ISSN: 1692-7257 - Volumen 2 – Número 40 - 2022

Revista Colombiana de Tecnologías de Avanzada

Recibido: 17 de abril de 2022
Aceptado: 23 de julio de 2022

APRENDIZAJE AUTOMATICO Y LA COLECCIÓN REUTERS-21578 EN LA CLASIFICACION DE DOCUMENTOS

MACHINE LEARNING AND THE REUTERS COLLECTION-21578 IN DOCUMENT CLASSIFICATION

Juan José Paniagua Medina[*], **Everardo Vargas Rodríguez**[*], **Rafael Guzmán Cabrera**[*].

[*] **Universidad de Guanajuato,** Departamento de Estudios Multidisciplinario Sede del Sur.
Av. Universidad S/N, Colonia Yacatitas, Yuriria, Gto.
Tel.: 445 458 90 40
E-mail: {jj.paniaguamedina, evr, guzmanc}@ugto.mx

Cómo citar: Paniagua Medina, J. J., Vargas Rodríguez, E., & Guzmán Cabrera, R. (2022). APRENDIZAJE AUTOMATICO Y LA COLECCIÓN REUTERS-21578 EN LA CLASIFICACION DE DOCUMENTOS. REVISTA COLOMBIANA DE TECNOLOGIAS DE AVANZADA (RCTA), 2(40), 39-46. https://doi.org/10.24054/rcta.v2i40.2344

Resumen: En la actualidad existe una gran facilidad para producir documentos, esto conlleva que exista demasiada información, toda esta información producida es casi imposible de organizar si no se utilizan métodos automáticos. La clasificación automática de documentos puede definirse como una acción ejecutada por un sistema artificial sobre un conjunto de documentos tanto estructurados o no estructurados. Esta acción se realiza utilizando las palabras contenidas en los documentos para definir la clase a la que pertenece el documento de prueba. En este trabajo presenta diversos experimentos de clasificación utilizando la base de datos Reuters-21578 con el fin de observar el comportamiento de los clasificadores naive bayes, máquinas de vectores de soporte (SVM por sus siglas en inglés) y regresión logística. Los resultados obtenidos permiten conocer el desempeño de los clasificadores, su comportamiento al aplicar técnicas de limpieza para la disminución de la dimensión de los documentos y diferentes escenarios de clasificación.

Palabras clave: Clasificación de documentos, naive bayes, regresión logística, SVM.

Abstract: Currently, it is very easy to produce documents, which means that there is too much information, and all this information produced is almost impossible to organize if automatic methods are not used. The automatic classification of documents can be defined as an action executed by an artificial system on a set of structured or unstructured documents. This action is performed by using the words contained in the documents to define the class to which the test document belongs. This paper presents several classification experiments using the Reuters-21578 database in order to observe the performance of naive Bayes classifiers, support vector machines (SVM) and logistic regression. The results obtained show the performance of the classifiers, their behavior when applying cleaning techniques to reduce the size of the documents and different classification scenarios.

Keywords: Document classification, naive bayes, logistic regression, SVM.

1. INTRODUCCIÓN

Durante las últimas décadas la cantidad de documentos en formato electrónico ha crecido de manera exponencial, la organización de esta información es una tarea casi imposible sin el uso de un sistema automático. Para organizar y aprovechar de manera eficiente la información contenida en los documentos existen diversas líneas de investigación como la minería de texto, aprendizaje automático (machine learning) y recuperación de información, entre otras. En este trabajo se utilizará el aprendizaje automático para clasificar de manera automática los documentos de la colección Reuters 21578 con el fin de clasificar con el menor margen de error cada uno de los documentos en sus respectivas clases. La clasificación de documentos se puede definir como la acción de formar grupos de documentos con las mismas características (Kaufman y Rousseeuw., 2009). Esta área contribuye con métodos para ordenar los documentos por clases, donde estas clases pueden ser idiomas, opiniones, temas, sentimientos solo por mencionar algunas (de Dios, 2009). La clasificación temática se usa para distinguir entre grupos con el fin de agrupar los documentos en la clase correspondiente (Sebastiani, 2005). El problema que presenta esta técnica es la dificultad de encontrar la categoría que mejor describa a un documento (Hearst y Pedersen, 1996). Por lo general un documento puede contener mas de un tema, por lo que la clasificación de documentos es una tarea muy subjetiva y compleja ya que dependiendo del enfoque de cada tópico puede encajar en una u otra categoría (Macskassy *et al.*, 1998).

Este trabajo tiene la finalidad de someter diversos clasificadores a diferentes escenarios de clasificación utilizando la colección Reuters 21578, con el fin de definir las mejores condiciones que permitan mejorar la clasificación automática de documentos.

2. ESTADO DEL ARTE

La colección Reuters 21578 es un referente para la clasificación automática de documentos, por lo que existen una gran cantidad de investigaciones que la utilizan (Paniagua, 2021; Bidi y Elberrichi, 2016; Eluri *et al.*, 2016) por ser un conjunto de documentos con varias clases, con traslape entre ellas, desbalanceada y escritas de manera estructurada, ya que todos los documentos de la colección son noticias periodísticas escritas por personas profesionales que recibieron un entrenamiento para realizar esta tarea.

La comparativa de los clasificadores es una de las tareas más importantes para conocer que clasificador funciona mejor para cada uno de los escenarios de clasificación (Suh, 2016). Adicionalmente se han realizado investigaciones de clasificación especificas en diversas áreas, por ejemplo, la clasificación de documentos con contenido de recetas de comida (Montero *et al.*, 2018), clasificación de documentos históricos por época (Ocampo, 2020) y la clasificación de documentos de acuerdo con su grado de relevancia dentro de un conjunto de clases definidas (Briseño, 2018).

Dentro de los trabajos relacionados con la clasificación automática existen diversos métodos que mejoran la eficiencia de los clasificadores, por ejemplo, Smalbil (2020) encontró que reducir el número de errores de etiquetado de los datos mejora el rendimiento computacional para diversos métodos de clasificación. Al-Tahrawi (2016) encontró que idioma en el que están redactados los documentos afecta el rendimiento de los clasificadores. Por último, Vala (2015) comparo el rendimiento de los clasificadores al trabajar con documento con dimensiones demasiado grande.

3. PROCESO DE CLASIFICACIÓN AUTOMÁTICA DE DOCUMENTOS

En la actualidad existen un sinfín de métodos implementados para realiza la clasificación automática de documentos, en la Fig. 1 se muestra la metodología que fue implementada en el presente trabajo. A continuación, se describe cada una de las etapas de la metodología propuesta.

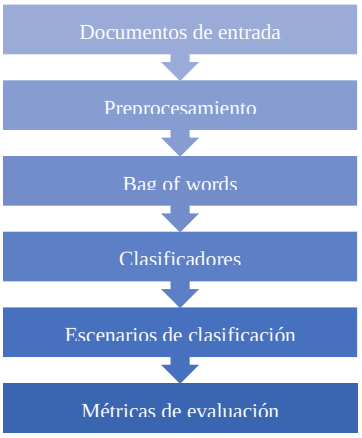

Fig. 1. Diagrama general del proceso.

3.1 Documentos de entrada

Para los documentos de entrada se utilizó la base de datos Reuters-21578 la cual fue recopilada por manualmente por el grupo Carnegie, esta base de datos es un conjunto de noticias reales de diferente índole que aparecieron en la agencia Reuters en el año 1987 (Lewis, 1997).

3.2 Preprocesamiento

El preprocesamiento es la etapa en que los datos son tratados con técnicas para la limpieza, integración, transformación y reducción de los datos con el fin de reducir la dimensionalidad de las matrices que nos servirán para llevar a cabo el proceso de clasificación, al reducir la dimensionalidad se reduce el tiempo de procesamiento (Hernández y Rodríguez, 2008).
A continuación, se menciona el preprocesamiento utilizado en el presente trabajo.

3.2.1 Baseline

El baseline representa a los documentos con su configuración y dimensionalidad original con el fin de contrastar los resultados al aplicar las diversas técnicas de preprocesamiento, es decir será el valor que tengamos como referencia para medir el impacto en la clasificación automática de documentos de las técnicas implementadas en el presente trabajo.

3.2.2 Lista de palabras de paro o Stopwords

En este trabajo se utilizó la técnica de limpieza para eliminar palabras vacías, también llamadas stopwords, es decir las palabras dentro de los documentos que no contienen información relevante para ayudar a realizar la clasificación, para ello se utiliza un enfoque basado en diccionario para eliminar palabras con ayuda de una lista genérica (Raulji y Saini, 2016).

3.3 Lematización

La lematización es una técnica que sirve para reducir las variantes morfológicas de las formas de una palabra a raíces comunes o lexemas, por ejemplo, “podrán, pudieron, poder → poder”, en palabras más simples, consiste en remover el plural, el tiempo, o los atributos finales de cada palabra (Balarkrishnan y Llloyd, 2014).

3.3 Bolsa de palabras o Bag of words

Para llevar a cabo la clasificación automática de documentos se utilizó la bolsa de palabras. Su función consiste en realizar un conteo de los puntos clave que se extraen de las palabras de un documento de texto (Zhang *et al., 2010*).

3.4 Clasificadores

En este trabajo se abordarán tres algoritmos de clasificación de aprendizaje supervisado, el algoritmo con el método de regresión logística, máquinas de vectores de soporte (SVM por sus siglas en inglés) y Naive Bayes.

3.4.1 Regresión logística

Dentro del aprendizaje automático la regresión logística (RL) es una técnica utilizada para la clasificación, es utilizada con condiciones donde es necesaria la descripción de las relaciones entre la variable categorica y los conjuntos de variables explicativas. Las probabilidades se modelan en función de variables explicativas utilizando una función logística (Williams *et al.*, 2005).

3.4.2 Naive Bayes

Los métodos de Naive Bayes (NB) son un conjunto de algoritmos de aprendizaje supervisado que se basan en la aplicación del teorema de Bayes con la suposición "Naive" de independencia condicional entre cada par de características dado el valor de la variable de clase (Webb et al., 2010).

3.5 Escenarios de clasificación

En este trabajo se utilizaron dos escenarios de clasificación validación cruzada y conjuntos de entrenamiento-prueba. La validación cruzada divide los documentos en k grupos, se toma uno de los grupos como datos de validación y el resto como valores de entrenamiento, este proceso es repetido k veces con cada uno de los grupos (Kohavi, 1995).
Los conjuntos de entrenamiento y prueba consisten en seleccionar una cantidad de los documentos y separarlos de los demás como datos de entrenamiento, una vez realizado el paso anterior se comienzan a realizar predicciones y con los datos restantes se realizan las pruebas (Sandoval, 2018).

3.6 Métricas de evaluación

Para la evaluación de los clasificadores se utilizaron dos métricas el F1-score y la precisión.

$$F1 = \frac{2 * precisión * recall}{precisión + recall}$$

La precisión se utiliza con el fin de conocer las predicciones realizada de manera correcta, es una medida de probabilidad de que un documento clasificado en la clase que corresponda realmente sea esa, la precisión se calcula como se muestra en la Ecuación (1).

El recall se utiliza con el fin de conocer las predicciones de que un documento pertenece a la clase es clasificado dentro de esa clase, el recall se calcula como se muestra en la Ecuación (2) (Melamed *et al.*, 2003).

$$precisión = \frac{Verdaderos\ Positivos}{Verdaderos\ Positivos + Falsos\ Positivos} \quad (1)$$

$$Recall = \frac{Verdaderos\ positivos}{Verdaderos\ Positivos + Falsos\ Negativos} \quad (2)$$

$$Recall = \frac{Verdaderos\ positivos}{Verdaderos\ Positivos + Falsos\ Negativos}$$

La métrica F1-score se utiliza para medir la exactitud en las pruebas, compara el rendimiento combinado de la precisión y el recall como se muestra en la Ecuación.

$$F1 = \frac{2 * precisión * recall}{precisión + recall} \quad (3)$$

4. METODOLOGIA

Para la realización de los experimentos desarrollados en el presente trabajo se utilizó un extracto de la colección con documentos. Esta sección fue de manera aleatoria y con la finalidad de reducir el tiempo de procesamiento, en la Tabla 1 se muestran la cantidad de documentos utilizados de la colección.

Tabla 1: Documentos utilizados de Reuters 21578.

Categoría	Documentos
earn	700
acq	568
corn	88
crude	149
grain	70
interest	108
money fx	100
ship	87
trade	112
wheat	79

Con la finalidad de medir el impacto en las métricas de evaluación utilizadas, así como el nivel de desbalanceo y traslapes se realizó la siguiente configuración del experimento:

1. Primero se realizó la clasificación de documentos utilizando todas las clases
2. Se usan únicamente las dos categorías con la mayor cantidad de documentos
3. Se utilizan las categorías con cantidad de documentos igual o mayor a 100

Categorías similares:

4. acq y ship
5. corn y grain
6. interest y money fx

Estos experimentos nos permitirán entender el impacto que tiene el desbalanceo, número de documentos por clase, número de clases y traslape en las métricas de evaluación utilizadas al realizar la clasificación automática de documentos en los diferentes escenarios implementados y con el procesamiento implementados.

Una vez establecido el orden de los experimentos se comenzó el proceso mostrado en la Fig.1.

En la Fig. 2 se muestra un diagrama del orden y las técnicas de preprocesamiento utilizadas para la reducción la dimensionalidad de los documentos.

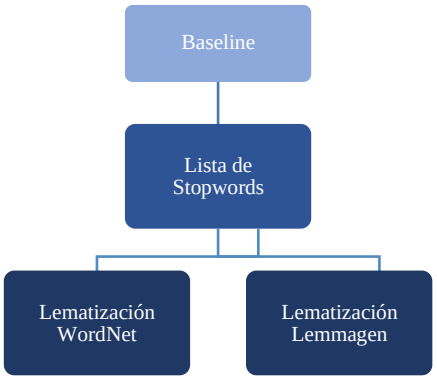

Fig. 2. Técnicas de preprocesamiento utilizadas.

La lista de las palabras vacías[1] utilizada contiene un total de 571 palabras.

Para la implementación de los escenarios de clasificación se utilizó la validación cruzada (VC) con 10 folds y para los conjuntos de entrenamiento-prueba (CE) se utilizó una proporción de 70-30 respectivamente, del total de documentos de cada experimento realizado. En cuanto a la validación de la clasificación se utilizaron las medidas de evaluación la precisión y el F1-score con ayuda de las Ecuaciones 1-3.

5. RESULTADOS

En la etapa de preprocesamiento se logró reducir la dimensionalidad de los documentos siguiendo el siguiente orden: baseline (BL), stopwords (SW), lematización WordNet (LW) y lematización Lemmagen (LL) tal y como se muestra en la Fig. 2. En la Tabla 2 se muestran los resultados obtenidos en la etapa de preprocesamiento.

Tabla 2: Reducción de la dimensionalidad.

	BL	SW	LW	LL
1	13893	13378	12225	10880
2	9448	8994	8359	7649
3	12220	11716	10763	9638
4	7878	7438	6821	6157
5	3886	3499	3199	2840
6	3060	2704	2490	2218

En la Tablas 3 y 4 se muestran los resultados obtenidos con la métrica F1-score para el clasificador regresión logística con el escenario de validación cruzada y conjuntos de entrenamiento, respectivamente.

Tabla 3: Resultados F1-score del método regresión logística con VC.

	BL	SW	LW	LL
1	0.903	0.911	0.915	0.916
2	0.971	0.971	0.971	0.973
3	0.957	0.957	0.961	0.963
4	0.977	0.978	0.979	0.979
5	0.772	0.759	0.766	0.779
6	0.947	0.933	0.937	0.933

Tabla 4: Resultados F1-score del método regresión logística con CE.

	BL	SW	LW	LL
1	0.898	0.907	0.905	0.906
2	0.973	0.973	0.976	0.978
3	0.942	0.952	0.956	0.955
4	0.968	0.978	0.978	0.978
5	0.736	0.759	0.773	0.777
6	0.933	0.933	0.940	0.938

En la Tablas 5 y 6 se muestran los resultados obtenidos con la métrica de precisión para el clasificador regresión logística con el escenario de validación cruzada y conjuntos de entrenamiento y prueba, respectivamente.

Tabla 5: Resultados precisión del método regresión logística con VC.

	BL	SW	LW	LL
1	0.902	0.911	0.915	0.916
2	0.971	0.971	0.971	0.973
3	0.957	0.957	0.962	0.963
4	0.977	0.978	0.981	0.981
5	0.772	0.759	0.766	0.779
6	0.947	0.933	0.938	0.933

Tabla 6: Resultados precisión del método regresión logística con CE.

	BL	SW	LW	LL
1	0.987	0.907	0.904	0.906
2	0.974	0.973	0.976	0.978
3	0.943	0.953	0.957	0.955
4	0.968	0.978	0.980	0.979
5	0.736	0.759	0.771	0.778
6	0.934	0.942	0.940	0.938

En la Tablas 7 y 8 se muestran los resultados obtenidos con la métrica F1-score para el clasificador naive bayes con el escenario de validación cruzada y conjuntos de entrenamiento, respectivamente.

Tabla 7: Resultados F1-score del método naive bayes con VC.

	BL	SW	LW	LL
1	0.703	0728	0.736	0.746
2	0.890	0.941	0.943	0.944
3	0.798	0.788	0.793	0.804
4	0.881	0.889	0.898	0.893
5	0.636	0.647	0.649	0.649
6	0.771	0.795	0.806	0.816

[1] https://github.com/manishkanadje/reuters-21578/blob/master/stopwords.txt

Tabla 8: Resultados F1-score del método naive bayes con CE.

	BL	SW	LW	LL
1	0.696	0.686	0.702	0.705
2	0.880	0.935	0.936	0.948
3	0.772	0.746	0.752	0.786
4	0.852	0.871	0.871	0.875
5	0.620	0.657	0.637	0.630
6	0.777	0.785	0.816	0.813

En la Tablas 9 y 10 se muestran los resultados obtenidos con la métrica de precisión para el clasificador naive bayes con el escenario de validación cruzada y conjuntos de entrenamiento y prueba, respectivamente.

Tabla 9: Resultados precisión del método naive bayes con VC.

	BL	SW	LW	LL
1	0.840	0.858	0.856	0.854
2	0.890	0.941	0.943	0.944
3	0.859	0.902	0.892	0.891
4	0.919	0.930	0.937	0.932
5	0.644	0.672	0.670	0.670
6	0.796	0.823	0.830	0.833

Tabla 10: Resultados precisión del método naive bayes con CE.

	BL	SW	LW	LL
1	0.837	0.849	0.854	0.840
2	0.880	0.935	0.936	0.949
3	0.842	0.910	0.898	0.882
4	0.911	0.924	0.925	0.926
5	0.622	0.666	0.645	0.636
6	0.805	0.810	0.839	0.835

En la Tablas 11 y 12 se muestran los resultados obtenidos con la métrica F1-score para el clasificador SVM con el escenario de validación cruzada y conjuntos de entrenamiento, respectivamente.

Tabla 11: Resultados F1-score del SVM con VC.

	BL	SW	LW	LL
1	0.352	0.397	0.421	0.417
2	0.593	0.617	0.656	0.650
3	0.420	0.504	0.533	0.524
4	0.805	0.805	0.805	0.805
5	0.398	0.398	0.398	0.398
6	0.355	0.355	0.355	0.355

Tabla 12: Resultados F1-score del método SVM con CE.

	BL	SW	LW	LL
1	0.342	0.377	0.332	0.398
2	0.621	0.671	0.726	0.571
3	0.462	0.503	0.431	0.515
4	0.807	0.807	0.807	0.807
5	0.405	0.405	0.405	0.405
6	0.360	0.360	0.360	0.360

En la Tablas 13 y 14 se muestran los resultados obtenidos con la métrica de precisión para el clasificador naive bayes con el escenario de validación cruzada y conjuntos de entrenamiento y prueba, respectivamente.

Tabla 13: Resultados precisión del método SVM con VC.

	BL	SW	LW	LL
1	0.454	0.498	0.496	0.498
2	0.769	0.779	0.787	0.783
3	0.586	0.633	0.639	0.632
4	0.752	0.752	0.752	0.752
5	0.310	0.310	0.310	0.310
6	0.270	0.270	0.270	0.270

Tabla 14: Resultados precisión del método SVM con CE.

	BL	SW	LW	LL
1	0.396	0.450	0.331	0.443
2	0.783	0.802	0.819	0.719
3	0.519	0.577	0.639	0.575
4	0.753	0.753	0.753	0.753
5	0.316	0.316	0.316	0.316
6	0.274	0.274	0.274	0.274

6. DISCUSIÓN

En la Tabla 2 se puede observar con claridad que al aplicar distintas técnicas de preprocesamiento de manera conjunta reduce de manera significativa la dimensionalidad de los documentos ayudando con esto a minimizar el tiempo de preprocesamiento.

A continuación, se compararán los resultados presentados previamente mostrados del desempeño de los clasificadores utilizados en este trabajo, con el fin de conocer sus ventajas y desventajas para cada uno de los experimentos realizados.

Para la clasificación con el método de regresión logística, las Tablas 3-6 muestran que los mejores resultados en la mayoría de los experimentos son obtenidos aplicando stopwords, lematización

lemmagen y la validación cruzada como escenario de clasificación, aunque la mejoría es muy pequeña (menor al 0.5% en cada experimento) en comparativa con el baseline. Los resultados muestran que el clasificador no presenta dificultades con desbalanceo, número de documentos y el traslape en los experimentos 4 a 6 los cuales cuentan con categorías similares.

Para la clasificación con el clasificador naive bayes, las Tablas 7-10 muestran que los mejores resultados son obtenidos aplicando stopwords, lematización lemmagen y la validación cruzada como escenario de clasificación, a diferencia de clasificador de regresión logística la mejoría es más relevante (mayor al 5% en la mayoría de experimentos) en comparativa con el baseline. Los resultados muestran que el clasificador comienza a empeorar el desempeño de la clasificación a medida que las categorías presentan mayor similitud.

Los resultados del clasificador SVM muestran que es el clasificador con el menor desempeño al clasificar los documentos en cada uno de los experimentos planteados en este trabajo, las Tablas 11-14 muestran para categorías similares el clasificador tiene un desempeño menor al 40%, además las técnicas de preprocesamiento no ayudan a mejorar el desempeño del clasificador. Este clasificador solo presenta un buen desempeño al clasificar clases con la misma cantidad de documentos y con clases con muy poca similitud.

A continuación, se mostrará los mejores resultados obtenidos por el clasificador de regresión logística con distintas técnicas de preprocesamiento. En la Fig. 3 se muestran los resultados obtenidos en el experimento 4 (acq-ship) con las métricas de evaluación para cada una de las etapas de preprocesamiento aplicadas en la base de datos, la cual presenta los mejores resultados de clasificación, con el fin de mostrar de manera grafica como a medida que se aumentan las técnicas de preprocesamiento mejora el desempeño del clasificador regresión logística, el cual obtuvo el mejor desempeño en cada uno de los experimentos de este trabajo.

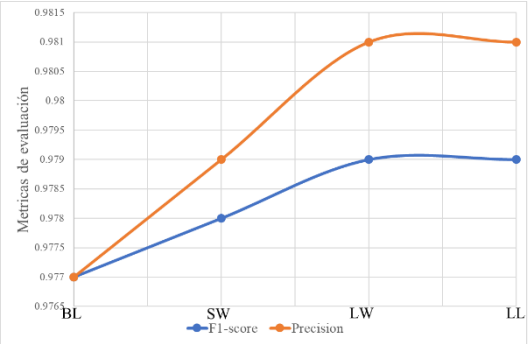

Fig. 3. Resultado de las métricas F1 y precisión (experimento 4) utilizando validación cruzada y regresión logística.

7. CONCLUSIONES

En primera instancia, los resultados muestran que al aplicar las técnicas conjuntas de preprocesamiento disminuye de manera considerable la dimensionalidad de los documentos sin perder la calidad de la clasificación.

De acuerdo con los resultados obtenidos por las métricas de evaluación en las Tablas 3 a 14 muestran que el mejor método de clasificación es utilizando regresión logística, adicional la aplicación de las técnicas de preprocesamiento stopwords y lematización Lemmagen en conjunto con el escenario de validación cruzada entregan los mejores resultados de las métricas F1-score y precisión en cada uno de los experimentos realizados en este trabajo. Por otra parte, se determinó que en el uso del clasificador SVM para categorías similares no realiza un buen trabajo de clasificación ya que una de las categorías absorbe las predicciones de la otra. Por lo tanto, se recomienda utilizar el método de regresión logística con aplicación de lista de stopwords y lematizadores ayuda aún más a mejorar la efectividad del clasificador, con ayuda de la Fig. 3 se puede observar el aumento en el desempeño a medida que se agregan técnicas de preprocesamiento.

Para finalizar se recomienda utilizar el clasificador de regresión logística, adicionalmente se recomienda si es necesario reducir el tiempo de procesamiento no aplicar técnicas de preprocesamiento ya que no se presentaron mejorías considerables para la mejora del desempeño del clasificador. También se recomienda no utilizar el clasificador SVM con clases con similitud ya que su desempeño es muy deficiente.

REFERENCIAS

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.

de Dios, J. (2009). Clasificación Automática de Textos usando Reducción de Clases basada en Prototipos.

Sebastiani, F. (2005). Text categorization. In *Encyclopedia of database technologies and applications* (pp. 683-687). IGI Global.

Hearst, M. A., & Pedersen, J. O. (1996, August). Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, 76-84

Macskassy, S. A., Banerjee, A., Davison, B. D., & Hirsh, H. (1998, August). Human Performance on Clustering Web Pages: A Preliminary Study. 264-268

Paniagua, J., Vargas, E., Guzmán, R. (2021). Clasificación automática de documentos utilizando aprendizaje automático y Reuters-21578. *CIENERGIA UG 2021,43-47*.

Bidi, N., & Elberrichi, Z. (2016, November). Feature selection for text classification using genetic algorithms. In *2016 8th International Conference on Modelling, Identification and Control (ICMIC)* (pp. 806-810). IEEE.

Eluri, V. R., Ramesh, M., Al-Jabri, A. S. M., & Jane, M. (2016, March). A comparative study of various clustering techniques on big data sets using Apache Mahout. In *2016 3rd MEC International Conference on Big Data and Smart City (ICBDSC)* (pp. 1-4). IEEE.

Suh, J. H. (2016). Comparing writing style feature-based classification methods for estimating user reputations in social media. *SpringerPlus*, *5*(1), 1-27.

Montero, S. C., Hernández, K. M., Murillo, É. C., de León, J. A. L., & Hernández-Delgado, M. (2018). Análisis de texto para la identificación automática de marcadores lingüísticos definicionales en recetas de gastronomía de Costa Rica. *Káñina*, *42*(3), 65-78.

Briceño Segovia, F. S. (2018). Clasificación automática de textos basado en ranking.

Ocampo Vargas, M. J. (2020). Análisis automático de documentos con contenido histórico en español.

Smalbil, J. (2020). Web-Based Economic Activity Classification: Comparing semi-supervised text classification methods to deal with noisy labels.

Vala, M., & Gandhi, J. (2015). Survey of text classification technique and compare classifier. *International Journal of Innovative Research in Computer and Communication Engineering*, *3*(11), 10809-10813.

Al-Tahrawi, M. M. (2016). Polynomial Neural Networks versus Other Arabic Text Classifiers. *J. Softw.*, *11*(4), 418-430.

Lewis, D. (1997). Reuters-21578 text categorization test collection, distribution 1.0. *http://www. research/. att. com*.

Hernández, C., & Rodríguez, J. E. R. (2008). Preprocesamiento de datos estructurados. *Revista vínculos*, *4*(2), 27-48.

Raulji, J. K., & Saini, J. R. (2016). Stop-word removal algorithm and its implementation for Sanskrit language. *International Journal of Computer Applications*, *150*(2), 15-17.

Balakrishnan, V., & Lloyd-Yemoh, E. (2014). Stemming and lemmatization: a comparison of retrieval performances, 174-179.

Zhang, Y., Jin, R., & Zhou, Z. H. (2010). Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, *1*(1), 43-52.

Webb, G. I., Keogh, E., & Miikkulainen, R. J. E. o. m. l. (2010). Naïve Bayes. 15, 713-714.

Cristianini, N., & Shawe-Taylor, J. (2000). An introduction to support vector machines and other kernel-based learning methods. Cambridge university press.

Williams, D., Liao, X., Xue, Y., & Carin, L. (2005, August). Incomplete-data classification using logistic regression. In *Proceedings of the 22nd International Conference on Machine learning* (pp. 972-979).

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Paper presented at the Ijcai.

Sandoval, L. (2018). Algoritmos de aprendizaje automático para análisis y predicción de datos. *Revista Tecnológica; no. 11*.

Melamed, I. D., Green, R., & Turian, J. (2003). Precision and recall of machine translation. In *Companion Volume of the Proceedings of HLT-NAACL 2003-Short Papers* (pp. 61-63).