Revista Venezolana de Computación
ISSN: 2244-7040
http://www.svc.net.ve/revecom
Vol. 4, No. 1, pp. 1-9, Junio 2017
Fecha recepción: 27/04/2017, Fecha aceptación: 05/08/2017

Implementación de un Método para la Clasificación Automática de Documentos Usando Tareas de Procesamiento de Lenguaje Natural y un Algoritmo de Máxima Entropía

Luis Molina[1], Javier Maldonado[1]
lm.molinab@gmail.com, jmaldo@unet.edu.ve

[1] Lab. de Computación de Alto Rendimiento (LCAR), Universidad Nacional Experimental del Táchira, Venezuela

Resumen: La presente investigación tiene como propósito la implementación de un método para la clasificación automática de documentos usando una aproximación de procesamiento de lenguaje natural (PLN), y un algoritmo basado en el principio de máxima entropía. Se consiguió una combinación de técnicas y parámetros que brinde la mayor eficacia posible en la clasificación de un texto ingresado con base en un conjunto de categorías preseleccionadas. El proceso investigativo inicia con la selección de las categorías y datos a usarse para los experimentos, obteniendo las cantidades de documentos que brinden mayor estabilidad al sistema seguido del pre-procesado de dichos datos mediante el uso de algoritmos de PLN, posteriormente se ejecuta el entrenamiento y luego las pruebas a cada experimento con las cuales se obtienen las medidas de evaluación para el clasificador. Finalmente se realiza un análisis comparativo de los resultados, determinando así la combinación de parámetros y técnicas de pre-procesado que brinde mayor eficacia en la clasificación para el conjunto de documentos estudiados. Todo este proceso está enmarcado en un entorno de noticias digitales, en el cual se consiguió una clasificación efectiva para el 91% de los documentos analizados, utilizando siete categorías con un total de 1400 noticias de entrenamiento por cada una y haciendo uso de la eliminación de palabras vacías y el *stemming* como técnicas de PLN, mostrando así la efectividad de los métodos utilizados en cuerpos de texto escritos en el idioma español.

Palabras Clave: Clasificación Automática de Documentos; Procesamiento de Lenguaje Natural; Máxima Entropía.

Abstract: The main purpose of this research is to obtain a method for automatic document classification using natural language processing (NLP), and an algorithm based on maximum entropy principle. The main goal is to obtain a combination of techniques and parameters that provide the greatest possible efficiency in text classification on a set of pre-selected categories. The research begins with the selection of the categories and data to be used for the experiments, obtaining the quantities of documents that provide greater stability to the system, followed by the data preprocessing through the use of NLP algorithms, then the training is performed based on a series of parameters, after that, some tests are executed to each experiment and we obtain the evaluation measures for the classifier. Finally, a comparative analysis of the results is carried out, in that way determining the combination of parameters and preprocessing techniques that provide the highest efficiency in the classification for the studied dataset. This entire process is focused on a digital news source, which achieve a 91% classification effectiveness of the processed documents using seven categories, with a training dataset of 1400 digital news for each one and using stop word removal and stemming as NLP techniques, showing the effectiveness of this methods on text bodies written on Spanish.

Keywords: Automatic Document Classification; Natural Language Processing; Maximum Entropy.

I. INTRODUCCIÓN

Con el constante aumento de la información en formato digital, surge la necesidad de encontrar métodos que permitan obtener conocimiento útil a partir de este gran número de datos. Una de las áreas encargadas de este tipo de análisis es la minería de texto, la cual consiste en realizar operaciones para analizar textos con la finalidad de extraer conocimiento como se cita en [1]; dentro de ella se encuentra la clasificación automática de documentos, la cual ha presentado un activo campo investigativo en los últimos años, en ésta se busca tomar un gran conjunto de textos y asignarlos a una o varias categorías.

Es necesario conocer que el proceso base para la clasificación automática de documentos consta de dos fases, en la primera, el usuario define las categorías en las cuales está interesado y proporciona al clasificador un conjunto de documentos de entrenamiento, con la finalidad de que el algoritmo aprenda. Después de esto, se procede a la fase de pruebas, en la cual se le entrega al clasificador un conjunto de documentos distinto al de entrenamiento, con el objetivo de que sean categorizados y de esta forma verificar que tan precisa es la clasificación.

Este proceso de clasificación depende directamente del contenido del documento que se busca categorizar, por lo tanto, es común agregar una fase previa de pre-procesamiento de los documentos, con la intención de ajustarlos de una mejor manera al proceso de clasificación que será utilizado en las dos fases posteriores (entrenamiento y pruebas), consiguiendo finalmente mejorar los resultados del clasificador. Esto se puede lograr aplicando distintas técnicas de procesamiento de texto de acuerdo a la necesidad del analista.

En la presente investigación se hace uso de un clasificador basado en la teoría de máxima entropía, así como la utilización de una serie de métodos de procesamiento de lenguaje natural (PLN), aplicados al texto como son el *stemming* y la eliminación de palabras vacías.

Uniendo el enfoque de uniformidad de la máxima entropía con el procesamiento de los documentos previo a su clasificación, se busca aumentar el rendimiento y mejorar los resultados arrojados por el algoritmo clasificador, indicando en última instancia la combinación de métodos de PLN que brinde los mejores resultados en el proceso de categorización para el conjunto de datos empleado.

Para el cuerpo de documentos estudiado, con un total de 1400 noticias de entrenamiento por cada una de las siete categorías seleccionadas, se obtuvo un 91% de efectividad en la clasificación, seleccionando la eliminación de palabras vacías y el *stemming* como técnicas de PLN para el pre-procesado de los documentos. De esta manera se consigue mostrar la efectividad de los métodos y técnicas utilizadas para conjuntos de documentos escritos en el idioma español, ya que la mayor parte de los trabajos previos en esta área han sido realizados sobre cuerpos de texto en inglés.

El artículo tiene la siguiente organización: en la Sección II se presentan algunas definiciones importantes para la comprensión de la investigación, en la Sección III se hace referencia a trabajos previos relacionados con esta investigación, en la Sección IV se indica la metodología empleada, en la Sección V se describen los experimentos realizados y el procedimiento utilizado para conseguir los resultados a ser analizados en la Sección VI, finalmente se presentan las conclusiones respectivas y algunos apuntes para trabajos futuros.

II. Conceptos Importantes

A. Procesamiento de Lenguaje Natural (PLN)

El PLN [2] es un conjunto de métodos de inteligencia artificial enfocados al entendimiento de la lingüística, su principal objetivo consiste en procesar un conjunto de frases en lenguaje humano (también llamados lenguajes naturales), de tal forma que puedan ser interpretados automática y eficientemente por un algoritmo computacional, dando respuestas con base en las instrucciones recibidas.

De acuerdo con [3], el uso de estos lenguajes naturales, facilita el desarrollo de programas que realicen tareas relacionadas con el lenguaje o bien, desarrollar modelos que ayuden a comprender los mecanismos humanos relacionados con el lenguaje.

El PLN posee una variada gama de aplicaciones, entre estas encontramos el análisis de sentimientos, traducción automática, recuperación de información y el resumen automático de textos.

Dentro de la gran variedad de tareas de PLN existentes, algunas de las más habituales y de las cuales se hace uso en la presente investigación son: reconocimiento de nombres [4], *stemming* [5] y eliminación de palabras vacías [6].

B. Clasificación Automática de Documentos

En un entorno comprendido por datos no estructurados como son los textos planos, la clasificación o categorización automática de documentos es considerada en [7] como probablemente el tema más común al analizar datos complejos.

Partiendo de un conjunto de categorías preestablecido, el clasificador asigna una de ellas al texto analizado con base en su contenido. Tal y como se indica en [8], el principal enfoque para el problema de categorización de textos se basa en técnicas de aprendizaje automático.

Este proceso está dividido en tres etapas las cuales son:

1) Pre-procesado: como se describe en [9], en esta etapa se obtienen las características o términos clave a partir de documentos almacenados y mejorar la relevancia entre palabra y documento, así como también entre palabra y categoría. En esta etapa, el texto es procesado y se determinan aquellas entidades que proporcionen información relevante al clasificador.

2) Entrenamiento: el usuario provee al clasificador un conjunto de textos previamente etiquetados, cada uno puede ser asociado a una o más categorías. A partir de esta información, el algoritmo de aprendizaje podrá realizar asociaciones, las cuales determinarán a que categoría pertenece un documento. En [10] se describe que la categorización solo cuenta las palabras que aparecen y, a partir de las cuentas, identifica los temas principales que cubre el documento.

3) Clasificación: De acuerdo con [11], aquí se introducen nuevos documentos no revisados por el clasificador, el cual retorna una asociación de clase a la que pertenece cada uno de ellos en función de las reglas previamente generadas.

C. Medidas de Evaluación en Algoritmos de Clasificación

Ante la diversidad de algoritmos que permiten la categorización automática de una serie de documentos, es necesario que existan medidas que permitan evaluar la exactitud de las operaciones de clasificación. Los más comunes son:

1) Precisión: Es una medida que evalúa la probabilidad de que las clasificación de un documento d en una clase c sea

correcta. También se describe en [12] como la fracción de los documentos recuperados en la clasificación que son relevantes.

2) Cobertura: También conocida por su nombre en inglés recall, es una medida que evalúa la proporción de documentos d que forman parte de una clase c, que son seleccionados por el algoritmo como pertenecientes a dicha clase. Se puede definir de igual manera como la fracción de los documentos relevantes que se recuperan en la clasificación, tal como lo hace [12].

3) Medida F: Es una medida que une la precisión y la cobertura para determinar la eficiencia del algoritmo, además, tal como indica [13], incluye el parámetro β ($0 \leq \beta \leq \infty$) para indicar el nivel de importancia relativo entre ambas medidas; si β es mayor a 1 se le da un mayor peso a la cobertura, mientras que si es menor a 1, la precisión será la que tenga más relevancia y si el valor de β es 1, ambas tendrá las misma importancia. La fórmula para calcular la medida F se representa en la Ecuación (1).

$$F_\beta = \frac{(\beta^2 + 1) precision * cobertura}{\beta^2 precision + cobertura} \quad (1)$$

Para los experimentos realizados en este trabajo el valor de β utilizado es 1, los motivos de tal selección están detallados en la fase experimental.

D. Máxima Entropía

Maxent, o máxima entropía, es una técnica para determinar distribuciones de probabilidad a partir de un conjunto de datos. Su principio base establece que ante el desconocimiento de información, la distribución deberá ser uniforme.

Con relación a esto, y según se define en [14], la idea detrás de la máxima entropía es que se debe preferir los modelos más uniformes que al mismo tiempo satisfagan cualquier restricción dada. Además, en su formulación más general, la máxima entropía puede ser usada para estimar cualquier distribución de probabilidad.

En un sistema que se encuentra sujeto a una serie de restricciones, la distribución se adaptará a las mismas, manteniendo siempre la mayor uniformidad posible en la distribución de probabilidad calculada, esto es, tener máxima entropía.

Para emplear la máxima entropía en la categorización automática de documentos, se debe usar un conjunto de textos previamente etiquetados, a partir de las cuales, el algoritmo creará las restricciones necesarias para estimar la distribución de probabilidad del modelo. Un ejemplo ilustrativo acerca del uso de la máxima entropía en el ámbito de la clasificación automática de documentos se puede observar en [14].

III. TRABAJOS RELACIONADOS

Con base en la revisión de literatura relevante a la presente investigación, se ha evidenciado que existen distintos trabajos realizados previamente referentes a la clasificación automática de documentos, procesamiento de lenguaje natural y máxima entropía, aplicadas a la obtención de información.

En [15] se presenta un enfoque diferente al tradicional para construir sistemas de extracción de información. La característica principal de la arquitectura es el escaso uso de recursos lingüísticos, los cuales son reemplazados por métodos de aprendizaje supervisado. Además se presentan algunas bases acerca de los métodos usados para la clasificación automática de documentos así como las distintas medidas para evaluarlos.

En [6] se realiza un estudio acerca de la aplicación de distintas técnicas de procesamiento de lenguaje natural enfocados en la recuperación de información. Se presentan los conceptos referentes a dichas técnicas, así como los resultados de su aplicación de manera individual y a través de combinaciones entre ellas, mostrando así como se pueden realizar experimentos que involucren más de una técnica para mejorar sus resultados.

Una implementación de técnicas de PLN en el ámbito de la clasificación automática de textos se puede observar en [8]. En dicho trabajo se pretende utilizar la clasificación de documentos, implementando distintas técnicas de PLN para pre-procesar los textos, todo esto con el objetivo de recuperar información relevante de los textos estudiados. Además se presentan conceptos relevantes a dichos temas y un enfoque para la construcción de los experimentos de clasificación.

En [14] se propone la implementación de técnicas de máxima entropía para la clasificación automática de documentos. En ésta investigación, se utiliza la técnica de máxima entropía para clasificación de textos, estimando la distribución condicional de la clase dado el documento. Se realizan experimentos con varios conjuntos de datos y se compara la precisión del algoritmo a la obtenida usando un clasificador de Bayes ingenuo (Naïve Bayes). Obtienen como resultado que el rendimiento de la máxima entropía es significantemente mejor en la mayoría de los casos. Con esto concluyen que aún queda mucho trabajo en esta área, pero los resultados indican que la máxima entropía es una técnica prometedora para la clasificación automática de textos.

IV. METODOLOGÍA

La metodología seguida para completar la investigación, tiene como base las fases del descubrimiento de conocimiento en bases de datos (Knowledge Discovery in Databases, KDD), siendo éste un concepto muy utilizado en la actualidad en la extracción y análisis de conocimiento a partir de un conjunto de datos, tal como se indica en [16], el KDD se refiere al proceso completo de descubrir conocimiento útil a partir de datos. A pesar de que el KDD comprende un gran número de fases dentro su aplicación [16], en este trabajo se emplean las más relevantes para la investigación las cuales son:

A. Selección de los Datos

En esta etapa, se seleccionan de la base de datos de noticias, aquellas que serán usadas tanto en el conjunto de entrenamiento como el de pruebas dentro del proceso de clasificación, y se generan los documentos respectivos para cada conjunto.

Para este trabajo se hace una selección aleatoria de una misma cantidad de noticias por cada categoría con la intención de

mantener la uniformidad de los datos y no brindar mayor importancia a alguna de las categorías estudiadas.

B. Pre-procesamiento

Durante esta fase, se realiza la selección de las distintas técnicas que serán utilizadas para pre-procesar las noticias, tanto las contenidas en el conjunto de entrenamiento como en el de pruebas.

Los documentos que serán procesados son los generados en la fase previa, produciendo internamente nuevos documentos de entrenamiento y pruebas durante la ejecución de cada experimento realizado.

C. Clasificación

Se realiza la categorización automática de las noticias, esto por cada una de las combinaciones de técnicas seleccionadas en la etapa anterior, primero se entrena el clasificador con el archivo de noticias de entrenamiento, generando así el modelo de clasificación.

En cuanto a la presente investigación, dicho modelo consiste de un archivo binario, generado por el método de entrenamiento del algoritmo de clasificación utilizado. Este modelo es usado seguidamente en la fase de pruebas, haciendo uso del respectivo archivo con las noticias pertenecientes al conjunto de pruebas y usando dicho modelo para la clasificación.

D. Análisis

Se toman los resultados obtenidos durante la etapa de clasificación y se realiza un análisis comparativo entre cada uno de ellos, el objetivo de esto es determinar cuáles combinaciones de técnicas producen mejoras en los resultados de clasificación al ser usadas en la etapa de pre-procesamiento, teniendo como referencia un experimento ejecutado sin su utilización.

V. Fase Experimental

En los experimentos realizados se utilizó un cuerpo de noticias provenientes de una base de datos, la cual contiene aproximadamente 2500 noticias por cada una de las siete categorías: deportes, economía, mundo, política, salud, sucesos y tecnología; tales textos son correspondientes a la versión web del diario El Nacional.

Dicha fuente se presenta como un conjunto de datos atractivo para el estudio ya que es un diario de gran circulación a nivel nacional, es de fácil acceso y tiene un repositorio de noticias abundante que cubren la variedad de categorías estudiadas.

Estas noticias representan la parte central de los experimentos, siendo la variación de las cantidades utilizadas para entrenamiento y pruebas un factor común en cada experimento configurado. Los lapsos de tiempo entre los cuales están comprendidas las noticias por categoría se presentan en la Tabla I.

Se toma en cuenta la cantidad de noticias necesarias para completar el total de 2500 noticias por categoría, las cuales no tienen el mismo volumen por lapso de tiempo, es por esta razón que fue necesario tomar espacios de tiempo independientes por categoría en estudio.

Tabla I: Espacio de Tiempo por Categoría de las Noticias Almacenadas

Categoría	Desde	Hasta
Política	25/11/2015	03/02/2016
Economía	28/11/2014	03/02/2016
Mundo	03/08/2015	03/02/2016
Deportes	22/08/2015	03/02/2016
Sucesos	28/07/2015	04/02/2016
Salud	12/04/2013	04/02/2016
Tecnología	27/01/2014	04/02/2016

Para la clasificación se hace uso del principio de máxima entropía enfocado en la clasificación de textos, este algoritmo es proporcionado por la librería OpenNLP de Apache [4], la cual presenta un conjunto variado de herramientas para el procesamiento de texto en lenguaje natural. Los parámetros de configuración para el clasificador son:

- *Iterations*: cantidad de iteraciones de entrenamiento que hará el clasificador antes de generar el modelo.
- *Cutoff*: número mínimo de ocasiones en las que una palabra debe aparecer en el conjunto de textos de entrenamiento para ser tomada en consideración en el modelo.

Las tareas de PLN utilizadas para procesar los textos antes de su clasificación son los anteriormente descritos: reconocimiento de nombres, provisto por OpenNLP [4]; *stemming*, implementado por medio del uso de *Snowball* [17], un software para procesamiento de texto el cual permite manejar el proceso de *stemming* para el idioma español; y eliminación de palabras vacías, utilizando un documento previamente creado con un listado de las palabras que serán eliminadas.

Por lo tanto, de acuerdo a estos componentes, las variables que se ven modificadas entre distintos experimentos son:

- Cantidad de noticias de entrenamiento y pruebas por cada categoría.
- Valores para los parámetros del clasificador, *iterations* y *cutoff*, los cuales son requeridos al momento de generar el modelo.
- Técnicas de PLN a utilizar y orden de ejecución de las mismas.

Para medir la efectividad de los experimentos se utiliza el Valor-F o Medida F con un parámetro $\beta = 1$, siendo esto también llamado F1. Esta decisión está motivada en que no se pretende dar una importancia mayor a la precisión o la cobertura, por lo tanto, se utiliza dicha medida para mostrar de forma equitativa la influencia de ambos parámetros en los resultados generales de los experimentos.

Debido a que el resultado deseado por parte de la investigación comprende la selección de una combinación de parámetros para una serie de variables definidas, es necesario determinar el mejor valor para cada una de dichos parámetros, esto se ha conseguido al realizar la experimentación atravesando las etapas que a continuación se describen:

A. *Selección de los Datos*

Esta etapa comprende la realización de diversos experimentos, con el objetivo de establecer una cantidad de noticias que ofrezca buenos resultados en las fases de entrenamiento y pruebas. El objetivo es conseguir una proporción de documentos que brinde un alto nivel de eficiencia en la clasificación, sin llegar a presenciarse un sobreajuste en los datos.

Se ha desarrollado una serie de experimentos con esta premisa para buscar el punto en el cuál se pueda obtener un buen nivel de eficiencia disminuyendo la posibilidad de sobreajuste. Para esto se han utilizado variaciones de 50 noticias de entrenamiento por cada una de las siete categorías utilizadas (política, economía, mundo, deportes, sucesos, salud y tecnología), y utilizando siempre 500 noticias para pruebas por cada una de ellas, manteniendo el resto de variables ajustándose a los valores por defecto del algoritmo de máxima entropía utilizada.

Cabe destacar que todos los valores presentados durante el desarrollo del trabajo, corresponden al valor de F1 promedio obtenido al probar cada experimento, sin hacer énfasis en una categoría en particular ya que el objetivo es hacer una comparativa experimental del rendimiento global de los métodos utilizados.

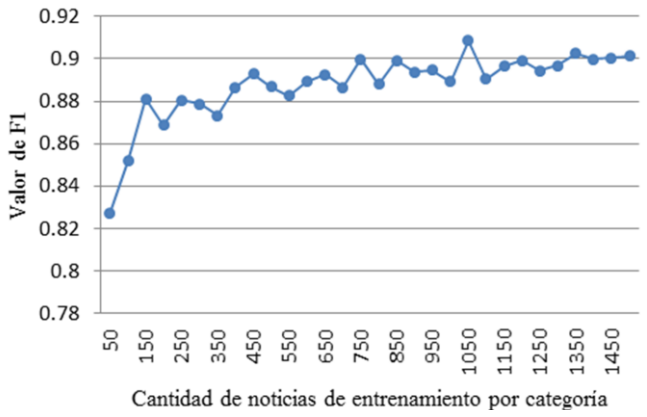

Figura 1: Variación de la Medida F1 con el Cambio en la Cantidad de Noticias de Entrenamiento por Categoría

En la Figura 1 se aprecia gráficamente el resultado de los experimentos descritos, se puede ver que inicialmente los valores van aumentando pero gradualmente tienden a alcanzar un cierto nivel de estabilidad luego de usar 1150 noticias de entrenamiento por categoría.

Nótese que la variación en los resultados luego de este punto no es muy significativa, a partir de aquí se considera el uso de entre 1150 y 1500 como rango de noticias de entrenamiento a usarse por cada uno de los experimentos que involucren modificación de técnicas a realizarse.

B. *Ajuste de Parámetros del Clasificador*

Seguidamente, se procede a definir valores para los parámetros base del clasificador, iteraciones y apariciones (*cutoff*).

Dado que en los trabajos revisados no se evidenció un análisis detallado de estos parámetros, se hace uso de los valores por defecto asignados por los desarrolladores del algoritmo como punto de partida, los cuales son: 100 iteraciones y un *cutoff* de 5; además se realizan variaciones independientes de dichos parámetros para verificar el impacto de cada uno de ellos en el resultado del clasificador.

Las variaciones sobre estos valores son realizados sobre los mismos experimentos ejecutados en la etapa anterior, cambiando en primer lugar el valor de iteraciones a 200, por otra parte el *cutoff* a 7, y en última instancia una mezcla de ambas variaciones, la comparación entre los resultados de éstos experimentos y el experimento base se observa en la Tabla II.

En la Tabla II, se presenta una comparación los valores de efectividad en término porcentual para el "estado estable" del clasificador en este experimento, se observa que en casi todos los casos los resultados son más altos usando 200 iteraciones y un *cutoff* de 5 para generar el modelo, alcanzando un valor promedio de F1 de 90,27% incrementando aproximadamente un 0,39% sobre la configuración predeterminada.

Por lo tanto, los experimentos creados a partir de este punto utilizarán 200 iteraciones y un *cutoff* de 5 como parámetros base para la generación del modelo.

Tabla II: Valores de F1 al Modificar Parámetros Base del Clasificador en el Estado Estable del Sistema

Cantidad de noticias de entrenamiento	**Variaciones de parámetros**			
	100 iteraciones / 5 cutoff	***200 iteraciones / 5 cutoff***	***100 iteraciones / 7 cutoff***	***200 iteraciones / 7 cutoff***
1150	89,6477	**90,1305**	89,3939	90,0186
1200	89,9158	90,353	90,0368	**90,3578**
1250	89,4282	**89,7412**	89,3981	89,7212
1300	89,6926	**90,0065**	89,6646	89,9205
1350	90,263	**90,9033**	90,1159	90,7889
1400	89,9722	90,3204	89,9171	**90,3521**
1450	90,0611	**90,239**	89,9144	90,0367
1500	90,1247	90,5267	90,0524	**90,6158**
Promedio	89,8881625	**90,277575**	89,81165	90,22645

C. *Pre-procesamiento y Clasificación*

Se realizan los experimentos correspondientes a una serie de combinaciones de las técnicas de PLN seleccionadas. En la Tabla III se listan las combinaciones de técnicas empleadas en experimentos, así como también el orden (representado numéricamente), en el que son ejecutados dentro de ellos, para posteriormente comparar sus resultados. Cada combinación es etiquetada con base en las técnicas que ejecuta y a su orden de aplicación; las siguientes etiquetas representan los nombres de las técnicas empleadas:

- N: Ninguna técnica.
- PVac: Eliminación de palabras vacías.
- RNom: Reconocimiento de nombres.
- Stem: *Stemming*.

La etiqueta "N" representa la ausencia de las tres técnicas en el experimento, marcando tal inexistencia con el texto "No

aplica"; la misma será utilizada como base de comparación para establecer la mejoría o empobrecimiento de los resultados obtenidos. Cada una de las combinaciones se probará en el estado de mayor estabilidad del clasificador.

Por ejemplo, la combinación RNom+Stem ejecutará en primer lugar el reconocimiento de nombres seguido del *stemming* (la eliminación de palabras vacías no se ejecuta ya que su columna indica el texto "No aplica"); en el caso de PVac+RNom+Stem se comienza ejecutando la eliminación de palabras vacías, posteriormente se utiliza el reconocimiento de nombres y finalmente el *stemming*.

Es necesario resaltar que debido a las propiedades del clasificador y buscando reducir el ruido en el análisis, al finalizar el procesamiento de los textos por cada combinación de técnicas, se eliminan los signos de puntuación y todas las palabras son transformadas a minúsculas.

Esto se debe a que los signos pueden causar ruido y una misma palabra se puede considerar en dos ocasiones si se presenta en mayúscula y minúscula, por ejemplo las palabas "Presidente" y "presidente" serían tomadas como dos entidades distintas, lo cual es solucionado al transformar todo el texto a letras minúsculas.

En la Sección VI se comparan los resultados obtenidos por cada una de las combinaciones usadas y se realiza un análisis detallado de los mismos, con la intención de obtener la configuración que brinde los mejores resultados al momento de clasificar documentos en el ámbito de estudio actual.

Tabla III: Combinaciones de Técnicas Usadas en los Experimentos

Combinación	**Técnicas aplicadas y orden**		
	Eliminación de palabras vacías	***Reconocimiento de nombres***	***Stemming***
N	No aplica	No aplica	No aplica
Stem	No aplica	No aplica	1
RNom	No aplica	1	No aplica
PVac	1	No aplica	No aplica
RNom+Stem	No aplica	1	2
PVac+Stem	1	No aplica	2
RNom+PVac	2	1	No aplica
PVac+RNom	1	2	No aplica
PVac+RNom+ Stem	1	2	3
RNom+PVac+ Stem	2	1	3

VI. Análisis de Resultados

En la Tabla IV se presentan los resultados obtenidos por cada combinación de técnicas de procesamiento de lenguaje natural probadas, al igual que antes los valores mostrados corresponden a la medida F1 en términos de porcentaje, la cual le asigna un peso igual a la precisión y cobertura, dando la posibilidad de evaluar ambas como un solo conjunto. Se pueden observar igualmente las combinaciones que brindan un mejor resultado por cada uno de los conjuntos de entrenamiento utilizados, así como el valor promedio de F1 para cada una de tales composiciones de técnicas.

Se puede observar en los resultados presentados, el incremento del valor promedio obtenido por las instancias que hacen uso de la eliminación de palabras vacías en comparación al resto. De hecho, al usar únicamente dicha técnica (columna *PVac*), se ha obtenido el segundo valor promedio más elevado con un F1 de 90,5863%, solamente por detrás de los experimentos bajo la combinación *PVac+Stem*, notándose así el impacto positivo de la eliminación de palabras vacías sobre el análisis.

Los experimentos realizados al utilizar únicamente el *stemming* (columna Stem), el reconocimiento de nombres (columna RNom) y la combinación de ambas (RNom+Stem), provocan una disminución en el valor medio de F1 con respecto al obtenido al promediar las instancias que no hacen uso de ninguna técnica. Sin embargo estos valores son mejorados al incorporar la eliminación de palabras vacías al análisis.

Un factor que puede influir en los resultados brindados por el clasificador, son las categorías de documentos tomadas en cuenta por el experimento y la correlación entre ellas. Si una clase guarda mucha relación con otra, varios de los documentos pertenecientes a ambas clases podrían ser asignados a la categoría incorrecta, un ejemplo de esto serían "política" y "economía", las cuales comúnmente están relacionadas principalmente en temas de carácter monetario.

Por lo tanto experimentos que se enfoquen en categorías con temas distanciados, como "política" y "deportes" tendrán mejores resultados con respecto a aquellos enfocados únicamente en clases con posibles correlaciones como "política" y "economía".

También se puede aumentar el error si se añaden categorías muy generales como "mundo", la cual presenta distintas noticias en diversas áreas siendo la única relación entre ellas su carácter internacional. Estos puntos se pueden observar en términos numéricos más adelante.

Al mismo tiempo se puede considerar que entre menor sea la cantidad de categorías estudiadas, se requerirá de menos documentos de entrenamiento por cada una para conseguir una efectividad elevada.

Esto se debe a que el clasificador necesitará hacer menos relaciones entre las entidades y las categorías, la misma línea de pensamiento se puede aplicar en el caso de que la cantidad de categorías estudiadas aumenten, situación en la cual el algoritmo requerirá de un número mayor de documentos de entrenamiento.

Siempre es necesario tener en consideración que en un entorno de clasificación automática de documentos, la calidad de los datos de entrenamiento es esencial. Por lo tanto se requiere tener una buena fuente de información para minimizar en lo posible el error de categorización ya que cualquier problema externo puede afectar también al clasificador, por ejemplo palabras mal escritas o el uso de documentos mal categorizados por la fuente en la fase de entrenamiento.

Habiendo considerado diversos factores que pueden afectar directamente los resultados de la clasificación, se prosigue con la selección de la configuración que brinda los mejores valores de F1 la cual será considerada como la propuesta para mejorar la tarea de categorización automática de documentos.

Tabla IV: Valor de F1 por cada Combinación de Técnicas de Procesamiento de Texto con Distintos Tamaños de Noticias de Entrenamiento

Noticias de entrenamiento	Combinación									
	N	***Stem***	***RNom***	***PVac***	***RNom+Stem***	***PVac+Stem***	***RNom+PVac***	***PVac+RNom***	***PVac+RNom+Stem***	***RNom+PVac+Stem***
1150	90,1305	90,2158	90,0817	90,4741	90,3085	**90,8743**	90,3705	90,2757	90,7568	90,6999
1200	90,353	90,7211	90,305	90,6105	90,5559	90,8211	90,3552	90,5495	90,9576	**91,0169**
1250	89,7412	89,6299	89,8577	**90,2848**	89,8389	90,1867	90,1694	90,202	90,1015	89,9291
1300	90,0065	89,78	89,9923	**90,4723**	89,569	90,1605	90,2841	90,2249	89,8604	90,0047
1350	90,9033	90,5394	90,5405	**90,9303**	90,691	90,9199	90,664	90,6665	90,6966	90,6115
1400	90,3204	90,5173	90,587	90,8141	90,3478	**91,0381**	90,4245	90,593	90,7567	90,6431
1450	90,239	90,1238	89,7342	90,5489	89,928	**90,5498**	89,9326	90,1836	90,2598	90,2691
1500	90,5267	89,8328	90,4647	90,5551	89,9779	90,2296	**90,7569**	90,6934	89,8856	89,8852
Promedio	90,2776	90,17	90,1954	90,5863	90,1522	**90,5975**	90,3697	90,4236	90,4094	90,3825

De la Tabla IV se extraen como los mejores resultados un 91,0381% obtenido al ejecutar la eliminación de palabras vacías seguida del *stemming* (*PVac+Stem*) en un entorno con 1400 noticias de entrenamiento por categoría, y un 91,0169% logrado al ejecutarse el reconocimiento de nombres, eliminación de palabras vacías y *stemming*, en ese orden (*RNom+PVac+Stem*), utilizando el conjunto de 1200 noticias de entrenamiento por cada categoría.

A simple vista, estos dos valores son prácticamente iguales, ambos representan una efectividad en la clasificación de 91%, por lo que cualquiera de las dos combinaciones podría ser tomada como la ideal, por lo tanto, se requiere un análisis más detallado de ambas opciones para determinar cuál de las dos debe ser seleccionada como la mejor.

En la Tabla V se presentan los resultados detallados de precisión, cobertura y F1 por cada categoría para la configuración correspondiente al usar la combinación *PVac+Stem* con 1400 noticias de entrenamiento por categoría. La misma clase de resultados se muestran en la Tabla VI para la combinación *RNom+PVac+Stem* con 1200 noticias por categoría para la fase de entrenamiento, de este modo se puede comparar con mayor detalle los valores de evaluación para estos experimentos. Al igual que en los casos anteriores los resultados son presentados en términos de porcentajes para todas las medidas de evaluación.

Tabla V: Valores de Evaluación Detallados al Usar la Combinación PVac+Stem con 1400 Noticias de Entrenamiento por Categoría

Categoría	Medida de evaluación		
	Precisión	***Cobertura***	***F1***
política	88,7129	89,6	89,1542
economía	89,4212	89,6	89,5105
mundo	87,2385	83,4	85,2761
deportes	96,6203	97,2	96,9093
sucesos	92,1606	96,4	94,2326
salud	93,2271	93,6	93,4132
tecnología	89,7541	87,6	88,664
Promedio	91,0192	91,0571	91,0381

Tabla VI: Valores de Evaluación Detallados al Usar la Combinación RNom+PVac+Stem con 1200 Noticias de Entrenamiento por Categoría

Categoría	Medida de evaluación		
	Precisión	***Cobertura***	***F1***
política	87,2624	91,8	89,4737
economía	89,6282	91,6	90,6034
mundo	85,1927	84	84,5921
deportes	95,9759	95,4	95,6871
sucesos	93,4263	93,8	93,6128
salud	93,5743	93,2	93,3868
tecnología	92,1776	87,2	89,6197

Categoría	Medida de evaluación		
	Precisión	***Cobertura***	***F1***
Promedio	91,0339	91	91,0169

En primer lugar, se puede notar que las categorías más distintivas o específicas brindan mejores resultados que otras con mayor correlación, esto se aprecia con "deportes", "sucesos" y "salud", las cuales presentan en ambos casos valores entre el 93 y el 96% mientras que otras como "política", "economía" y "tecnología" se acercan al 90% y "mundo", la menos específica de las categorías consideradas presenta los valores más bajos, alrededor de 85%.

Para verificar la igualdad de ambas configuraciones enunciada anteriormente, se empleó una prueba de hipótesis para comparación de medias con varianza conocida y muestras normales, o prueba Z cuya fórmula se enuncia en [18] y [19], utilizando los valores respectivos de este experimento como se muestra a continuación:

$$Z = \frac{91{,}0228 - 90{,}9965}{\sqrt{\frac{15{,}8392}{7} + \frac{13{,}3243}{7}}} = \frac{0{,}0263}{2{,}0411} = 0{,}0129 \approx 0{,}01 \quad (2)$$

La hipótesis nula (H_0) planteada corresponde a la igualdad de ambas medias, significando esto que ambos métodos son equivalentes; mientras que la hipótesis alternativa (H_1) equivale a que dichos valores son diferentes siendo de la misma manera distintas ambas combinaciones.

Siendo la regla de decisión rechazar H_0 si:

$$|z| > z_{\alpha/2} \quad (3)$$

Para la prueba, se ha tomado un valor de $\alpha = 0{,}05$, siendo entonces el punto crítico $z_{\alpha/2} = 1{,}96$. Entonces, debido a que $Z = 0{,}01$ es menor que $z_{\alpha/2} = 0{,}025$, se encuentra dentro de la zona de aceptación de la hipótesis nula para el nivel de significancia de 0,05 por lo que se acepta H_0.

Por lo tanto, dado que dicha hipótesis establece que ambas medias son iguales, se puede concluir que no se han encontrado diferencias estadísticamente significativas entre ambas medias, o análogamente, no existe evidencia que demuestre que las medias de ambas muestras sean diferentes. Este argumento se asegura con un nivel de confiabilidad de 99,2% de acuerdo al valor-p calculado en la Ecuación (4).

$$p = P(|Z| > 2{,}16) = 0{,}992 \quad (4)$$

A causa de esto, se puede requerir de un mayor grado de subjetividad por parte del analista al momento de decidir cuál de las dos combinaciones será seleccionada por sobre la otra.

Para determinar cuál configuración se ajusta mejor al enfoque de un investigador, se pueden considerar los siguientes criterios adicionales:

A. Resultados para una Clase Específica

Puede que, por algún motivo, el interés del analista o el usuario interesado en la información sea mayor para una categoría en especial, por lo tanto, si en este caso se le da mayor importancia a "economía", la elección recaería en utilizar la combinación *RNom+PVac+Stem* aplicada en un entorno con 1200 noticias de entrenamiento por categoría, ya que presenta un valor de F1 de 90,6% a comparación del 89,51% presentado por la otra opción.

Sin embargo, si se considera más importante la categoría "deportes" debería ser seleccionada la combinación *PVac+Stem* con 1400 noticias de entrenamiento por cada categoría ya que brinda un F1 de 96,9% frente al 95,68% presentado por la segunda alternativa.

B. Cantidad de Información de Entrenamiento

La base de datos utilizada contiene alrededor de 2500 noticias disponibles por cada una de las categorías estudiadas. La primera de las configuraciones comparadas utiliza 1400 noticias de entrenamiento por categoría, esto corresponde a un valor cercano al 56% del total de noticias, mientras que el segundo objeto de estudio presenta un entorno con 1200 noticias de entrenamiento por cada categoría, siendo esto alrededor del 48% del total disponible.

Por lo tanto esta última debería ser la configuración seleccionada bajo este criterio, ya que requiere de una menor cantidad de documentos de entrenamiento para alcanzar un valor alto de F1, evitando así de la mejor manera la probabilidad de un sobreajuste durante el entrenamiento del clasificador.

C. Complejidad del Pre-procesamiento

Si se considera que entre más técnicas se utilicen para procesar el texto antes de pasarlo al clasificador (sea en fase de entrenamiento o pruebas), mayor será el tiempo necesario para clasificar los documentos deseados y obtener los resultados del proceso, en situaciones como la que se presenta lo más natural sería seleccionar aquella configuración que utilice la menor cantidad de técnicas posible.

En este caso sería seleccionada la primera de las dos configuraciones utilizadas ya que además de utilizar menos técnicas (dos contra tres), hace uso de las menos complejas en cuanto a requerimientos de procesamiento.

D. Media del Resto de Experimentos para la Combinación

Si se desea tomar en consideración los otros experimentos, con cantidad de noticias de entrenamiento distintas, para cada combinación se puede obtener un valor promedio de F1.

La primera configuración utiliza la combinación *PVac+Stem* la cual, como se muestra en la Tabla IV tiene un valor promedio de 90,5975% mientras que la segunda hace uso de la combinación *RNom+PVac+Stem*, esta presenta un resultado medio de 90,3825% siendo 0,215% peor que la anterior, por lo tanto en este caso sería seleccionada la primera configuración ya que presenta mejores resultados de forma general.

Al tomar en consideración cada uno de los puntos anteriores, sin darle mayor importancia a una categoría sobre otra, se considera que la mejor elección corresponde a la primera configuración comparada, ya que presenta una mejor media general utilizando un procesamiento del texto menos extenso. Por lo tanto, para el caso de estudio actual la configuración completa de los parámetros requeridos es la presentada en la Tabla VII.

Tabla VII: Configuración Utilizada para Obtener el Mejor Resultado en la Clasificación

Variable	Valor
Iteraciones	200
Cutoff	5
Porcentaje de noticias de entrenamiento por categoría	56%
Técnica de PLN #1	Eliminación de palabras vacías
Técnica de PLN #2	*Stemming*

Es necesario recalcar que los resultados son afectados directamente por la calidad de la información contenida en los documentos, ya que las técnicas de PLN mejoran o empeoran los valores dependiendo de la información que deben procesar. De la misma manera el clasificador, en este caso de máxima entropía, también depende completamente de los datos de entrada.

Es importante señalar que en este trabajo se utilizaron los parámetros ajustados a una prueba, los cuales no aplican necesariamente para todos los contextos. Por lo tanto, con un conjunto de datos de entrada diferentes, puede existir un conjunto de parámetros más ajustados a dicho cambio, presentándose finalmente esta investigación como antecedente y base para futuras investigaciones.

VII. CONCLUSIONES

La cantidad de información usada en la fase de entrenamiento es un factor de vital importancia en un sistema de clasificación automática de documentos, por lo tanto, es necesario encontrar y establecer un tamaño de datos de entrenamiento que logre una eficacia elevada evitando al mismo tiempo la posibilidad de sobreajuste.

Cada técnica de procesamiento de lenguaje natural tiene un efecto distinto sobre los textos, su efectividad puede variar dependiendo del entorno de información sobre el cual es aplicada, por lo tanto es necesario probar distintas técnicas y combinaciones de las mismas, verificando cuál de ellas es capaz de brindar el mayor incremento en la eficacia de la clasificación para el contexto estudiado.

Debido a que el clasificador depende totalmente de los datos etiquetados de entrenamiento y, además de esto, toma como principio el no asumir nada usando la uniformidad en las probabilidades, es necesario que las categorías estudiadas y la información a clasificar sea lo suficientemente específica o bien pre-procesada para producir resultados de mayor calidad, mostrando así la necesidad de una buena selección de

parámetros y procesos de PLN antes de empezar a clasificar los documentos.

Debido a las características de los métodos estudiados, es importante resaltar que los cambios realizados sobre cualquiera de las variables presentadas, pueden generar diversos cambios sobre los resultados de la clasificación.

Al concluir la clasificación, es tarea del analista seleccionar la combinación de parámetros y técnicas que mejor se ajuste a sus criterios de selección, ya que estos pueden variar de acuerdo al ámbito en el cual se realizan los experimentos, por lo que un análisis detallado de los resultados obtenidos será necesario constantemente.

VIII. Trabajo Futuro y Aportes

Con la presente investigación se ha conseguido obtener un método alternativo para la clasificación automática de documentos, el cual puede ser usado como base para futuros trabajos en dicho ámbito de investigación.

La mayor parte de antecedentes investigativos relevantes a la categorización automática de documentos se enfocan en textos en inglés, por lo que este trabajo muestra la efectividad de los métodos utilizados en cuerpos de texto escritos en el idioma español.

Ya que en la presente investigación solo se hizo una variación sobre la cantidad de iteraciones y el *cutoff*, es factible probar de forma más exhaustiva las variaciones de estos parámetros con la intención de encontrar la combinación de valores que mejor se ajuste al entorno de investigación estudiado.

El porcentaje de documentos aplicados para entrenamiento no es necesariamente igual en todos los entornos, para otros ámbitos se podría obtener un valor mayor o menor al 56% propuesto en esta investigación, por lo que es recomendable partir de este valor y probar distintas disminuciones e incrementos del mismo con la intención de encontrar el mejor ajuste para el entorno de textos estudiados.

Si bien solo se han probado tres técnicas de procesamiento de lenguaje natural, se ha podido verificar que la eliminación de palabras vacías es un primer paso al momento de procesar texto para su posterior análisis.

De igual manera, es altamente recomendable probar otras técnicas además de las usadas en esta investigación, pues cada una tiene un efecto diferente sobre el texto y no es totalmente posible determinar cual dará resultados positivos o negativos en el ámbito estudiado.

Algunas técnicas que podrían ser tomadas en consideración además de las que se han manejado durante esta investigación son: etiquetado gramatical, para realizar una desambiguación del significado de una palabra en un contexto; resolución de correferencias, con la intención de descubrir relaciones directas entre las entidades previamente encontradas en el texto; lematización, cuyo objetivo es procesar un grupo de inflexiones para obtener una única palabra o lema que las represente a todas.

Finalmente es necesario destacar que todo el procedimiento correspondiente a la creación de los experimentos, la selección y variación de los valores para cada uno de los parámetros utilizados, así como la determinación de los mejores valores en cada etapa, es realizado de forma manual. Por lo tanto, un avance altamente considerable e interesante, es la inclusión de tareas de inteligencia artificial orientadas a la selección de los valores óptimos para cada una de las variables estudiadas. De esta manera el algoritmo se encargaría de realizar las variaciones en los valores y ejecutar los experimentos para finalmente determinar la mejor configuración para el ámbito de investigación estudiado, haciendo una sintonización de parámetros de forma automática.

Referencias

[1] R. Eíto Brun and J. A. Senso, *Minería Textual*, El profesional de la información, pp. 11-27, 2004.

[2] S. Vijayarani, M. J. Ilamathi, and M. Nithya, *Preprocessing Techniques for Text Mining-An Overview*, International Journal of Computer Science & Communication Networks, vol. 5, no. 1, pp. 7-16, 2015.

[3] A. C. Vásquez, H. V. Huerta, and J. P. Quispe, *Procesamiento de Lenguaje Natural*, Revista de Investigación de Sistemas e Informática, vol. 6, no. 2, pp. 45-54, 2009.

[4] Apache OpenNLP. https://opennlp.apache.org

[5] A. F. Anta, L. N. Chiroque, P. Morere, and A. Santos, *Sentiment Analysis and Topic Detection of Spanish Tweets: A Comparative Study of NLP Techniques*, Procesamiento del Lenguaje Natural, no. 50, pp. 45-52, 2013.

[6] P. G. Otero and M. G. González, *Técnicas de Procesamiento del Lenguaje Natural en la Recuperación de Información*, Novática, no. 219, pp. 42-47, 2012.

[7] R. Feldman and J. Sanger, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. New York: Cambridge University Press, 2007.

[8] M. A. P. Abelleira and C. A. Cardoso, *Minería de Texto para la Categorización Automática de Documentos*, Cuadernos de la Facultad, Universidad Católica de Salta, no. 5, pp. 11-45, 2010.

[9] K. Nalini and L. J. Sheela, *Survey on Text Classification*, International Journal of Innovative Research in Advanced Engineering (IJIRAE), vol. 1, no. 6, pp. 412-417, Julio 2014.

[10] V. Gupta and G. S. Lehal, *A Survey of Text Mining Techniques and Applications*, Journal of Emerging Technologies in Web Intelligence, vol. 1, no. 1, pp. 60-76, Agosto 2009.

[11] C. Goller, J. Löning, T. Will, and W. Wolff, *Automatic Document Classification: A Thorough Evaluation of Various Methods*, 2000.

[12] A. G. Ramírez de la Rosa, *Clasificación Automática de Resumenes de Tesis Basada en Algoritmos de Agrupamiento Jerarquicos*, Universidad Tecnológica de la Mixteca, Oaxaca, México, Tesis de pregrado 2008.

[13] M. Emms and S. Luz, *Machine Learning for Natural Language Processing*, European Summer School of Logic, Language and Information, Dublin, Irlanda, 2011.

[14] K. Nigam, J. Lafferty, and A. McCallum, *Using Maximum Entropy for Text Classification*, in IJCAI-99 Workshop on Machine Learning for Information Filtering, pp. 61-67, 1999.

[15] A. Téllez, *Extracción de Información con Algoritmos de Clasificación*, INAOE, Tonatzintla, Tesis de maestría 2005.

[16] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, *From Data Mining to Knowledge Discovery in Databases*, AI magazine, 1996.

[17] Snowball. http://snowball.tartarus.org

[18] R. E. Walpole, R. H. Myers, S. L. Myers, and K. Ye, *Probabilidad y Estadística para Ingeniería y Ciencias*, Octava edición, Pearson Educación.

[19] T. Bartz-Beielstein, *Experimental Research in Evolutionary Computation*, Springer, 2006.