Un Enfoque Híbrido para la Clasificación Automática de Obras Literarias

Guillermo Rodriguez[1] Julián Litman[2] Alejandro Bolaños[3] Gabriel Hugo Taboada[4]
[1]ISISTAN-CONICET Research Institute, Argentina, Tandil.
[2]Universidad de Palermo, Argentina, Buenos Aires.
[3,4]Universidad CAECE, Argentina, Buenos Aires.
guillermo.rodriguez@isistan.unicen.edu.ar, julianlitman@gmail.com,
alejandro.daniel@gmail.com, gabriel.taboada@liveware.com.ar

Abstract. Al contar cada vez con mayor volumen de datos a analizar y que gran parte de estos sea en formato texto, resulta muy dificultoso para las personas poder entender y aprovechar el valor que ofrecen. La clasificación automática de textos consiste en asignar a un documento de texto una serie de clases utilizando técnicas de Machine Learning basado en su contenido y los temas que lo componen. La clasificación automática tiene importantes aplicaciones en la administración de contenido, la minería de opinión, el análisis de reseñas de un producto, filtros de SPAM y análisis de sentimiento en redes sociales. Este trabajo explica y detalla paso a paso una estrategia híbrida entre aprendizaje no supervisado y clasificación automática de textos basada en obras clásicas y literatura técnica trabajando sobre textos no estructurados y seleccionando las técnicas apropiadas para llegar a una aplicación concreta. Luego de realizar evaluaciones con conjuntos de libros, los resultados obtenidos permitieron verificar que nuestro enfoque es efectivo para la asignación automática de categorías a obras literarias.

Keywords: Categorización Automática de Textos; Text Mining; Machine Learning No Supervisado.

1 Introducción

En la actualidad se generan más de 2.5 quintillones de bytes por día [1]. Éstos son generados por cada persona cuando escucha música en el colectivo, compra algo por internet o dice que le gusta una foto en Facebook, entre muchas otras tareas que se realizan con los teléfonos o computadoras.

El 80% de estos datos son no estructurados, es decir, están compuestos por texto en lenguaje natural (hablado, escrito o visual), sonidos o imágenes y no son comprensibles para las computadoras de la misma forma que lo son para los humanos. Estos textos incluyen videos, correos electrónicos, artículos, y publicaciones en blogs revistas y redes sociales y canciones entre otros. Por otro lado, los datos estructurados, son los que encontramos en la mayoría de las bases de datos o planillas. Generalmente organizados en filas y columnas, y se procesan y ordenan fácilmente.

Teniendo en cuenta que el acceso a la lectura y escritura a través de nuevos formatos es el máximo visto en la historia y continúa creciendo de modo exponencial, del mismo modo que lo hacen las tecnologías, los datos no estructurados se vuelven una fuente de información cada vez más valiosa para analizar y entender.

En este contexto, se propone trabajar sobre datos no estructurados en formato de texto. Nuestro enfoque pretende estudiar si se pueden identificar temáticas de libros mediante el uso de técnicas automáticas de clasificación y procesamiento del lenguaje natural. Libros escritos desde el año 1600 hasta la actualidad nos permiten entender las realidades y conflictos que marcaron cada época. La relevancia histórica de estos textos son los que estipulan, justamente, la importancia de poder analizarlos con técnicas científicas de avanzada. De esta forma, es posible clasificarlos por temáticas basándose en su contenido.

Poder clasificar automáticamente textos por su género resulta útil en diversas áreas de estudio. La búsqueda veloz de información puede parecer la más obvia, es decir, poder encontrar eficientemente lo que uno necesita en grandes volúmenes de información, pero también vemos su aplicación día a día en filtros de SPAM, atribución de autoría en textos, entre otros. Poder identificar el género de un texto, sin tomarse el tiempo de leerlo, analizándolo computacionalmente, facilita la comprensión del mismo, ayuda al público en general a encontrar textos relacionados u obras del mismo autor. En el área profesional esto es utilizado, por ejemplo, para hacer la realización de auto-resúmenes en trabajos de investigación o en la búsqueda y categorización de un documento legal para un juzgado.

Este trabajo propone un enfoque híbrido basado en Sigular Value Decomposition (SVD) y Latent Dirichlet Allocation (LDA) para la clasificación automática de textos literarios. Para la evaluación del enfoque se recurrió a un corpus de 3545 libros en ingles. El resto del trabajo se organiza de la siguiente manera. La Sección 2 describe el marco teórico. La Sección 3 reporta el enfoque propuesto. La Sección 4 ilustra los resultados experimentales. La Sección 5 describe los principales trabajos relacionados en el área. Finalmente, la Sección 6 concluye el artículo y discute futuras líneas de trabajo.

2 Marco Teórico

Uno de los primeros ejemplos de resumen y clasificación fueron los catálogos de las bibliotecas públicas, siendo el más antiguo atribuido a Thomas Hyde (1764) para la Biblioteca Bodeliana de la Universidad de Oxford. Otro paso en el desarrollo del procesamiento de texto fue la generación automática de resúmenes para generar *abstracts*. Uno de sus primeros exponentes fue Hans Peter Luhn con su paper titulado "The Automatic Creation of Literature Abstracts" [11] donde escribe sobre aplicar métodos computacionales para generar *abstracts* de papers científicos haciendo un análisis de frecuencia de las palabras, es decir, cuantas veces se repite cada una para determinar la importancia relativa de cada una. La cantidad de veces que se repite cada palabra relacionada con la cantidad de frases donde está cada una genera una

métrica que define la importancia de cada frase en el paper extrayendo finalmente las frases más importantes y utilizandolas como *abstract* para el documento.

Las siguientes sub-secciones describen los principales conceptos relacionados con Text Mining (TM), Data Mining (DM) y Clasificación automática de textos.

2.1 Text Mining

TM se refiere al proceso de extraer sentido, patrones y estructuras ocultas dentro de texto no estructurado. Como la forma más natural de almacenar información es texto, se cree que el TM tiene un potencial comercial que todavía no está siendo explotado al máximo. De hecho, se estima que el 80% de la información que contiene una empresa es contenido por documentos de texto. Los procesos de TM suelen ser complejos ya que se involucran con datos no estructurados muchas veces confusos. Es un campo de aplicación multidisciplinario que involucra entre otros la búsqueda y recuperación de información, la categorización de documentos basadas en su contenido según temas predefinidos, procesos de extracción de información de interés en textos basado en su relevancia y el clustering o agrupación de textos basándose en sus características.

2.2 Data Mining

Según Mark Brown y John Brocklebank de la empresa SAS Institute [4] puede definirse como el proceso de seleccionar, explorar y modelar grandes volúmenes de datos con el fin de encontrar patrones desconocidos para obtener ventajas competitivas.

Data Mining (DM) aplica un conjunto de técnicas estadísticas y computacionales para realizar análisis exploratorios de datos. Esto no es algo novedoso, analistas vienen utilizando estadística y el armado de modelos similares hace años con el mismo fin. Lo que genera el auge que estamos viviendo es la reducción del costo del Hardware que hace posible acceder más rápido a grandes volúmenes de datos de manera económica.

Existen cientos de casos de uso dentro del DM en varias industrias como pueden ser la prevención de fraudes en bancos, la detección de comportamientos anómalos dentro de un grupo de clientes [5], análisis de canasta de mercado para detectar los hábitos de consumo en un supermercado [6] o el cálculo del churn dentro de una empresa telefónica para saber cuál es la probabilidad de que un cliente abandone la misma [7].

2.3 Clasificación Automática de Textos

La clasificación automática de textos consiste en asignar a un documento de texto una serie de clases utilizando técnicas de Machine Learning (ML) [8]. La clasificación se realiza generalmente basándose en la significancia o peso de las palabras o la extracción de características de documentos de texto.

La clasificación automática de textos puede realizarse de dos maneras distintas. La primera, llamada comúnmente "clasificación supervisada" [9], se asigna un texto a un conjunto de categorías predefinidas. La tarea del clasificador es asignar un nuevo documento a la categoría que corresponda.

En la segunda, llamada "clasificación no supervisada" o clustering, las categorías no están definidas previamente por lo que los documentos se agrupan según características que comparten en común, su contenido u otros criterios [10]. Todas estas prácticas pueden verse englobadas dentro de Data Mining, y más precisamente, Text Mining.

3 Enfoque Propuesto

El enfoque propuesto consta de cuatro etapas: Recolección de Datos, Preprocesamiento, Modelado de Documentos y Descubrimiento de Tópicos. Las siguientes sub-secciones describen las mencionadas etapas.

3.1 Recolección de datos

Todo proyecto de TM generalmente comienza con la recolección de datos, que van a ser la base en la que se van a realizar los análisis posteriores. Es fundamental que éstos sean de calidad para poder obtener conocimiento valioso de los mismos. La colección de texto a analizar se llama comúnmente Corpus. Éste puede estar formado por cualquier tipo de textos y, usualmente, puede ser representado a través de una matriz. La primera columna será el nombre del texto y la segunda su contenido, mientras que cada fila va a representar un texto.

3.2 Preprocesamiento

Existe una ley, llamada "Ley de Zipf" [12] que estipula que en una determinada lengua, la frecuencia de aparición de distintas palabras sigue una distribución que puede aproximarse de la siguiente manera:

$$Pn \sim {}^{1}/_{n^{a}}$$

donde Pn representa la frecuencia de la n-ésima palabra más frecuente y el exponente a es un número natural, en general ligeramente superior a 1. El segundo elemento se repetirá con una frecuencia de 1/2 del primero, el tercero con frecuencia 1/3 y así sucesivamente. En la Fig.1 puede verse una representación de esta ley, donde el eje de abscisas representa cada palabra ordenada por la cantidad de apariciones.

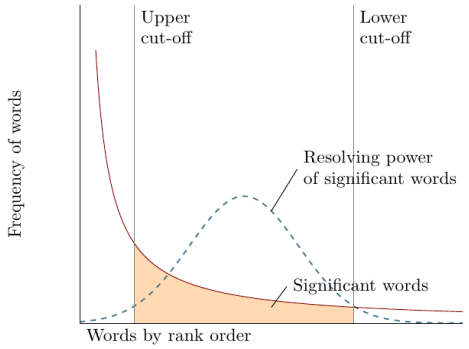

Figura 1. Diagrama de palabras por frecuencia.

Por ejemplo, según la RAE [13], las 10 palabras más frecuentes del español son *de, la, que, el, en, y, a, los, se y del* (Fig. 2). Esto quiere decir que la palabra "la" aparece la mitad de veces que la palabra "de" y que la palabra "que" aparece un tercio de veces que "de", al menos en textos normales.

Figura 2. Nube de palabras más frecuentes del español según la Ley de Zipf.

La ley de Zipf nos lleva a pensar que muchas de las palabras que se repiten sucesivamente en nuestros textos no van a tener valor significativo para la categorización, pero nos conviene removerlas para poder hacer el análisis correctamente.

El preprocesamiento consiste transformar las partes del corpus que no aportan significado al análisis y consta de los siguientes pasos:

- Eliminación de signos de puntuación, espacios innecesarios y conversión del texto a formato "plano", es decir, únicamente caracteres sin distintos tipos de letra, mayúsculas o negritas.
- Eliminación de stop words. Las *stop words* son un tipo de palabras generalmente cortas y muy utilizadas en el lenguaje, casi siempre ocupan un lugar fuertemente

gramatical en las oraciones y no agregan mucha información del contexto de la misma (pronombres, artículos, preposiciones, conjunciones, conectores, etc.).
- Lematización o normalización de palabras. En este paso se eliminan los sufijos y prefijos de una palabra para agruparlas según su raíz léxica. Por ejemplo, los vocablos *medicina*, *médico* y *medicinal* tienen la raíz léxica *medic.* Este proceso resulta interesante cuando se quiere ver la cantidad de veces que aparece una familia de palabras dentro de un documento.
- Identificación de nombres propios. Para realizar un buen análisis de TM es fundamental identificar nombres propios, sean de personas, organizaciones o empresas, así como las relaciones que existen entre ellos y los términos que aparecen en el document, como por ejemplo, términos como "La Casa Rosada".

3.3 Modelado de documentos

Para poder realizar análisis sobre los documentos es necesario representar su contenido mediante un modelo. En este trabajo se usaron 3 modelos: Modelo de Espacio Vectorial, Singular Value Decomposition y Latent Dirichlet Allocation.

Modelo de Espacio Vectorial

El modelo más popular para representación estructurada de texto es el Modelo de Espacio Vectorial (VSM), por sus siglas en inglés, el cual representa el texto como un vector, donde los elementos del mismo indican la cantidad de ocurrencias de cada palabras, o término, en el texto. Esto naturalmente resultará en vectores que contienen entre 10 mil y 35 mil elementos.
En nuestro caso, cada libro será un vector de palabras en un espacio de n dimensiones, siendo n el número de palabras en el texto.
El modelo VSM, por defecto, asume que el orden de las palabras en el documento no importa. Esto parece una gran suposición ya que naturalmente el texto debe ser leído en un orden específico para ser entendido. Al estar trabajando con clasificación de textos esto no va a ser un problema ya que las palabras o términos que aparezcan en el documento (en el orden que sea) generalmente son suficientes para ser tomados como diferentes conceptos.
Existen muchas maneras de determinar la importancia de un término dentro de un documento. La mayoría de las aproximaciones están basadas en dos afirmaciones:

- Cuanto más se repite una palabra en un documento, más importante es para el tema del mismo.
- Cuanto más se repite la palabra en todos los documentos de la colección, menos importante es.

Singular Value Decomposition & Latent Dirichlet Allocation

Para no tener que trabajar con toda la matriz completa en nuestro algoritmo de clasificación se utilizan técnicas para reducir la dimensionalidad de los datos, y aun

así, mantener la información significativa. Algunas de las técnicas utilizadas son Singular Value Decomposition (SVD) y el Latent Dirichlet Allocation (LDA).
Singular Value Decomposition (SVD) es una técnica de descomposición de matrices utilizada para separar los datos originales en componentes linealmente independientes. Estos componentes son una abstracción de los términos correlacionados que existen en la fuente de datos original. Generalmente los términos poco utilizados o de pesos despreciable pueden ser ignorados, lo que resulta en menos dimensiones a analizar [14].
En el modelo LDA cada documento es visto como una mezcla de los "tópicos" que estén presentes en ese corpus. El modelo propone que cada palabra del documento puede atribuirse a uno de los tópicos del documento. Es una técnica de reducción de dimensiones que opera sobre la matriz término-documento creada en el paso anterior [15]. La Fig. 3 muestra el clásico modelo de LDA dónde α es el parámetro de Dirichlet que indica la distribución a priori de palabras por tópico; β es el parámetro del Dirichlet previo a las distribuciones de tópicos por documento; θ es la distribución del tópico para el documento D; φ es la distribución de probabilidad por tópico T; Z es el tópico de la palabra W en el documento D.

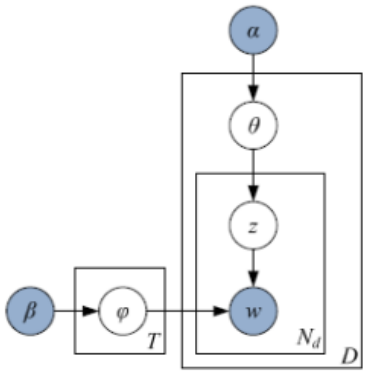

Figura 3. Modelo de LDA.

3.4 Descubrimiento de Tópicos

Un tópico es un vector con palabras o términos que cuentan con el peso que tiene cada palabra con el mismo. Por ejemplo, considerando cada uno de los siguientes libros como un corpus:

- Saga original de Harry Potter (7 libros)
- Saga original del Señor de Los Anillos (3 libros)
- Algunos libros de Shakespeare (King Lear, Othello, Romeo y Julieta y Macbeth)

El modelo LDA descubre los diferentes tópicos que el documento representa y cuánto de esos tópicos está presente en cada uno de los documentos. LDA produjo los siguientes tópicos ilustrados en la Tabla 1.

La Tabla 2 describe la probabilidad de ocurrencia de un tópico en cada uno de los documentos mencionados. A los tópicos construidos sobre una serie de textos pueden aplicarse nuevos documentos para saber a cuáles de los tópicos preestablecidos tienen mayor presencia.

4 Resultados Experimentales

La prueba de concepto se realizó en dos herramientas, SAS Enterprise Miner [1], un software propietario de la empresa SAS que permite realizar modelos descriptivos y predictivos basados en grandes volúmenes de datos, y RStudio[2], un entorno de desarrollo para el lenguaje estadístico R. En un principio, durante la etapa de carga y preparación de datos, el desempeño de ambas herramientas fue el deseado, pero cuando aumentamos la cantidad de libros, de unos pocos cientos a miles, optamos por SAS Enterprise Miner debido a que resultó ser más eficiente.

Tabla 1. Listado de tópicos de Harry Potter (Tópico 1), Lord of the Rings (Tópico 2) y Shakespeare (Tópico 3).

Topico 1: +harry,+wand,+professor,ministry		Tópico 2: +ring,+rider,+folk,+road,grey	
0.068	harry	0.08	ring
0.068	wand	0.079	rider
0.062	professor	0.075	folk
0.054	potter	0.074	road
0.054	ministry	0.072	grey
0.049	student	0.069	fellowship

Tópico 3: thou,thee,lady,dost,st	
0.142	thou
0.137	thee
0.11	lady
0.108	dost
0.104	st
0.098	hast

Tabla 2. Libros con su peso relativo por tópico.

Titulo	Topico 1	Topico 2	Topico 3
Harry Potter y la Orden del Fénix	0.523	0.207	0.163
El Retorno del Rey	0.023	0.675	0.311
Othello	0.017	0.043	0.562

[1] "SAS Enterprise Miner." https://www.sas.com/en_gb/software/enterprise-miner.html. Accedida en Abril de 2019.

[2] https://www.rstudio.com/

En nuestro caso, el corpus está formado por 3545 libros en inglés descargados de internet. Estos están en formato texto, con extensión .txt. y ocupan entre 1KB y 3MB. En el experimento, se seleccionó LDA como método no supervisado para la extracción de tópicos en grandes colecciones de documentos.
Dividimos los libros en dos grandes grupos, el primero servirá de para entrenar y validar los modelos de clasificación a realizar y el segundo para clasificar los libros en las categorías previamente definidas. Los libros del primer grupo van a ser un total de 3500 y 45 los del segundo. Luego, estos dos grupos serán preprocesados para su utilización en el armado de modelos.
Para entender mejor nuestros datos procedimos a armar un diagrama de dispersión con los términos y la cantidad de ocurrencias que tiene cada uno. La Figura 4 ilustra la Ley de Zipf, la cual permite entender que muchas de estas palabras no van a ser significativas y se procede a eliminarlas.

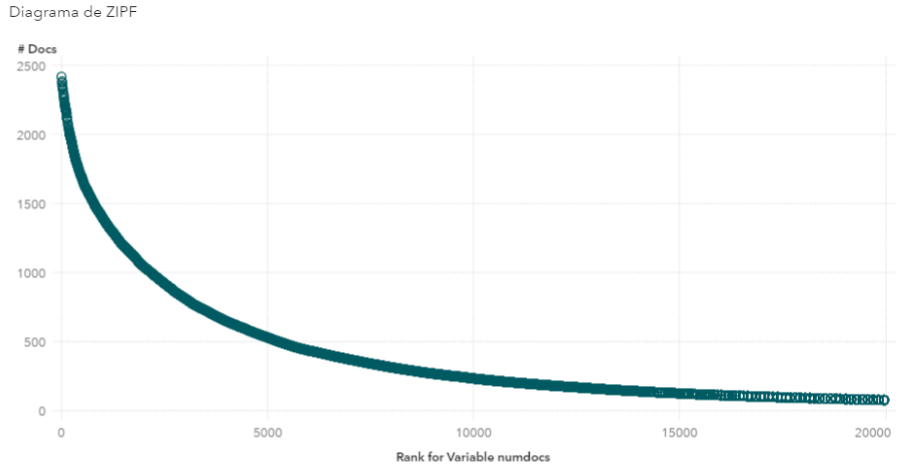

Figura 4. Representación de la Ley de Zipf realizada con los datos del caso de studio.

A su vez, dentro del entrenamiento de nuestro modelo encontramos que varios de nuestros documentos están fuertemente relacionados con este tópico. Todo el texto va a estar en minúscula, sin formato o signos de puntuación. Ya que trabajamos con libros y su contenido, eliminamos las stop word, o palabras que tienen principalmente utilidad gramatical en el idioma inglés (*I, me, we, the, then, a, an,...*) las preposiciones (*about, above, across, but…),* abreviaturas (por ejemplo, *etc.*) y valores numéricos (*one, two, three, four,…*).

Además, identificamos y removimos también los nombres propios, ya que no van a sumar a nuestro análisis al querer generalizar las categorías. Para la clasificación utilizamos principalmente sustantivos, verbos y adjetivos. Estas palabras pasaron por el proceso de lematización, donde extrajimos sus prefijos y sufijos, para obtener sus raíces.

El objetivo del próximo paso es obtener la matriz término-documento con la que vamos a trabajar en el modelado. Para esto, redujimos la cantidad de términos

aplicando técnicas de descomposición de dimensiones. Muchos de los términos con los que estamos trabajando son sinónimos o no son relevantes para los documentos con los que estamos trabajando, por ejemplo, palabras que pertenezcan a las categorías que vamos a analizar pero que se repiten consistentemente en todos los documentos *(Man, thing, time, hand).* Se establece el peso de cada palabra en cada documento, descartando las que tengan pesos más bajos.

Luego, obtuvimos finalmente los tópicos. Tras realizar diversas pruebas se decidió trabajar con 100 tópicos. Tomar muy pocos tópicos generará tópicos muy genéricos, mientras que usar demasiados va a resultar en tópicos demasiado segmentados.

Cada uno de estos tópicos están compuestos por términos pertenecientes a los diferentes documentos. Tomamos como ejemplo el tópico "Literatura científica y de Investigación". Éstos generalmente reciben el nombre de sus términos más fuertes por lo que decidimos renombralo manualmente tras conocer sus tópicos y textos más representativos (Tabla 3), donde el número de la izquierda es la probabilidad de que aparezca ese término en el tópico.

Tabla 3. Tópico 1 – Textos científicos y de investigación.

Tópico 1: Textos científicos y de Investigación	
0.125	species
0.109	commerce
0.097	commodity
0.09	industry
0.077	reasoning
0.077	manufacture
0.073	maxim
0.07	requisite

A su vez, dentro del entrenamiento de nuestro modelo encontramos que varios de nuestros documentos están fuertemente relacionados a este tópico. Para tomar otro ejemplo (Tabla 4), obtuvimos un tópico relacionado a historias de detectives cuyos términos principales son "Detective, Monk, Fiction, Novel y Deduction" y a documentos relacionados a la informática cuyos términos son "Software, Program, User, Computer y Technology".

Tabla 4. Documentos relacionados al Tópico 1.

Documentos relacionados al Tópico 1	
0.699	Of Money, and Other Economic Essays
0.682	Dialogues Concerning Natural Religion
0.335	On the Origin of Species
0.275	The Technique of the Mystery Story

Para finalizar con el análisis, tomamos una pequeña muestra de los 3500 del principio, exactamente 45 de diversas temáticas para evaluar la certeza de nuestro algoritmo clasificador. Se arma un modelo de scoring donde "comparando" el contenido de estos documentos con los tópicos creados anteriormente para obtener el valor de correspondencia entre estos. Así, podemos comprobar que cuando ingresamos un libro técnico a nuestro modelo se dará más peso a los tópicos relacionados con estos temas y menos a los de poesía, por ejemplo.

Tomamos, por ejemplo, el caso del libro Free Culture: The Nature and Future of Creativity de Lawrence Lessig cuyo subtítulo es "Cómo los grandes medios usan la tecnología y las leyes para encerrar la cultura y controlar la creatividad" y sus principales temáticas son los regímenes actuales de derechos de autor, la piratería informática y el copyleft según su reseña de Amazon [16]. No utilizamos este libro para la generación de tópicos por lo que es interesante analizar los resultados del scoring para saber cuánto se parece su texto a cada uno de los tópicos generados.

Tras procesar el libro podemos ver que este libro se compone mayoritariamente por los tópicos relacionados a textos técnicos (+computer, software, techology, user, copyright) con un 0.461 de probabilidad mientras no lo hace por tópicos relacionados a literatura oriental (Mandarin, tael, shan, dragon) con una probabilidad de 0.043. A su vez, resulta informativo analizar otros tópicos relacionados como es el caso de tópicos relacionados a textos económicos (economic, socialism, capitalist, industrial, labor) al que está relacionado y lo compone con una probabilidad del 0.167.

Fácilmente, sin conocer el contenido del libro o haberlo leído, podemos determinar que es un texto de género científico fuertemente relacionado a conceptos de libertad económica, relacionado a la industria del software y el copyright.

5 Trabajos relacionados

Numerosos métodos y técnicas híbridas han sido aplicados en el área de Machine Learning y minería de textos. El concepto de combinar clasificadores se propone como una estrategia para la mejora del rendimiento de clasificadores individuales. Recientemente muchos métodos han propuesto el ensamble clasificadores. Los mecanismos que se utilizan para construir un ensable de clasificadores incluyen: i) usar diferentes subconjuntos de datos de entrenamiento con un solo método de aprendizaje, ii) usar diferentes parámetros entrenamientos con un solo método de entrenamiento y iii) usar diferentes métodos de aprendizaje [18].

Los beneficios de los conjuntos de atributos locales y globales, y diccionarios locales versus globales han sido examinados en [19]. Los atributos locales son atributos dependientes de la clase, mientras que los atributos globales son atributos independientes de la clase. Los diccionarios locales son diccionarios dependientes de la clase mientras que los diccionarios globales son diccionarios de clase independiente. La mejor categorización del texto se obtiene utilizando los atributos locales y los diccionarios locales [19].

Un nuevo enfoque híbrido de clasificación de documentos de texto es propuesto en [20], que utiliza Naïve Bayes para la vectorización de datos crudos de texto, junto con

un clasificador SVM para clasificar los documentos con la categoría correcta. Los autores demostraron que el enfoque híbrido basado Naive Bayes y el clasificador SVM ha mejorado la precisión de clasificación en comparación con el enfoque de clasificación basado puramente en Naïve Bayes. En [21], los autores presentan otro método híbrido de Naive Bayes con mapas autorganizados (SOM). El clasificador Naïve Bayes es propuesto en la primera etapa, mientras que SOM realiza los pasos de indexación para recuperar los mejores casos de coincidencia.

En el contexto de la combinación de múltiples clasificadores para categorización de texto, numerosos trabajos han demostrado que combinar diferentes clasificadores puede mejorar la precisión de la clasificación [22]. Comparando el mejor clasificador individual y el método combinado, el rendimiento del método combinado es superior [23].

En [24], los autores proponen un método híbrido para clasificación de texto usando redes neuronales de propagación hacia atrás y la técnica de descomposición de valores singulares (SVD) para reducir la dimensionalidad y construir la semántica latente entre términos. La técnica SVD no solo pudo reducir en gran medida la alta dimensionalidad, sino también mejorar el desempeño. De esta manera, SVD contribuye en mejorar aún más los sistemas de clasificación de documentos de textos en términos de precisión y eficiencia. Finalmente, en [25], los autores proponen un algoritmo híbrido basado en k-NN y técnicas de Rocchio para mejorar la clasificación del texto. De esta manera, los autores demostraron que el algoritmo híbrido superó al algoritmo de Rocchio en términos de preicisón.

6 Conclusiones

Debido al creciente número de blogs, sitios y utilización de texto para comunicarse, la importancia de la clasificación automática de texto está tomando relevancia como nunca antes. La clasificación de textos puede ser automatizada, siempre que se hayan preprocesado y preparado los documentos con que se vaya a trabajar. Contar con datos de calidad, y grandes volúmenes, resulta crucial ya que la entradas que reciba el clasificador va a afectar la efectividad que tenga.

En este caso, al tener una buena base de libros para entrenar generamos un clasificador efectivo que puede discriminar entre distintos géneros y determinar, por ejemplo, si un documento es técnico, o no, y cuál es su contenido. Además, durante la realización del análisis encontramos libros que componen sagas, sin definirlo previamente, e influencia entre autores solamente por las palabras que utilizan. Podríamos afirmar que el modelado y la ejecución fueron correctos al ver las clases generadas y cómo se corresponden con los nuevos documentos.

Como trabajo futuro, se planea (1) buscar y obtener referencias en textos, facilitando encontrar cuando fue la primera vez que se mencionó un tema específico; (2) atribuir autoría e influencias textos científicos; (3) generar automática resúmenes para textos técnicos y literarios; (4) utilizar el lenguaje natural para la generación de contratapas para textos literarios, fortaleciendo el trabajo de los editores; (5) establecer una evaluación con sitios web de referencia (ej. Amazon y Ebay) y encontrar la si-

militud entre los tópicos asignados por nuestro efoque a cada libro de texto, y las categorías asignadas por dichos sitios a esa misma colección de libros.

Referencias

1. R. Jacobson, www.ibm.com, 24 Abril 2013. [Online]. Available: https://www.ibm.com/blogs/insights-on-business/consumer-products/2-5-quintillion-bytes-of-data-created-every-day-how-does-cpg-retail-manage-it/.
2. Proyecto Gutenberg, [En línea]. Available: http://www.gutenberg.org.
3. M. d. Cultura, Ley de Propiedad Intelectual. 1996.
4. J. y. B. M. Brocklebank, Data Mining 2008. [En línea]. Available: http://www2.sas.com/proceedings/sugi22/DATAWARE/PAPER128.PDF.
5. S. Wang, A Comprehensive Survey of Data Mining-Based Accounting-Fraud Detection Research, IEEE, Changsha, 2010.
6. A. Trnka, Market Basket Analysis with Data Mining methods, IEEE, Trnava, 2010.
7. S. &. W. H. Hung, Applying Data Mining to Telecom Churn Management, PACIS, 2004.
8. Tellez, E. S., Moctezuma, D., Miranda-Jiménez, S., & Graff, M. (2018). An automated text categorization framework based on hyperparameter optimization. Knowledge-Based Systems, 149, 110-123.
9. F. y. S. F. Debole, Supervised term weighting for automated text categorization, *Proceedings of the 2003 ACM symposium on Applied computing*, pp. 784-788 , 2003.
10. M. K. D. y. M. A. Zaveri, Automatic Text Classification: A Technical Review, *International Journal of Computer Applications*, vol. 28, nº 2, pp. 37-40, 2011.
11. H. P. Luhn, The Automatic Creation of Literature Abstracts, *IBM Journal*, pp. 159-165, 1958.
12. ILCE, El caos ordena la linguística. La Ley de Zipf., Instituto Latinoamericano de la Comunidad Educativa, [En línea]. Available: http://bibliotecadigital.ilce.edu.mx/sites/ciencia/volumen3/ciencia3/150/htm/sec_23.htm.
13. 1000 Palabras más frecuentes del español, RAE, [En línea]. Available: http://corpus.rae.es/frec/1000_formas.TXT.
14. R. Wicklin, The singular value decomposition: A fundamental technique in multivariate data analysis, 28 Agosto 2017. [En línea]. Available: https://blogs.sas.com/content/iml/2017/08/28/singular-value-decomposition-svd-sas.html.
15. D. M. Blei, A. Y. Ng y M. I. Jordan, Latent Dirichlet Allocation, *Journal of Machine Learning Research 3*, pp. 993-1022, 2003.
16. Amazon - Free Culture: The Nature and Future of Creativity, [En línea]. Available: https://www.amazon.es/Free-Culture-Nature-Future-Creativity/dp/0143034650.
17. SAS Enterprise Miner., [En línea]. Available: https://www.sas.com/en_gb/software/enterprise-miner.html.
18. M. Ikonomakis, S. Kotsiantis, V. Tampakas. Text Classification Using Machine Learning Techniques, Wseas Transactions on Computers, issue 8, volume 4, pp. 966- 974, 2005.
19. How, B. C. and Kiong, W. T. An examination of feature selection frameworks in text categorization. In AIRS. 558–564. 2005.
20. Dino Isa, Lam Hong lee, V. P Kallimani, R. RajKumar. Text Documents Preprocessing with the Bayes Formula for Classification using the Support vector machine, IEEE Transactions on Knowledge and Data Engineering, Vol-20, N0- 9 pp-1264-1272, 2008.
21. Dino Isa, V. P Kallimani Lam Hong Lee. Using Self Organizing Map for Clustering of Text Documents, Expert System with Applications. 2008.

22. Bao Y. and Ishii N. Combining Multiple kNN Classifiers for Text Categorization by Reducts, LNCS 2534, 340- 347, 2002.
23. Bi Y., Bell D., Wang H., Guo G., Greer K. Combining Multiple Classifiers Using Dempster's Rule of Combination for Text Categorization, MDAI, 2004, 127- 138, 2004.
24. C. Hua Li, S. C. Park, “An efficient document classification model using an improved back propagation neural network and singular value decomposition”. Expert Systems with Applications, 36, 3208–3215, 2009.
25. Miao, D., Duan, Q., Zhang, H., & Jiao, N. Rough set based hybrid algorithm for text classification. Expert Systems with Applications, 36(5), 9168-9174. 2009.