texto.

4 Resultados

Para visualizar y analizar los resultados del dataset procesado, se utilizó la herramienta Power BI, en la cual se hizo una transformación de los datos para obtener los resultados vistos en la Fig. 4., en esta podemos ver la columna text

la cual muestra el texto y la columna Sentiment, la cual nos dice la polaridad del texto ya procesado.

text	Sentiment
Wow, the first reviews of #AvengersEndgame are <U+0001F631><U+0001F631><U+0001F631>https://t.co/bZubeR7uAK	Positive
With #AvengersEndgame in just a few days, I'd like to bring this back... https://t.co/luoQuMSzK	Positive
With #AvengersEndgame coming so soon, I want to know: what do you want to see from the #Avengers game coming fromâ€¦ https://t.co/oAh6snNEGE	Positive
With #AvengersEndgame and a new Thanos title out this week, @prbates36 recommends 10 Mad Titan comic-book runs thatâ€¦ https://t.co/9DfoOwJF72	Negative
Who's watching #AvengersEndgame on its first day tomorrow, April 24?	Neutral
Here's where you can watch it as early as 6:01â€¦ https://t.co/NOIANwP6xG	
whoever that in the captain marvel suit is: WHERE DID YOU GET THAT OR HOW DID YOU MAKE IT? It looks so awesome I neâ€¦ https://t.co/uvtho9mAM1	Positive
Who wants 1x ticket for #AvengersEndGame?	Neutral
8:00pm session at Melbourne Central Hoyts Xtremescreen 24th of April. Seaâ€¦ https://t.co/hFy7ELNIZ9	
Who knew pre-booking tickets for #AvengersEndgame could be so stressful	Negative
When your fiancâ€¦ surprises you with #AvengersEndgame tickets <U+0001F60D><U+0001F60D> https://t.co/dfdDgRih9IN	Positive

Fig. 4. Imagen de textos clasificados en Power BI.

La Fig. 5 muestra un gráfico de anillos con la cantidad de tweets procesados y la clasificación de los mismos por polaridad.

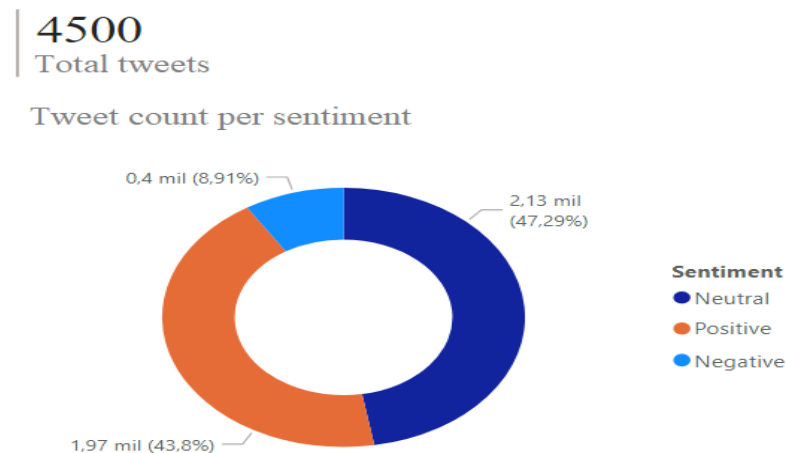


Fig. 5. Gráfico de anillos con total de tweets clasificados y polaridad

Para evaluar nuestro modelo, se utilizó el algoritmo de regresión logística, el cual es un método estadístico que sirve para examinar un grupo de datos en el cual existe una o más variables independientes que determinan un resultado [16], es decir, toma cierto número de clases en donde se consideró 3 categorías de sentimientos: neutro, positivo y negativo.

A partir de los resultados experimentales, se generó la siguiente matriz de

confusión (Fig. 6) la cual nos permite visualizar el desempeño del algoritmo que se emplea en el aprendizaje supervisado y el test de exactitud (Fig. 7.) que nos permite probar los algoritmos de aprendizaje en el software Orange.

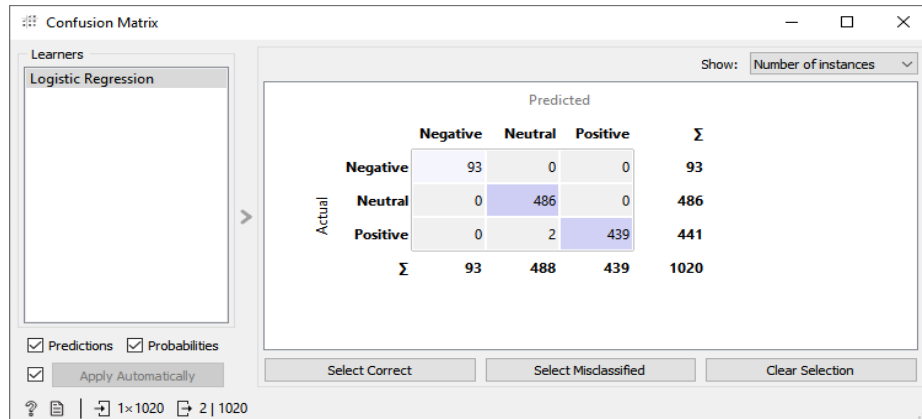


Fig. 6. Matriz de confusión

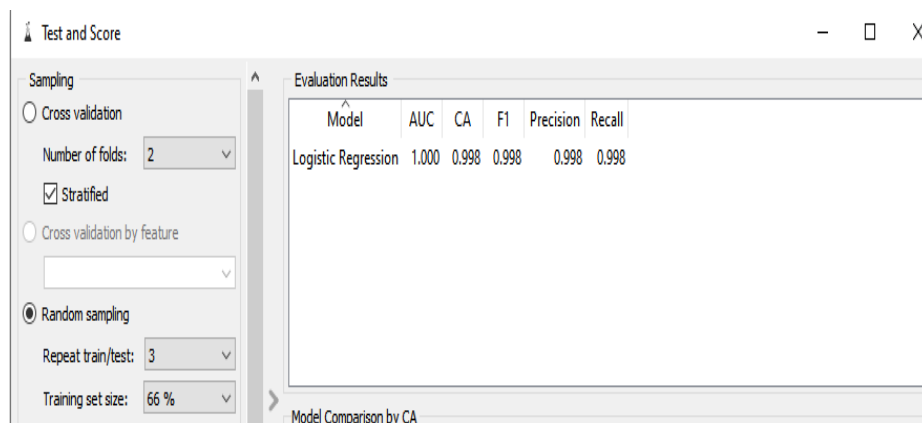


Fig. 7. Test and Score

Con esto se demuestra que el modelo planteado cumple con las expectativas, obteniendo una precisión del 0.998%.

El dataset, pipeline y demás archivos se encuentran en el repositorio para sus pruebas y verificación que se adjuntan a continuación https://github.com/jcarriel/text_analysis.

5 Conclusiones

De acuerdo al estado del arte realizado se determinó las herramientas, los algoritmos y el software específico para realizar este artículo, y gracias a eso se logró realizar el proceso con la obtención de los resultados esperados.

Con los estudios y revisión realizados se logró desarrollar un modelo basado en la técnica de procesamiento de lenguaje natural automático supervisado para determinar la polaridad de un texto, el análisis experimental ayudó en gran manera con técnicas gramaticales en la etapa del preprocesamiento, logrando una mejora en la clasificación.

Realizando una validación del modelo de clasificación con el algoritmo de la Regresión Logística, se concluye que con una precisión del 98%, el modelo planteado cumple con las expectativas de asertividad.

Referencias

1. Krouska, A., Troussas, C., Virvou, M.: The effect of preprocessing techniques on Twitter sentiment analysis. IISA 2016 - 7th International Conference on Information, Intelligence, Systems and Applications. (2016). <https://doi.org/10.1109/IISA.2016.7785373>.
2. Jianqiang, Z., Xiaolin, G.: Comparison research on text pre-processing methods on twitter sentiment analysis. IEEE Access. 5, 2870–2879 (2017). <https://doi.org/10.1109/ACCESS.2017.2672677>.
3. Alhajj, R., Rokne, J.: Encyclopedia of Social Network Analysis and Mining. (2018). <https://doi.org/10.1007/978-1-4939-7131-2>.
4. Zhou, D., Kong, H.: Encyclopedia of Applied and Computational Mathematics. (2015). <https://doi.org/10.1007/978-3-540-70529-1>.
5. Kumar, R.: Natural Language Processing. In: Machine Learning and Cognition in Enterprises. pp. 65–73. Apress, Berkeley, CA (2017). https://doi.org/10.1007/978-1-4842-3069-5_5.
6. Allen, L.K., Crossley, S.A., McNamara, D.S.: Predicting misalignment between teachers' and students' essay scores using natural language processing tools. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 9112, 529–532 (2015). https://doi.org/10.1007/978-3-319-19773-9_54.
7. Cedeno-moreno, D., Vargas-lombardo, M.: Aprendizaje automático aplicado al análisis de sentimientos Machine learning applied to the sentiment analysis. 16, (2020).
8. Etchegoyhen, T., Garcia, E.M., Azpeitia, A., Alegria, I., Labaka, G., Otegi, A., Sarasola, K., Cortes, I., Jauregi, A., Ellakuria, I., Calonge, E., Martin, M.: QUALES: Machine translation quality estimation via supervised and unsupervised machine learning. Procesamiento de Lenguaje Natural. 61, 143–146 (2018). <https://doi.org/10.26342/2018-61-18>.
9. Khan, D.M., Rao, T.A., Shahzad, F.: The Classification of Customers' Sentiment using Data Mining Approaches. Global Social Sciences Review. IV, 146–156 (2019). [https://doi.org/10.31703/gssr.2019\(iv-iv\).19](https://doi.org/10.31703/gssr.2019(iv-iv).19).
10. Rodrigues Chagas, B.N., Nogueira Viana, J.A., Reinhold, O., Lobato, F., Jacob, A.F.L., Alt, R.: Current Applications of Machine Learning Techniques in CRM: A Literature Review and Practical Implications. Proceedings - 2018 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2018. 452–458 (2019). <https://doi.org/10.1109/WI.2018.00-53>.
11. Thomas, J.R., Bharti, S.K., Babu, K.S.: Automatic keyword extraction for text summarization in e-newspapers. ACM International Conference Proceeding Series. 25-26-Aug, (2016). <https://doi.org/10.1145/2980258.2980442>.
12. Gauchi Risso, V.: Estudio de los métodos de investigación y técnicas de recolección de datos utilizadas en bibliotecología y ciencia de la información. Revista española de Documentación Científica. 40, 175 (2017). <https://doi.org/10.3989/redc.2017.2.1333>.
13. Sánchez-Holgado, P., Martín-Merino Acera, M., Blanco Herrero, D.: Del data-driven al data-feeling: análisis de sentimiento en tiempo real de mensajes en español sobre divulgación científica usando técnicas de aprendizaje automático. Anuario Electrónico de Estudios en Comunicación Social “Disertaciones.” 13, 35–58 (2020). <https://doi.org/10.12804/revistas.urosario.edu.co/disertaciones/a.7691>.
14. Ali, R., Lee, S., Chung, T.C.: Accurate multi-criteria decision making methodology for recommending machine learning algorithm. Expert Systems with Applications. 71, 257–278 (2017). <https://doi.org/10.1016/j.eswa.2016.11.034>.
15. Trupthi, M., Pabboju, S., Narasimha, G.: Sentiment analysis on twitter using streaming API. Proceedings - 7th IEEE International Advanced Computing Conference, IACC 2017. 915–919 (2017). <https://doi.org/10.1109/IACC.2017.0186>.
16. Pérez Obregón, J.M., Romero Díaz, T.: Análisis del rendimiento académico mediante regresión logística y múltiple. Revista Electrónica de Conocimientos, Saberes y Prácticas. 1, 33–42 (2018). <https://doi.org/10.30698/recsp.v1i2.10>.