INSTITUTO POLITÉCNICO NACIONAL

ESCUELA SUPERIOR DE CÓMPUTO

Trabajo Terminal

CLASIFICACIÓN DE NOTICIAS DE DIARIOS DE CIRCULACIÓN NACIONAL MEDIANTE APRENDIZAJE AUTOMÁTICO

2017-A042

Presentan:

José Alejandro García Molina

Luis Enrique Ramírez Roque

Miguel Angel Sanchéz Ramírez

Directores

Dra. Consuelo Varinia García Mendoza

M. en C. Joel Omar Juárez Gambino

Ciudad de México,16 de Mayo de 2018

Instituto Politécnico Nacional

Escuela Superior de Cómputo

No. de registro : 2017-A042 Mayo 2018

Documento Técnico

“Clasificación de noticias de diarios de circulación nacional mediante aprendizaje automático”

Autores:

García Molina José Alejandro[1]

Ramírez Roque Luis Enrique[2]

Sanchéz Ramírez Miguel Angel[3]

RESUMEN

En este trabajo terminal se propone clasificar mediante técnicas de aprendizaje automático, noticias de diarios de circulación nacional en las diferentes secciones en que en estos se dividen, por ejemplo: cultura, deportes, política. La tarea de clasificar un diario en secciones se realiza manualmente, lo cual implica tiempo y esfuerzo por parte del editor. El trabajo contempla recolectar noticias de tres diarios de circulación nacional en las cuales el editor ya ha marcado a qué sección pertenecen. De estas noticias se extraerán diferentes características que servirán para utilizarlas en algoritmos de aprendizaje automático. Se probarán diferentes técnicas de extracción de características junto con diferentes algoritmos de aprendizaje automático, y al final se seleccionarán aquellos que obtengan mejores resultados. Con las técnicas seleccionadas, se podrán clasificar nuevas noticias en las secciones correspondientes de los diarios de forma automática.

Palabras clave: Clasificación de texto, aprendizaje automático, procesamiento de lenguaje natural.

Directores

Juárez Gambino Joel Omar **García Mendoza Consuelo Varinia**

1) alealejandromolina@gmail.com
2) luis.e.3194@gmail.com
3) miguelangelsanchezr2014@gmail.com

ESCUELA SUPERIOR DE CÓMPUTO
SUBDIRECCIÓN ACADÉMICA

DEPARTAMENTO DE FORMACIÓN INTEGRAL E INSTITUCIONAL

ESCOM

COMISIÓN ACADÉMICA DE TRABAJO TERMINAL

México, D.F. a 30 de Mayo de 2018.

LIC. ANDRÉS ORTIGOZA CAMPOS
PRESIDENTE DE LA COMISIÓN ACADÉMICA
DE TRABAJO TERMINAL
P R E S E N T E

Por medio del presente, se informa que los alumnos que integran el **TRABAJO TERMINAL:** 2017-A042, titulado **"Clasificación de noticias de diarios de circulación nacional mediante aprendizaje automático"** concluyeron satisfactoriamente su trabajo. Los discos (DVDs) fueron revisados ampliamente por su (s) servidor (a/es/as) y corregidos, cubriendo el alcance y el objetivo planteados en el protocolo original y de acuerdo a los requisitos establecidos por la Comisión que Usted preside.

Los discos (DVDs) fueron revisados ampliamente por sus servidores y corregidos, cubriendo el alcance y el objetivo planteados en el protocolo original y de acuerdo a los requisitos establecidos por la Comisión que Usted preside.

A T E N T A M E N T E

Omar Juárez Gambino

Juárez Gambino Joel Omar

Consuelo Varinia García Mendoza

Advertencia

”Este documento contiene información desarrollada por la Escuela Superior de Cómputo del Instituto Politécnico Nacional, a partir de datos y documentos con derecho de propiedad y por lo tanto, su uso quedará restringido a las aplicaciones que explícitamente se convengan.” La aplicación no convenida exime a la escuela su responsabilidad técnica y da lugar a las consecuencias legales que para tal efecto se determinen. Información adicional sobre este reporte técnico podrá obtenerse en: La Subdirección Académica de la Escuela Superior de Cómputo del Instituto Politécnico Nacional, situada en Av. Juan de Dios Bátiz s/n Teléfono: 57296000, extensión 52000.

Índice

Índice de figuras

Índice de tablas

1. Capítulo 1 Introducción

Hoy en día existen muchas fuentes de información disponibles en internet como son los blogs, redes sociales, foros, etc. Toda esta información requiere estar clasificada y etiquetada, para que los motores de búsqueda como Google o Yahoo puedan acceder a ella y recuperarla. Por lo tanto, la tarea de clasificar la información de acuerdo a las temáticas que aborda es muy importante, sin ella no se podría consultar dicha información.
La clasificación de información textual se puede hacer de forma manual utilizando a un experto que analiza su contenido y define que tópicos abordan. Sin embargo, el costo de esta tarea ha motivado el desarrollo de algoritmos para la clasificación de texto. Estos algoritmos han permitido automatizar tareas como la de clasificación bibliotecaria [1].

1.1. Problemática

La clasificación de textos ayuda a organizar y facilitar la recuperación de la información contenida en estos. Esta tarea se aplica a múltiples tipos de información documental como los libros de una biblioteca o notas periodísticas de un diario. En el caso particular de los diarios, la clasificación de la información en secciones es realizada generalmente de forma manual por el editor. El editor debe leer cada nota y de acuerdo a su contenido decide a que sección corresponde.
Por ejemplo, la nota de un concierto aparecerá en la sección cultural, mientras que el resultado de un partido de futbol aparecerá en la sección deportiva. Esta tarea requiere tiempo y esfuerzo para lograr clasificar correctamente la información antes de publicarla.

1.2. Solución Propuesta

Se propone desarrollar un clasificador de noticias el cual permita determinar de forma automática las secciones a las cuales pertenecen las noticias proporcionadas.
El trabajo contempla recolectar noticias de los diarios *El Universal*, *Excélsior* y *La Jornada* que ya han sido previamente clasificadas en secciones por los mismos, y siguiendo un enfoque supervisado, extraer de ellas características que puedan ser utilizadas para entrenar al clasificador.
Después de probar con varias características y algoritmos de clasificación, se seleccionarán aquellos que generen mejores resultados y se creará un modelo que será utilizado para clasificar nuevas noticias.
El trabajo propuesto es novedoso porque, aunque la tarea de clasificación de noticias de acuerdo a su temática ya se ha abordado antes, no se ha realizado con la intención de clasificarlas en las secciones de los diarios de circulación nacional. Esto permitirá que la tarea que actualmente se realiza de forma manual sea realizada de manera automática.

1.3. Objetivo

Clasificar de forma automática noticias de diarios de circulación nacional de acuerdo a su contenido en las diferentes secciones definidas por los mismos.

1.4. Objetivos específicos

1. Recolectar noticias de los diarios: *El Universal*, *Excélsior* y *La Jornada* (1500 noticias aproximadamente por cada diario).

2. Extraer características de las noticias como son la frecuencia de palabras y vectores que las caractericen.

3. Entrenar un clasificador probando diferentes algoritmos de clasificación como Naive Bayes, Máquinas de Soporte Vectorial y Regresión Logística.

4. Seleccionar las técnicas que obtengan los mejores resultados de acuerdo a la medida de exactitud. Esta medida determina la razón de las noticias clasificadas correctamente con respecto al total de noticias.

5. Crear un modelo para clasificar nuevas noticias a partir de las técnicas seleccionadas.

1.5. Estructura del documento

El presente trabajo consta de 6 capítulos organizados de la siguiente manera:

- **Capítulo 1:** Introducción, se especifica el objetivo del documento, se muestran los antecedentes y la justificación del proyecto.

- **Capítulo 2:** Estado del Arte, se mencionan los proyectos nacionales, internacionales y herramientas disponibles que se asemejan al cometido buscado.

- **Capítulo 3:** Marco Teórico, se describen los conceptos teóricos en los cuales se sustenta este trabajo.

- **Capítulo 4:** Análisis y diseño del sistema, se muestra el análisis del sistema, describe a los actores y roles que pueden interactuar con el sistema, los requerimientos funcionales y no funcionales, reglas de negocio, diseño de interfaces y todo aquello relacionado con el funcionamiento del sistema.

- **Capítulo 5:** Desarrollo, describe la implementación de la solución propuesta a la problemática planteada.

- **Capítulo 6:** Conclusiones, indica las conclusiones finales sobre el trabajo elaborado.

2. Capítulo 2 Estado del arte

2.1. Introducción

A continuación, se muestran trabajos nacionales, internacionales y herramientas desarrolladas que realizan un trabajo similar al que se propone.

2.2. Trabajos nacionales

2.2.1. Clasificación Automática de Textos de Desastres Naturales en México

Los autores de este trabajo [2] diseñaron un sistema que combina métodos de búsqueda de información y de clasificación de textos, con el objetivo de clasificar de forma automática textos de desastres naturales en la República Mexicana. Los resultados reportados indican que es posible clasificar una página web dentro de las categorías de huracán, inundación, sequía y no relevante con una exactitud aproximada del 97 %. Para ello crearon un repositorio de información de desastres naturales en México, el cual fue elaborado gracias a la información contenida en la página del diario *El Reforma*, donde recolectaron un total de 375 documentos de los cuales solo un 88.5 % fueron noticias relevantes. Primeramente, los textos pasan por un proceso de pre procesamiento el cual busca reducir el tamaño del mismo y gracias a ello lograron reducir un 52 % del texto original: posteriormente el texto es representado por vectores de palabras, pasa por un proceso de reducción de dimensionalidad donde finalmente empleando técnicas de clasificación estadística y aprendizaje automático como arboles de decisión, máquinas de vectores de soporte, Naive Bayes, etc. logran clasificar la información. Al combinar esos métodos se logró clasificar una página web dentro de las categorías de huracán, inundación, sequía y no relevante con una exactitud del 97 % utilizando un algoritmo simple de Bayes sobre el conjunto de entrenamiento reducido en su dimensionalidad mediante la técnica de ganancia de información.

2.3. Trabajos internacionales

2.3.1. Category Classification and Topic Discovery of Japanese and English News Articles

En este trabajo [3] se presenta un conjunto de algoritmos para clasificar noticias tanto en el idioma inglés como el japonés. Extrae las palabras clave de un texto mediante el algoritmo de extracción de palabras clave propuesto por Bracewell [4], calcula las similitudes entre los temas conocidos con las palabras clave, dichas palabras crean un vector de palabras clave donde los valores son las puntuaciones de palabra clave. Para comparar el vector de palabras clave del artículo y el vector de palabras clave del tema, los dos se transforman en el mismo espacio vectorial para ser comparados y con base en ello se determina si pertenece a un tema previamente declarado o uno nuevo.
En dicho trabajo se delimitaron 8 secciones posibles (para ambos idiomas) a las que puede ser clasificado el artículo proporcionado, las cuales son: Business, Politics, Crime and Misfortune, Health, Sports, Entertainment, Technology y Science and Nature.
Para las pruebas, utilizaron un conjunto de entrenamiento de 1000 artículos, los resultados de este trabajo mostraron una exactitud de 97.22 % en la clasificación para noticias en inglés y 95.8 % en las noticias en japonés.

2.3.2. Document Classification for Newspaper Articles

En este trabajo [5] se clasificaron de manera automática artículos periodísticos (específicamente del Instituto Tecnológico de Massachusetts). Se probaron los algoritmos de Naive Bayes, máxima entropía y de examinación de estructura de texto, junto con analizadores gramaticales probabilísticos. Se establecieron 6 secciones a las que puede pertenecer el artículo proporcionado, las cuales son:

- Arts
- Features
- News
- Opinion
- Sports
- World

Trabajaron con un corpus de 3000 artículos, cada una de las seis secciones tuvo 500 artículos, para las pruebas tuvieron un conjunto de entrenamiento de 120 noticias (20 por cada sección). Los autores reportan un 77 % de exactitud a la hora de clasificar noticias de dicho instituto utilizando un clasificador Naive Bayes multinomial.

2.4. Herramientas disponibles

Entre las herramientas de trabajo que son de gran utilidad para el procesamiento de lenguaje natural y aprendizaje automático se encuentran:

2.4.1. Microsoft Azure

Microsoft Azure [6], el cual es un conjunto de servicios en la nube ofrecido por Microsoft, entre los cuales se encuentra el servicio de Cognitive Services [7] del cual se extiende la aplicación llamada *Text Analytics* la cual brinda un servicio de procesamiento de lenguaje natural avanzado. Dicha aplicación cuenta con 3 funciones principales: análisis de sentimientos (brinda una respuesta sentimental del texto, ya sea positiva o negativa), extracción de frases clave y detección de idioma.

2.4.2. IBM Watson

Otra herramienta de gran utilidad es *IBM Watson* [8] Watson es un servicio inteligente de análisis y visualización de datos, el cual tiene la capacidad de responder a cuestionamientos de lenguaje natural, descubrir patrones y análisis predictivo automatizado hacia un texto en específico, esto gracias a la base de datos con la que cuenta, la cual tiene una gran cantidad de información de diversas fuentes y a los algoritmos que utiliza para el procesamiento de la información.

De acuerdo a la revisión realizada de los trabajos similares, ya se ha abordado el problema de clasificar de forma automáticas las noticias de los diarios en las diferentes secciones que los componen. Sin embargo, no se ha hecho para el idioma español y con diarios de circulación nacional. Por lo anterior, este trabajo terminal contribuirá en el desarrollo del estado del arte.

3. Capitulo 3 Marco teórico

En este apartado se expondrán algunos conceptos esenciales para el desarrollo de este trabajo terminal.

3.1. Lenguaje

De manera formal, el lenguaje se considera un conjunto generalmente finito de frases conformado por combinaciones de elementos de un conjunto (comúnmente infinito) llamado alfabeto, siendo aplicado un conjunto de reglas sintácticas (formación) y semánticas (sentido). Existen dos clases distintas de lenguaje: natural y formal [9].

3.1.1. Lenguaje natural

El lenguaje natural es el lenguaje que ha estado en constante cambio a través del tiempo con el objetivo de ser usado para la comunicación humana, como es el caso del inglés, italiano, español, etc. Se define a partir de una gramática, la cual se modifica constantemente. Cuenta con una capacidad expresiva gracias a la semántica [9].

3.1.2. Lenguaje formal

Es el lenguaje creado para poder expresar todo aquello relacionado con el conocimiento científico. Permite modelar teorías en diversas ramas de estudio ya que eliminan ambigüedades y eso es de gran ayuda en las ciencias exactas. Cuenta con un conjunto de componentes léxicos, semántica y reglas gramaticales. Al igual que el lenguaje natural cuenta con una gramática previamente establecida, pero con la diferencia que cuenta con un mínimo componente semántico el cual se puede incrementar con base en la teoría a realizar, no tiene ambigüedad y entre sus características principales esta la gran utilidad en el ámbito computacional debido a las características previamente dichas [9].

Para los fines de este proyecto se trabajará con el lenguaje natural, ya que las noticias están escritas en dicho lenguaje. Cabe destacar que el lenguaje natural es más difícil de procesar por una computadora que el lenguaje formal ya que presenta una gran cantidad de ambigüedades que pueden dificultar el tratamiento computacional.

3.2. Procesamiento de lenguaje natural

Debido a las características del lenguaje natural, la computadora no puede procesarlo de manera directa. Por lo tanto, es necesario utilizar técnicas que permitan que la máquina pueda manejar la información escrita en este lenguaje. El procesamiento de lenguaje natural permite desarrollar modelos que facilitan esta tarea [9]. A continuación, se describirán algunas de estas técnicas.

3.2.1. Tokenización

Es el proceso que descompone los textos de una colección en sus unidades mínimas, las palabras o términos propiamente dichos. A tales elementos se les denomina tokens que conforman una lista de ítems que se utiliza para su análisis estadístico, lingüístico, de almacenamiento y posteriormente de recuperación de información. Los tokens a su vez pueden ser identificados mediante una codificación ASCII o en su defecto hexadecimal, con el objeto de facilitar la identificación uno a uno cada caracter que compone la palabra. De hecho, este proceso permite la identificación de cadenas de caracteres de forma unívoca, de cara a posteriores tratamientos de depuración, eliminación de signos de puntuación o la reducción morfológica [10].

Ejemplo:
Precio máximo de la gasolina Magna: $16.28.

Tokens resultantes:

ID	1	2	3	4	5	6	7	8	9	10
Token	Precio	máximo	de	la	gasolina	magna	:	$	16.28	.

Tabla 1: Ejemplo tokenización

3.2.2. Lematización

Proceso de eliminación automática de partes no esenciales de los términos (sufijos, prefijos) para reducirlos a su parte esencial (lema) y facilitar la eficacia de la indización y la consiguiente recuperación de información [11]. Existen diversos grados de lematización debido a la ambigüedad de ciertas palabras. Por ejemplo, se puede realizar una lematización morfológica en la que únicamente se basa en la estructura de la palabra y no es su significado, en cambio, también se puede hacer una lematización semántica en la que se toma en consideración el contexto en el que aparece la palabra [12].

Ejemplo

$$soy, son, es \longrightarrow ser \tag{1}$$

$$coche, coches, cochecitos \longrightarrow coche \tag{2}$$

3.2.3. Stop words

Las palabras vacías, irrelevantes o stop words son aquellas que por sí solas carecen de significado y que, por su altísima frecuencia de aparición en los textos, generan un ruido innecesario para la recuperación de información. La eliminación de estos términos generalmente mejora la afinación en los modelos de recuperación [10]. Entre las palabras de este tipo en el idioma español se encuentran las preposiciones y los artículos principalmente.

En este trabajo se utilizarán las técnicas anteriormente mencionadas para procesar el texto contenido en las noticias.

3.3. Aprendizaje Automático

Aprendizaje Automático (AA) es la rama de la Inteligencia Artificial que tiene como objetivo desarrollar técnicas que permitan a las computadoras aprender. De forma más concreta, se trata de crear algoritmos capaces de generalizar comportamientos y reconocer patrones a partir de una información suministrada en forma de ejemplos. Es un proceso de inducción del conocimiento, es decir, un método que permite obtener por generalización un enunciado general a partir de enunciados que describen casos particulares [13].
Una de las tareas del AA es intentar extraer conocimiento sobre algunas propiedades no observadas de un objeto basándose en las propiedades que sí han sido observadas de ese mismo objeto (o incluso de propiedades observadas en otros objetos similares), es decir, predecir comportamiento futuro a partir de lo que ha ocurrido en el pasado.

3.3.1. Aprendizaje no supervisado

Es un tipo particular de AA en el cual no se tienen ejemplos previamente clasificados de los cuales aprender. Los algoritmos de este tipo son diseñados para desarrollar nuevos conocimientos mediante el descubrimiento de regularidades en los datos (data-driven). Estos métodos de aprendizaje no están dirigidos por las metas(goal-driven) [14].

3.3.2. Aprendizaje supervisado

La suposición fundamental de este tipo de método de aprendizaje es que los ejemplos proporcionados como entrada son necesarios para cumplir las metas del aprendizaje. En este tipo de método se dan ejemplos y estos se asocian a las categorías o clases a las cuales pertenecen. Una de las tareas de aprendizaje automático más importante es la clasificación, la cual consiste en determinar a que categoría o clase pertenece una instancia, a partir de las características que la definen. A continuación, se describen algunos de los métodos de clasificación más utilizados. En el trabajo terminal se utilizarán el enfoque de aprendizaje automático supervisado ya que se cuenta con ejemplo de noticias que ya han sido clasificadas en secciones previamente. A continuación, se describen algunos de los métodos de aprendizaje supervisados que se consideran en este trabajo.

3.3.2.1 Naive Bayes

La clasificación Naive Bayes son aproximaciones probabilísticas, las cuales hacen especulaciones sobre cómo deben ser generados los datos. Generalmente utilizan aprendizaje supervisado sobre el conjunto de entrenamiento para poder estimar los parámetros del modelo generativo, en tanto el conjunto de datos de entrada nuevos se realiza el teorema de Bayes, seleccionando la probable categoría que se ha generado [15].
El clasificador Naive Bayes es el más simple de estos clasificadores, en el que se supone que todos los ejemplos son independientes entre sí dado el contexto de la categoría.

3.3.2.2 Árboles de decisión

Un árbol de decisión es un clasificador conformado por un conjunto de nodo unidos por arcos, cada uno de los nodos simboliza una comparación/decisión, por lo general dicho nodo recibe un arco entrante y es el origen de un nodo saliente. Cuenta con nodos específicos llamados nodos raíz el cual como su nombre indica, no tiene un arco entrante pero es el origen de un nodo saliente.
A la hora de clasificar, se realiza una comparación al nodo raíz, el cual dirige el camino a un nodo descendiente, de dicho nodo cuelga una rama. Dicho proceso es recursivo ya que un nodo actúa como raíz de una nueva rama inferior, así hasta que se llega a un nodo terminal (también conocido como hoja) del cual ya no sale arco alguno. Las hojas tienen asociadas las clases que el clasificador va a asignar a las instancias que lleguen a dicha hoja [16].

3.3.2.3 Máquinas de soporte vectorial

Las máquinas de soporte vectorial (Support Vector Machines) son sistemas de aprendizaje los cuales se basan en el uso de un espacio de funciones lineales en un espacio de mayor dimensión inducido por un kernel, en el que las hipótesis son entrenadas por un algoritmo. [17]
En su origen, las máquinas de soporte vectorial fueron creadas con el objetivo de resolver problemas de clasificación binaria, pero se han utilizado para otro tipos de problemas (regresión, agrupamiento). Han sido implementadas en clasificación de imágenes, reconocimiento de caracteres, detección de proteínas, clasificación de patrones, identificación de funciones,etc.
Pertenecen a la categoría de los clasificadores lineales, debido a que inducen separadores lineales (también conocidos como hiperplanos), ya sea en el espacio original de los ejemplos de entrada, si

éstos son separables o cuasi-separables (ruido), o en un espacio transformado (espacio de características), si los ejemplos no son separables linealmente en el espacio original. La búsqueda del hiperplano de separación en estos espacios transformados, normalmente de muy alta dimensión, se hará de forma implícita utilizando las denominadas funciones kernel. Mientras la mayoría de los métodos de aprendizaje se centran en minimizar los errores cometidos por el modelo generado a partir de los ejemplos de entrenamiento (error empírico), el sesgo inductivo asociado a la SVM radica en la minimización del denominado riesgo estructural.
La idea es seleccionar un hiperplano de separación que equidista de los ejemplos más cercanos de cada clase para, de esta forma, conseguir lo que se denomina un margen máximo a cada lado del hiperplano. Además, a la hora de definir el hiperplano, sólo se consideran los ejemplos de entrenamiento de cada clase que caen justo en la frontera de dichos márgenes [17].

3.3.2.4 Regresión Logística

La regresión logística es una técnica estadística multivariante que nos permite estimar la relación existente entre una variable dependiente no métrica (donde la variable es binaria o también conocida como dicotómica, es decir, solo va a dar como resultado dos alternativas posibles) y un conjunto de variables independientes métricas o no métricas [18]. Un modelo de regresión logística permite, por ejemplo, predecir o estimar la probabilidad de que un individuo vaya a la Universidad una vez acabada a Enseñanza Secundaria (acudir a la Universidad, Y, sí/no) en función de determinadas características individuales (X,): nivel económico de la familia, edad en años, género (masculino/femenino), zona de residencia (rural/urbana, etcétera) [19].

3.4. Representación de texto

Los métodos de AA requieren que la información de la cual aprenderán esté representada en un formato que facilite su procesamiento. Generalmente esta representación es mediante vectores de valores numéricos. Cuando se requiere utilizar estos métodos con información en forma de texto, dicha información debe ser transformada para generar una representación más adecuada. A continuación, se describe uno de los modelos de representación más utilizados para información textual.

3.4.1. Modelo vectorial

En el modelo vectorial [20] se intenta recoger la relación de cada documento Di de una colección de N documentos, con el conjunto de las m características de la colección. Formalmente un documento puede considerarse como un vector que expresa la relación del documento con cada una de esas características. Este modelo es utilizado en operaciones de recuperación de información, al igual que en operaciones de categorización automática, filtrado de información entre otros. Dicho modelo busca recoger la relación de un documento Di perteneciente a una colección de documentos N.

$$D_i \rightarrow d_i = (c_{i1}, c_{i2}, ...c_{im}).$$

Es decir, ese vector identifica en que grado el documento D_i satisface cada una de las m características. En ese vector, C_{ik} es un valor numérico que expresa en qué grado el documento D_i posee la característica k. El concepto *característica* suele concretarse en la ocurrencia de determinadas palabras o términos en el documento, aunque nada impide tomar en consideración otros aspectos.

3.5. Selección de características

La selección de características es el proceso de detectar las características relevantes de una muestra de datos, así se puede eliminar todas aquellas características que no sean relevantes (no proveen suficiente información) o son redundantes (no proveen más información que con la que actualmente se tiene) [21].
En términos de aprendizaje supervisado, el feature selection brinda un subconjunto de datos con características seleccionadas usando uno de los siguientes enfoques:

- El tamaño especifico del subconjunto de datos que optimizan la medida de evaluación.
- El tamaño más pequeño de un subconjunto de datos que satisfacen ciertas restricciones sobre las medidas de evaluación.
- En términos generales, el subconjunto de datos entre la relación entre tamaño y medida de evaluación.

El uso correcto de los diversos algoritmos de de selección de características para seleccionar las mejores características incrementa el aprendizaje inductivo, la velocidad de aprendizaje e incluso puede reducir la complejidad del modelo.
Cabe destacar que dicho proceso en ocasiones puede tornarse complicado en el caso que el número de muestras sea mucho menor que las características ya que la búsqueda de dichas características se hará en un conjunto escaso, generando así que no se puedan diferenciar las características relevantes.

3.6. Métricas de evaluación de un modelo de aprendizaje automático

Una vez generando un modelo de clasificación, es importante medir el desempeño del mismo, con la intención de mejorar su eficiencia. Una de estas técnicas es la llamada matriz de confusión.

Matriz de confusión
La matriz de confusión es una representación de la información de los resultados obtenidos por un clasificador, dicha matriz suele ser de tamaño n x n, donde n es es el número de clases diferentes con las que se están trabajando [22].

	Predicción	
	Negativo	Positivo
Negativo	a	b
Positivo	c	d

Tabla 2: Ejemplo matriz de confusión 2x2

En el cuadro 2 se muestra un ejemplo de matriz de confusión con dos clases, la cual ejemplifica de manera adecuada las diferentes entradas de la misma, entra las que se encuentran:

- **a:** Número de predicciones correctas para la clase negativo
- **b:** Número de predicciones incorrectas para la clase positivo
- **c:** Número de predicciones correctas para la clase positivo
- **d:** Número de predicciones incorrectas para la clase negativo

La diagonal principal en cualquier matriz de confusión n x n representa el número de predicciones correctas para cada una de las n secciones.

Gracias a la matriz de confusión, es posible obtener ciertas métricas que nos ayudan a evaluar el modelo de aprendizaje. Entre las que se encuentran:

Exactitud: es la proporción del número total de predicciones que son correctas respecto al total. Se determina utilizando la ecuación:

$$P = (a+d)/(a+b+c+d) \tag{3}$$

Recall: Es la proporción de predicciones positivas que fueron correctamente clasificadas. Se determina utilizando la ecuación:

$$P = (d)/(c+d) \tag{4}$$

Precision Es la proporción de predicciones positivas que se clasificaron correctamente. Se determina con la siguiente ecuación:

$$P = (d)/(a+d) \tag{5}$$

F-Measure (F1): Se interpreta como la media armónica entre Precision y Recall. Se determina con la siguiente ecuación:

$$P = 2(precision * recall)/(precision + recall) \tag{6}$$

3.7. Validación Cruzada

La validación cruzada es una técnica utilizada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y prueba.
Consiste en repetir y calcular la media aritmética obtenida de las medidas de evaluación sobre diferentes particiones. Se utiliza en entornos donde el objetivo principal es la predicción y se quiere estimar que tan preciso es un modelo que se llevará a cabo a la práctica [23].

3.7.1. Validación cruzada de k iteraciones

En la validación cruzada de K iteraciones, los datos de muestra se dividen en K subconjuntos. Uno de los subconjuntos se utiliza como datos de prueba y el resto (K-1) como datos de entrenamiento. El proceso de validación cruzada es repetido durante k iteraciones, con cada uno de los posibles subconjuntos de datos de prueba. En la Figura 2 se muestra un ejemplo de los subconjuntos de prueba y entrenamiento creados con k igual a 4.

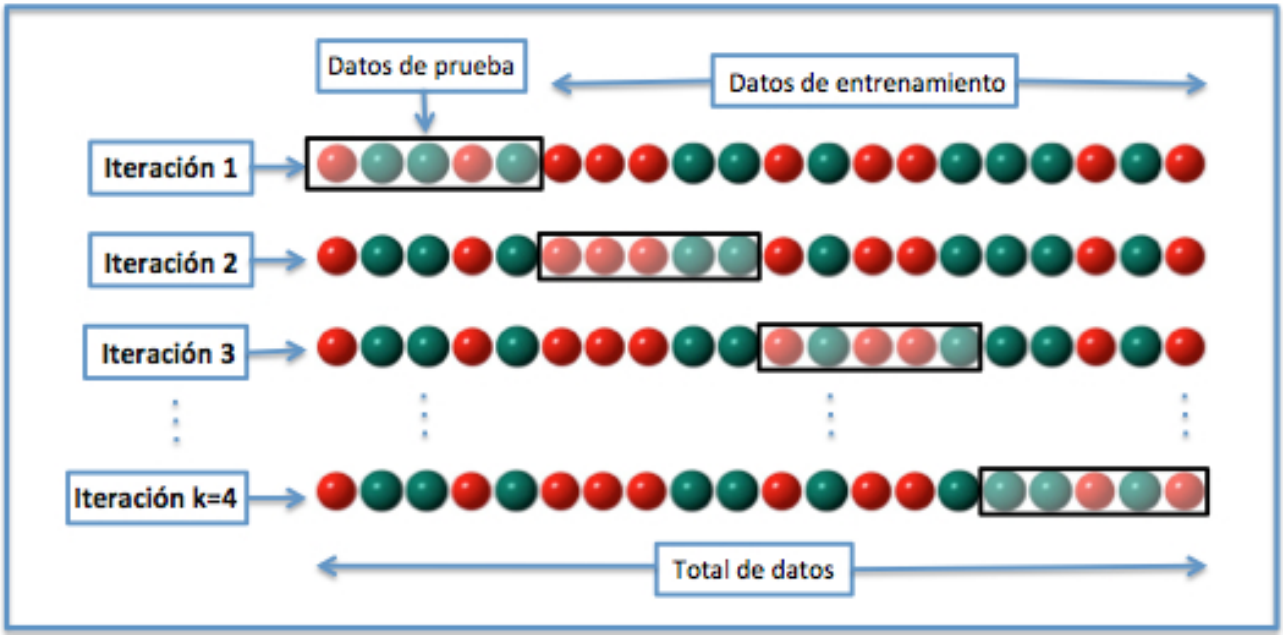

Figura 1: Ejemplo de validación cruzada con 4 iteraciones

Finalmente se realiza la media aritmética de los resultados de cada iteración para obtener un único resultado. Este método es muy preciso puesto que evaluamos a partir de K combinaciones de datos de entrenamiento y de prueba, pero aun así tiene una desventaja, y es que, a diferencia del método de retención, es lento desde el punto de vista computacional. En la práctica, la elección del número de iteraciones depende de la medida del conjunto de datos. Lo más común es utilizar la validación cruzada de 10 iteraciones, es una norma no establecida pero muy común en muchos trabajos relacionados al que se esta presentando [23].

4. Capitulo 4 Análisis y diseño del sistema

En esta sección se describe el análisis y diseño del sistema propuesto para la clasificación de noticias en secciones.

4.1. Actores y roles

- Usuario: Cualquier persona que esté interesada en clasificar noticias

4.2. Requerimientos funcionales

A continuación, se enlistan los requerimientos funcionales del sistema:

Número	Requerimiento	Descripción
RF01	Ingresar noticia	Se requiere de una interfaz para poder ingresar una noticia al clasificador, ya sea por medio de archivo o por cuadro de texto.
RF02	Procesar noticia ingresada	El sistema procesará las noticias ingresadas correctamente con una serie de algoritmos de tokenización y lematización.
RF03	Clasificar noticias ingresadas	El sistema clasificará las noticias ingresadas.
RF04	Mostrar resultados	El sistema deberá generar un archivo txt (el cual se encontrará en la misma carpeta en la que está contenido el sistema) el cual contendrá el resultado de la clasificación a la(s) noticia(s) ingresada(s), al igual que la cantidad de palabras y contenido de la(s) misma(s).

Tabla 3: Requerimientos funcionales

4.3. Requerimientos no funcionales

A continuación, se enlistan los requerimientos no funcionales del sistema:

Número	Requerimiento	Descripción
RNF01	La clasificación de una noticia no debe tardar más de 1 segundo	El resultado por noticia se debe mostrar en un máximo de 1 segundo.
RNF02	Codificación de caracteres	El sistema deberá procesar noticias codificadas en formato UTF-8.

Tabla 4: Requerimientos no funcionales

4.4. Reglas del negocio

A continuación, se enlistan las reglas del negocio del sistema:

Regla de negocio	Tipo	Descripción
RDN01	Restricción	Las noticias ingresadas al clasificador deben estar en el idioma español mexicano.
RDN02	Restricción	Las noticias ingresadas deberán tener un mínimo de 180 palabras en ellas.
RDN03	Restricción	El sistema solo permite ingresar archivos en formato txt.
RDN04	Restricción	El sistema solo podrá recibir un archivo como entrada para la clasificación.

Tabla 5: Reglas de negocio

4.5. Casos de Usos

A continuación, se muestran los casos de uso del sistema y se detallan sus trayectorias principales y alternativas.

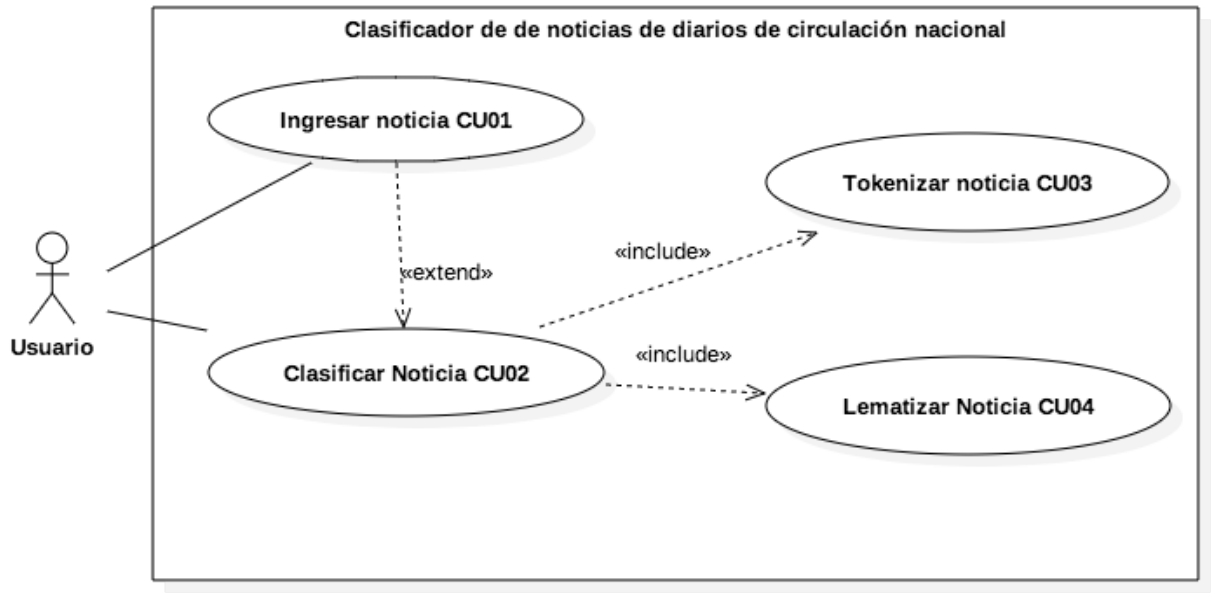

Figura 2: Diagrama de casos de uso

4.5.1. CU01 Ingresar noticia

Resumen

El usuario puede ingresar una noticia al clasificador para que este la evalué. Para ingresar una noticia se puede hacer dicha acción de dos maneras: ingresando directamente el texto en un campo y desde un archivo txt o csv.

Descripción

Nombre:	**CU01 Ingresar noticia.**
Descripción:	El usuario ingresa al sistema una noticia para que sea clasificada.
Precondiciones:	Agregar archivo o tener contenido en el campo de texto agregar noticia con un mínimo de 180 palabras.
Postcondiciones:	El sistema podrá analizar la noticia que se acaba de ingresar.
Situaciones de error:	Archivo dañado, tipo de formato no compatible, no se ingresaron noticias, no se ingresaron un mínimo de 180 palabras
Actores:	Usuario
Entradas:	Noticia(s)
Salidas:	Mensaje MSG01

Tabla 6: Caso de uso 01

Trayectoria Principal

1. El usuario solicita ingresar noticia desde archivo presionando el botón A1 de la pantalla IU-Menu. [Trayectoria A]
2. El sistema muestra la pantalla IU-CargarArchivo.
3. El usuario presiona el botón B1. [Trayectoria B]
4. El sistema muestra la pantalla IU-SeleccionarArchivo.
5. El usuario selecciona el archivo que desea clasificar.
6. El usuario presiona el botón C1. [Trayectoria C]
7. El sistema verifica que la noticia ingresada cumpla con la regla del negocio RND03. [Trayectoria D]
8. El sistema muestra la pantalla IU-IngresarArchivo.
9. El usuario solicita ingresar noticia presionando el botón B2. [Trayectoria B]
10. El sistema verifica que la noticia ingresada cumpla con la regla del negocio RND02. [Trayectoria E]
11. El sistema verifica que la noticia ingresada cumpla con la regla del negocio RND01.[Trayectoria F]
12. El sistema separa las noticias contenidas en el archivo.
13. El sistema carga el archivo seleccionado.

14. El sistema muestra el mensaje MSG01.
15. El sistema muestra la pantalla IU-Menu.
16. El sistema habilita el botón A3.
17. *– – Fin del caso de uso*

Trayectoria Alternativa A

1. El usuario solicita ingresar noticia desde texto presionando el botón A2 de la pantalla IU-Menu.
2. El sistema muestra la pantalla IU-IngresarNoticiaDesdeTexto.
3. El usuario ingresa en el cuadro de texto la noticia a clasificar.
4. El usuario presiona el botón D1. [Trayectoria G]
5. El sistema verifica que la noticia ingresada cumpla con la regla del negocio RND02. [Trayectoria E]
6. El sistema verifica que la noticia ingresada cumpla con la regla del negocio RND01. [Trayectoria F]
7. El sistema separa las noticias contenidas en el archivo.
8. El sistema ejecuta el CU01 desde el paso 14 de la trayectoria principal.
9. *– – Fin de la trayectoria.*

Trayectoria Alternativa B

1. El usuario presiona el botón B3.
2. El sistema muestra la pantalla IU-Menu.
3. *– – Fin de la trayectoria.*

Trayectoria Alternativa C

1. El usuario presiona el botón C2.
2. El sistema muestra la pantalla IU-Menu.
3. *– – Fin de la trayectoria.*

Trayectoria Alternativa D

1. El sistema muestra el mensaje MSG03.
2. El usuario presiona el botón Aceptar.
3. El sistema muestra la pantalla IU-Menu.
4. *– – Fin de la trayectoria.*

Trayectoria Alternativa E

1. El sistema muestra el mensaje MSG07.
2. El usuario presiona el botón Aceptar.
3. El sistema muestra la pantalla IU-Menu.
4. *– – Fin de la trayectoria.*

Trayectoria Alternativa F

1. El sistema muestra el mensaje MSG06.
2. El usuario presiona el botón Aceptar.
3. El sistema muestra la pantalla IU-Menu.
4. *– – Fin de la trayectoria.*

Trayectoria Alternativa G

1. El usuario presiona el botón D2.
2. El sistema muestra la pantalla IU-Menu.
3. *– – Fin de la trayectoria.*

4.5.2. CU02 Clasificar noticia

Resumen

El sistema clasifica una noticia en las secciones correspondientes al diario seleccionado.

Descripción

Nombre:	**CU02 Clasificar noticia.**
Descripción:	El usuario ingresa al sistema una noticia para que sea clasificada.
Precondiciones:	Cargar noticia adecuadamente al sistema.
Postcondiciones:	Resultado de clasificación elaborada.
Situaciones de error:	Archivo dañado, tipo de formato no compatible, no se ingresaron noticias
Actores:	Usuario
Entradas:	Noticia(s) ingresada(s)
Salidas:	Noticia(s) clasificada(s)

Tabla 7: Caso de uso 02

Trayectoria Principal

1. El usuario solicita clasificar la noticia ingresada presionando el botón A3.
2. El sistema muestra la pantalla IU-ClasificarNoticia.
3. El usuario selecciona el diario del cual desea saber la sección de la clasificación presionando la pestaña correspondiente E0.
4. El usuario presiona el botón E1. [Trayectoria A]
5. El sistema muestra el mensaje MSG02.
6. El sistema realiza el CU03 Tokenizar noticia.
7. El sistema realiza el CU04 Lematizar noticia.
8. El sistema muestra el mensaje MSG4. [Trayectoria C]
9. El sistema muestra los resultados de la clasificación en el archivo resultados.txt encontrado en la carpeta de descargas.
10. *– – Fin de la trayectoria.*

Trayectoria Alternativa A

1. El usuario presiona el botón E2.
2. El sistema muestra la pantalla IU-Menu.
3. *– – Fin de la trayectoria.*

Trayectoria Alternativa B

1. El sistema muestra el mensaje MSG05.
2. El usuario presiona el botón Aceptar.
3. *– – Fin de la trayectoria.*

4.5.3. CU03 Tokenizar noticia

Resumen

El sistema aplica el proceso de tokenización a la noticia a clasificar.

Descripción

Nombre:	**CU03 Tokenizar noticia.**
Descripción:	El sistema realiza el proceso de tokenización a una noticia a clasificar.
Precondiciones:	Cargar noticia adecuadamente al sistema, caso de uso CU05.
Postcondiciones:	Resultado de clasificación elaborada.
Situaciones de error:	Archivo dañado, tipo de formato no compatible, no se ingresaron noticias
Actores:	Usuario
Entradas:	Noticia(s) ingresada(s)
Salidas:	Noticia(s) clasificada(s)

Tabla 8: Caso de uso 03

Trayectoria Principal

1. El sistema realiza el proceso de tokenización descrito en el diagrama Diagrama-tokenizacion. [Trayectoria A]
2. El sistema ejecuta el caso de uso CU02 desde el paso 7 de la trayectoria principal.
3. *– – Fin de la trayectoria.*

Trayectoria Alternativa A

1. El sistema muestra el mensaje MSG05.
2. El usuario presiona el botón aceptar.
3. El sistema muestra la pantalla IU-Menu.
4. *– – Fin de la trayectoria.*

4.5.4. CU04 Lematizar noticia

Resumen

El sistema aplica el proceso de lematización a la noticia a clasificar.

Nombre:	**CU04 Lematizar noticia.**
Descripción:	El sistema realiza el proceso de lematización a una noticia a clasificar.
Precondiciones:	Noticia ingresada para clasificar tokenizada.
Postcondiciones:	Noticia ingresada para clasificar lematizada.
Situaciones de error:	Archivo dañado, tipo de formato no compatible, no se ingresaron noticias
Actores:	Usuario
Entradas:	Noticia ingresada para clasificar tokenizada.
Salidas:	Noticia ingresada para clasificar tokenizada y lematizada.

Tabla 9: Caso de uso 04

Trayectoria Principal

1. El sistema realiza el proceso de lematización descrito en el diagrama Diagrama-lematizacion. [Trayectoria A]
2. El sistema ejecuta el caso de uso CU02 desde el paso 8 de la trayectoria principal.
3. *– – Fin de la trayectoria.*

Trayectoria Alternativa A

1. El sistema muestra el mensaje MSG05.
2. El usuario presiona el botón aceptar.
3. El sistema muestra la pantalla IU-Menu.
4. *– – Fin de la trayectoria.*

4.6. Catálogo de mensajes

En esta sección se lista la relación de los mensajes utilizados para la herramienta de clasificación de noticias.

Notificación: Los mensajes de esta sección se refieren a todos aquellos que la herramienta muestra por diversas razones al usuario en la pantalla.

Alerta: Estos mensajes se muestran para notificar al usuario cuando se completa alguna acción.

Error: Estos mensajes se muestran cuando ocurre algún fallo en la herramienta o cuando no se pudo realizar alguna acción.

4.6.1. MSG01 Éxito al cargar el archivo

- Tipo: Notificación
- Indicar al usuario que el archivo que se cargó con éxito.
- Se ha cargado el archivo RUTA.
- El mensaje se muestra con base en los siguientes parámetros:
- RUTA: Ruta del archivo cargado en el sistema.

4.6.2. MSG02 Número de noticias agregadas

- Tipo: Notificación
- Indicar al usuario cuantas noticias fueron agregadas correctamente en el archivo ingresado.
- Agregando: N noticia(s).
- El mensaje se muestra con base en los siguientes parámetros:
- N: Número de noticias agregadas correctamente en el archivo ingresado.

4.6.3. MSG03 Error al cargar noticia

- Tipo: Error
- Objetivo: Notificar al usuario que el archivo que intentó ingresar una noticia que no cumple con el formato requerido.
- Redacción: Error al cargar la noticia n: No cuenta con el formato especificado.
- Parámetros: El mensaje se muestra con base en los siguientes parámetros:
- N: Número de noticia con error de formato especificado.

4.6.4. MSG03 Error al cargar noticias

- Tipo: Error
- Objetivo: Notificar al usuario que el archivo que intentó ingresar no cuenta con ninguna noticia con el formato requerido.
- Redacción: Ninguna noticia cuenta con el formato requerido.

4.6.5. MSG04 Ayuda de carga de noticias desde texto

- Tipo: Notificación
- Objetivo: Apoyar al usuario a cargar noticia desde texto.
- Redacción: Agrega el contenido de la noticia sín titulo en la caja de texto. Al dar click en agregar noticia, si el contenido tiene el formato necesario, se agregará la noticia y se limpiara la caja de texto. Las noticias ingresadas se acumulan, de manera que cada vez que se agregue una nueva noticia se acumulara mientras ls ventana esta abierta.

4.6.6. MSG05 Cantidad de palabras en noticia

- Tipo: Notificación
- Objetivo: Mostrar la cantidad de palabras con las que cuenta la noticia ingresada.
- Redacción: La noticia tiene: N palabras.
- Parámetros: El mensaje se muestra con base en los siguientes parámetros:
- N: Número de palabras de la noticia ingresada.

4.6.7. MSG06 Cantidad de palabras mínima en noticia

- Tipo: Error
- Objetivo: Informar que la noticia que se intenta clasificar mediante no cuenta con el mínimo número de palabras establecidas.
- Redacción: La noticia no cuenta con el mínimo número de palabras establecidas (180).

4.6.8. MSG08 No se encontró noticia

- Tipo: Error
- Objetivo: Notificar al usuario que no se ha encontrado la(s) noticia(s) a clasificar debido a que no la ha ingresado.
- Redacción: No se ha encontrado noticia para clasificar.

4.7. Pantallas

4.7.1. IU-Menu

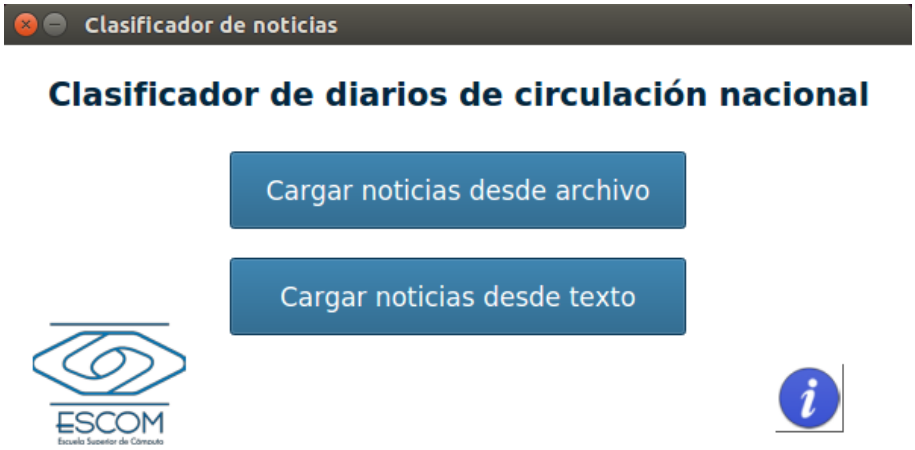

Figura 3: Pantalla IU-Menu

Objetivo

Llevar al usuario por los diferentes apartados que tiene la aplicación para su interacción y uso.

Diseño

La pantalla está conformada principalmente por cuatro botones, de los cuales es posible acceder a distintos apartados.

Salidas

- Mensajes de confirmación.
- Mensajes de ayuda.
- Acceso a pantallas del sistema.

Comandos

- Botón A1 Cargar noticias desde archivo: Permite acceder a la pantalla UI-CargarArchivo.
- Botón A2 Cargar noticias desde texto: Permite acceder a la pantalla IU- CargarNoticiaDesde-Texto.
- Botón A3 Información: Muestra mensaje de ayuda para orientar al usuario a realizar el proceso de clasificación de noticias.

Referencia

Caso de uso CU01, CU02.

4.7.2. IU-CargarArchivo

Figura 4: Pantalla IU-CargarArchivo

Objetivo

Ingresar un archivo existente en formato txt al sistema para que este posteriormente pueda clasificarlo.

Diseño

La pantalla está conformada principalmente por tres botones, de los cuales es posible acceder a distintos apartados del sistema.

Salidas

- Mensajes de confirmación
- Mensajes de ayuda.
- Acceso a pantallas del sistema

Comandos

- Botón B1 Agregar archivo: Permite ingresar una noticia selecionada previamente para poder ser analizada posteriormente.
- Botón B2 Clasificar noticias: Permite realizar el proceso de clasificación a la(s) noticia(s) previamente ingresada(s) correctamente.
- Botón B3 Regresar: Permite regresar a la pantalla IU-Menu.
- Botón B4 Información: Muestra mensaje de ayuda para orientar al usuario a realizar el proceso de carga de archivo.

Referencia

Caso de uso CU01.

4.7.3. IU-SeleccionarArchivo

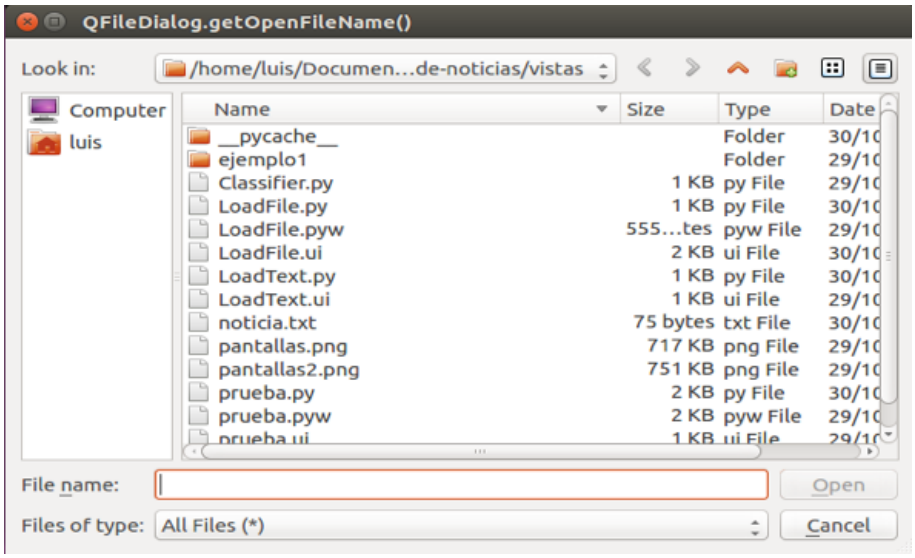

Figura 5: Pantalla IU-SeleccionarArchivo

Objetivo

Seleccionar un archivo existente en el sistema en formato txt que se desea ingresar para que este posteriormente pueda clasificarlo.

Diseño

La pantalla está conformada por 2 botones, una barra de navegacion donde se puede ubicar el archivo a ingresar y un combobox para poder filtrar por tipo de archivo todo lo contenido en la carpeta en la que se situa.

Salidas

- Se muestran todos los archivos contenidos en tu computadora y en discos externos conectados a la misma.
- Mensajes de confirmación

Comandos

- Botón C1 Abrir/Open: Permite al usuario ingresar una noticia desde archivo al sistema
- Botón C2 Cancelar/Cancel: Permite cancelar el proceso de ingresar noticia cerrando la pantalla IU-SeleccionarArchivo.

Referencia

Caso de uso CU01.

4.7.4. IU- IngresarNoticiaDesdeTexto

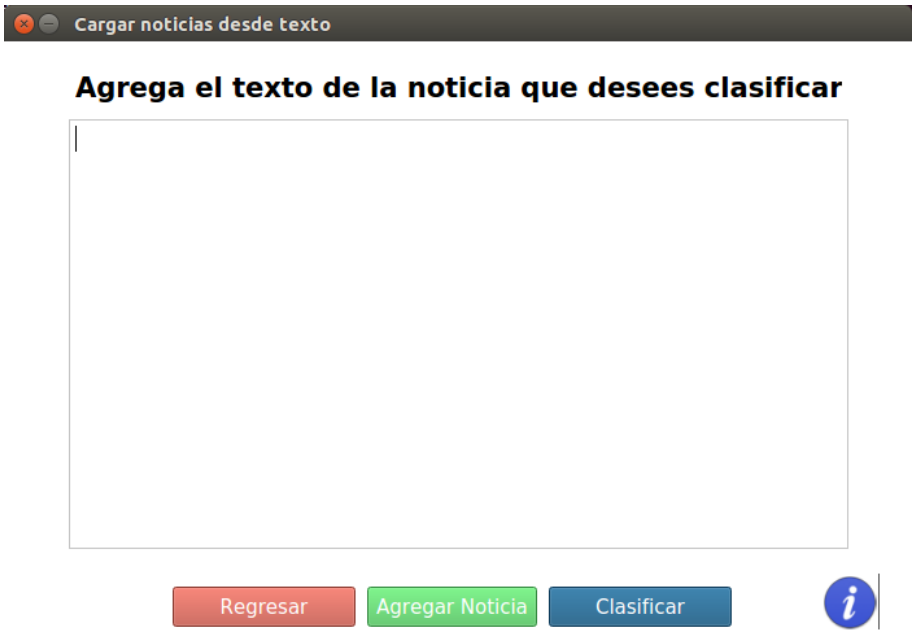

Figura 6: Pantalla IU- IngresarNoticiaDesdeTexto

Objetivo

Ingresar una noticia desde texto al sistema para que este posteriormente pueda clasificarla.

Diseño

La pantalla está conformada por un botón, que se encarga de ingresar la(s) noticia(s) desde texto y regresar a la pantalla UI-Menu para continuar con el proceso de clasificación.

Salidas

- Se muestra el contenido de la(s) noticia(s) de texto a clasificar.
- Mensajes de confirmación

Comandos

- Botón D1: Clasificar noticias: Permite realizar el proceso de clasificación a la(s) noticia(s) previamente ingresada(s) correctamente.
- Botón D2: Agregr noticia: Permite al usuario ingresar una noticia desde texto al sistema.
- Botón D3 Regresar: Permite cancelar el proceso de ingresar noticia cerrando la pantalla IU-SeleccionarArchivo.
- Botón D4 Información: Muestra mensaje de ayuda para orientar al usuario a realizar el proceso de agregar noticias desde texto..

Referencia

Caso de uso CU01.

4.7.5. IU-ClasificarNoticia

Figura 7: Pantalla IU-ClasificarNoticia

Objetivo

Clasifica una noticia previamente ingresada (mediante archivo o escrita) de acuerdo a la eleccion del usuario (El Universal, Excélsior o La jornada).

Diseño

La pantalla está conformada por 3 pestañas, donde cada una de ellas representa el diario destino del cual quiere obtener la clasificación; cuentan con 2 botones para clasificar o cancelar dicho proceso.

Salidas

- Noticia clasificada.
- Mensaje

Comandos

- Pestaña E0: Indica el diario del cual desea saber la sección la noticia ingresada.
- Botón E1 Clasificar: Inicia el proceso de clasificación de la noticia ingresada.
- Botón E2 Cancelar: Vuelve a la pantalla principal.
- Botón A3 Información: Muestra mensaje de ayuda para orientar al usuario a realizar el proceso de clasificación de noticias.

Referencia

Caso de uso CU02, CU05.

4.8. Mapa Navegación

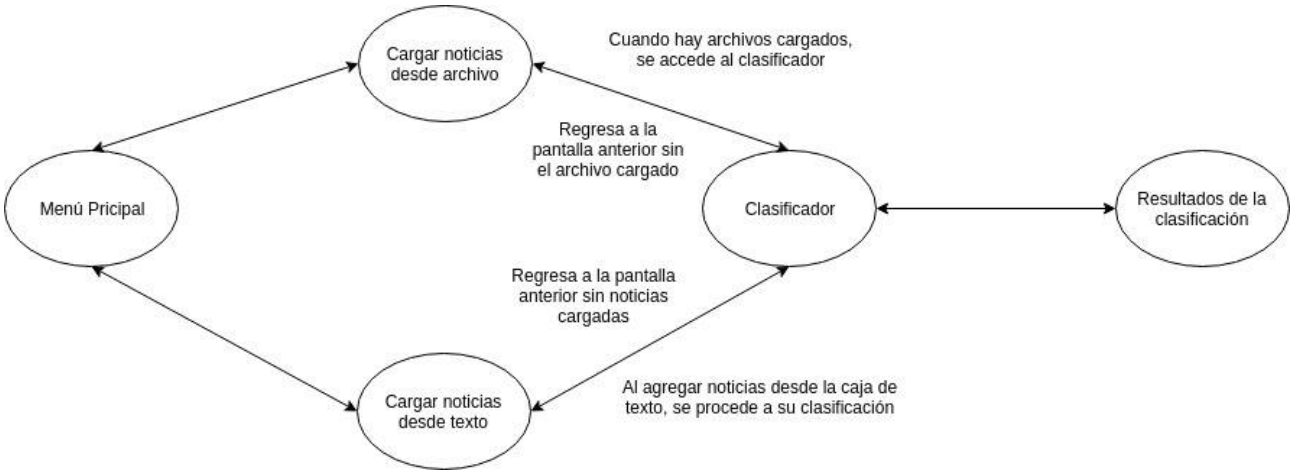

Figura 8: Mapa de navegación

En la Figura 8 se muestra el mapa de navegación, ilustrando los posibles rumbos que puede tomar el usuario por las diferentes pantallas del sistema.

4.9. Diagrama de actividades

Los diagramas de actividades muestran el procedimiento de las operaciones que se efectúan en el sistema.

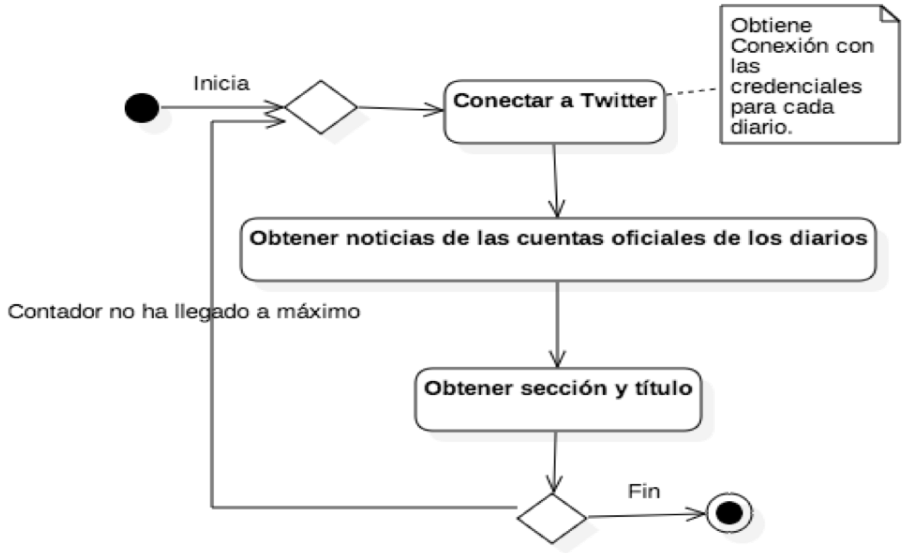

Figura 9: Diagrama de actividades proceso de obtención de noticias

En la figura 8 la primera actividad que se muestra es referente a la obtención de noticias de los principales diarios del país, la cual fue la primera tarea de este proyecto. Se empieza conectando a Twitter donde obtiene los tweets de dicho diario, el paso siguiente es obtener la url de cada tweet ya que representa el acceso a la noticia, finalmente obtiene la sección, título y noticia de dicha url.

4.10. Diagrama de actividades proceso de tokenización de noticias

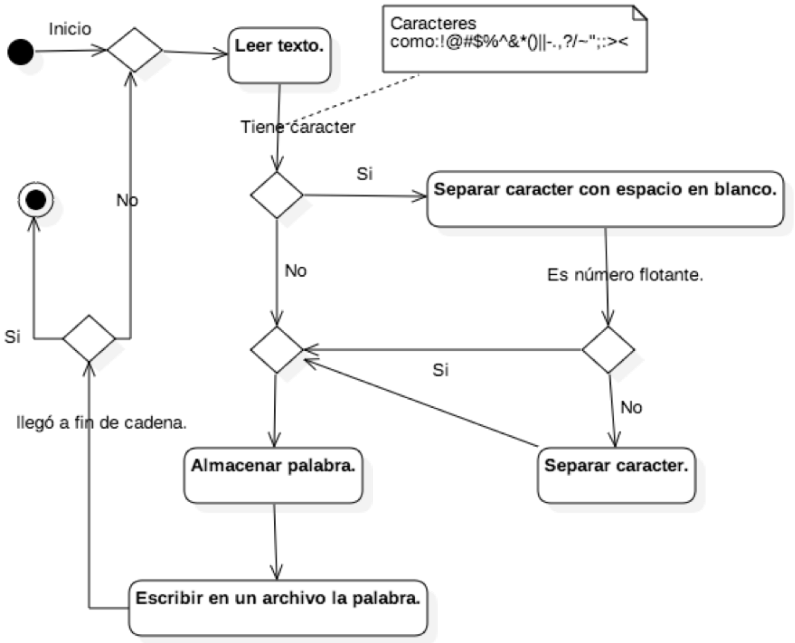

Figura 10: Diagrama de actividades proceso de tokenización de noticias

En la figura 9 el primer proceso que se hace es leer de un archivo de texto y valida si tiene algún carácter ilustrados en la figura, se separa con un espacio en blanco validando que no sea número flotante para posteriormente almacenar la palabra en un archivo, el proceso se repite mientras no sea fin de archivo.

4.11. Diagrama de actividades proceso de lematización de noticias

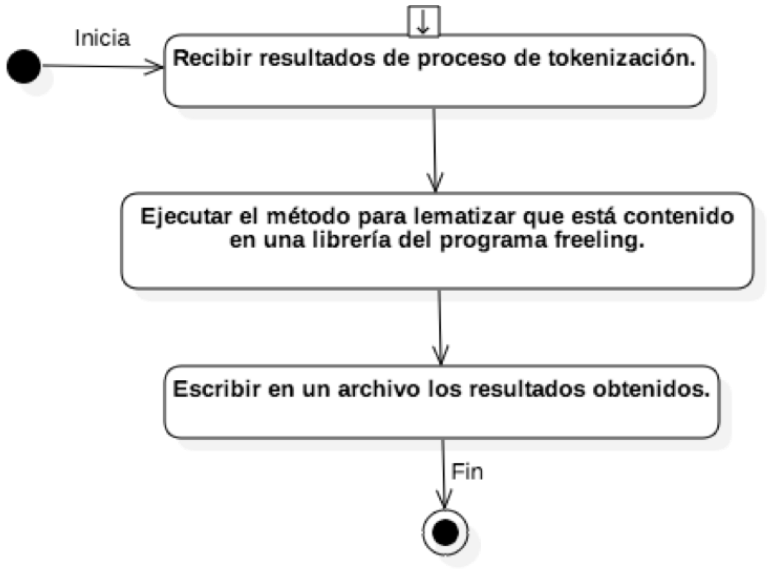

Figura 11: Diagrama de actividades proceso de lematización de noticias

En el proceso descrito en la Figura 11 recibe el resultado del proceso de tokenización descrito en la figura 10, después ejecuta el método de lematización implementado por las librerías de Freeling para al final escribir en un archivo los resultados obtenidos.

5. Capítulo 5 Desarrollo

En este capítulo se describen las etapas de cada uno de los procesos que se llevan a cabo para la clasificación de las noticias.

5.1. Elaboración de corpus

5.1.1. Selección de diarios de circulación nacional

Se han seleccionado los diarios de El Universal, Excélsior y La Jornada para trabajar con ellos como base para nuestro proyecto; cada uno fue elegido debido a factores como el reconocido prestigio de estos diarios, impacto en la sociedad mexicana, seriedad en cuanto a su periodismo, importancia como empresa y trayectoria.

5.1.2. Proceso de recolección de noticias

Primeramente, se tiene que definir un método y un medio para poder recolectar noticias. Debido a que en años recientes las redes sociales se han convertido en uno de los medios de comunicación más importantes, se ha decidido que la red social Twitter sea la fuente de información, esto se debe a dos principales razones:

- **Cambio en la hoja de estilo:** Si se buscan noticias directamente de la página web de cada uno de los diarios se pueden encontrar con diferentes problemas, como no encontrar url que permitan acceder a las diferentes noticias de la página o encontrarlas pero que no tengan el contenido deseado debido a que son publicidad u otro tipo de enlaces. Para ello se ha decidido conectarse a la cuenta oficial de Twitter de cada uno de los diarios y recuperar los tweets de los mismos. Esto debido a que la mayoría de las publicaciones de esas cuentas son las noticias más recientes e importantes de esos diarios y se incluye la url para acceder al sitio web del diario donde se muestra el contenido de dicha noticia. Este mecanismo permite recolectar las noticias y obtener la url para navegar por el gran catálogo de noticias de cada diario, obteniendo diferentes tipos de noticias que sirven para hacer más variado el conjunto de entrenamiento para el clasificador.

- **Twitter como medio de comunicación:** Un estudio [24] realizado en España muestra el resultado de una investigación en la que se busca es conocer el uso profesional que hacen los periodistas de diversas redes sociales (Facebook, LinkedIn, Twitter, YouTube, etc.), al igual que analizar para qué las utilizan y saber cuál es su percepción de las mismas y sus expectativas. La investigación se basa en una encuesta en profundidad realizada a medio centenar de periodistas españoles con perfiles activos en Twitter, con una edad media de 38 años y que llevan un promedio de 15 años ejerciendo la profesión. Los resultados revelan, por ejemplo, que los periodistas usan frecuentemente esta herramienta para publicar y distribuir información (95 %), identificar tendencias (86 %), buscar información (82 %), 'viralizar' información de sus propios medios de comunicación (82 %) o fidelizar a los usuarios (78 %). Sin embargo, sólo un 25 % de los encuestados dice utilizarla para realizar periodismo de investigación. Se puede concluir que en estos momentos Twitter representa un medio de comunicación para informar y ser informado al momento, gracias a su facilidad y velocidad de distribución de noticias que pueden ser compartidas o entregadas.

Para poder recolectar noticias de los diarios antes mencionados, se utilizó un script programado en el lenguaje de programación Perl el cual consiste en obtener las noticias publicadas en la red social Twitter. En la Figura siguiente se muestra el diagrama de flujo que describen los pasos que se siguen para obtener las noticias de los diarios antes mencionados.

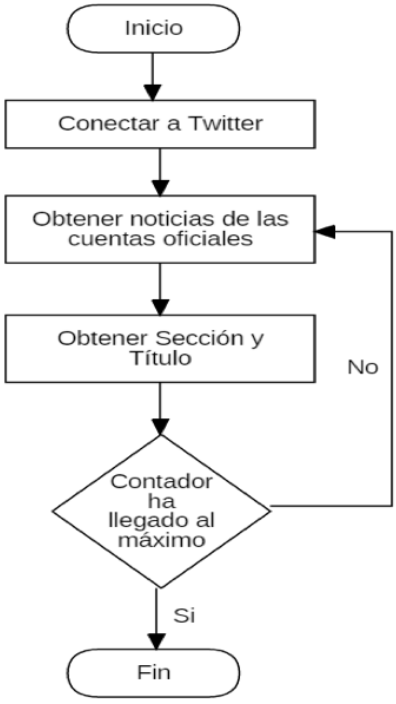

Figura 12: Proceso de recolección de noticias

5.1.3. Procesamiento de la noticia

Proceso de tokenización

Una vez recolectadas las noticias se pre-procesan con otro programa que tokeniza el archivo para poder manejarlo de una manera más simple. La tokenización tiene como propósito la separación de los elementos que conforman el texto a analizar. Un tokenizador o segmentador es el primero de los componentes que se necesitan para procesar las noticias en este proyecto.
En este caso, la tokenización se llevó a cabo a través de un programa realizado en lenguaje Python que se encarga de separar los caracteres alfabéticos de todos los demás. En el programa se generará como salida un archivo de texto que contiene el mismo texto de entrada pero esta vez separando todos aquellos signos de puntuación y símbolos, dejando así todas las palabras, signos y símbolos separados entre si por un espacio en blanco.
Para una muestra más clara del proceso de tokenización tenemos el siguiente ejemplo: Este es el texto de entrada.

Titulo: Presentan el prototipo de la primera moto voladora rusa

Cuando Radio Futura en su canción Enamorado de la moda juvenil decía aquello de "el futuro ya está aquí" justo antes del estribillo, tenía toda la razón. El futuro es hoy, tal y como va la velocidad de desarrollo y las nuevas tecnologías. Hace poco años, pensar en una moto voladora era más bien algo de ciencia ficción. Hoy, ya no lo es. Este es el prototipo de lo que se pretende sea la primera moto voladora rusa. Este modelo se llama Hoverbike S3 y la ha construido la empresa Hover Surf. Este vídeo fue grabado en el circuito de carreras de Moscú por Alexandr Atamanov, el director general de la empresa, en el que se ve cómo el aparato se eleva a un metro del suelo y comienza a volar. El mayor banco de Rusia, el VEB, ya ha invertido varios millones de dólares para el desarrollo de la misma. "El dispositivo se parece a una motocicleta, pero de hecho es más bien un prototipo de plataforma en la que se puede construir una multitud de vehículos. Su futuro es multifacético, existe la posibilidad de crear toda una familia de medios de transporte" aseguró Serguei Gorkov, Presidente del banco. Hover Surf ya ha recibido muchísimos prereservas de este modelo de Asia y de Oriente Medio, a pesar de que su precio es aún muy elevado, entre los 50.000 y 80.000 dólares, dependiendo de su configuración.

Figura 13: Noticia de prueba

Este es el texto que tendrá como salida el programa:

Titulo : Presentan el prototipo de la primera moto voladora rusa

Cuando Radio Futura en su canción Enamorado de la moda juvenil decía aquello de " el futuro ya está aquí " justo antes del estribillo , tenía toda la razón . El futuro es hoy , tal y como va la velocidad de desarrollo y las nuevas tecnologías . Hace poco años , pensar en una moto voladora era más bien algo de ciencia ficción . Hoy , ya no lo es . Este es el prototipo de lo que se pretende sea la primera moto voladora rusa . Este modelo se llama Hoverbike S3 y la ha construido la empresa Hover Surf . Este vídeo fue grabado en el circuito de carreras de Moscú por Alexandr Atamanov , el director general de la empresa , en el que se ve cómo el aparato se eleva a un metro del suelo y comienza a volar . El mayor banco de Rusia , el VEB , ya ha invertido varios millones de dólares para el desarrollo de la misma . " El dispositivo se parece a una motocicleta , pero de hecho es más bien un prototipo de plataforma en la que se puede construir una multitud de vehículos . Su futuro es multifacético , existe la posibilidad de crear toda una familia de medios de transporte " aseguró Serguei Gorkov , Presidente del banco . Hover Surf ya ha recibido muchísimos prereservas de este modelo de Asia y de Oriente Medio , a pesar de que su precio es aún muy elevado , entre los 50.000 y 80.000 dólares , dependiendo de su configuración .

Figura 14: Noticia Tokenizada

Como se puede observar en la figura anterior, se destaca que los caracteres que se contenian en la noticia original ahora se encuentran separados por un espacio en blanco (Por ejemplo, el titulo de la noticia contiene el signo de puntuación : pegado a la palabra Titulo, y despues del proceso de tokenización dicho signo de puntuacion y palabra se encuentran separados por un espacio en blanco. En la Figura siguiente se muestra el diagrama que describe el proceso de tokenización.

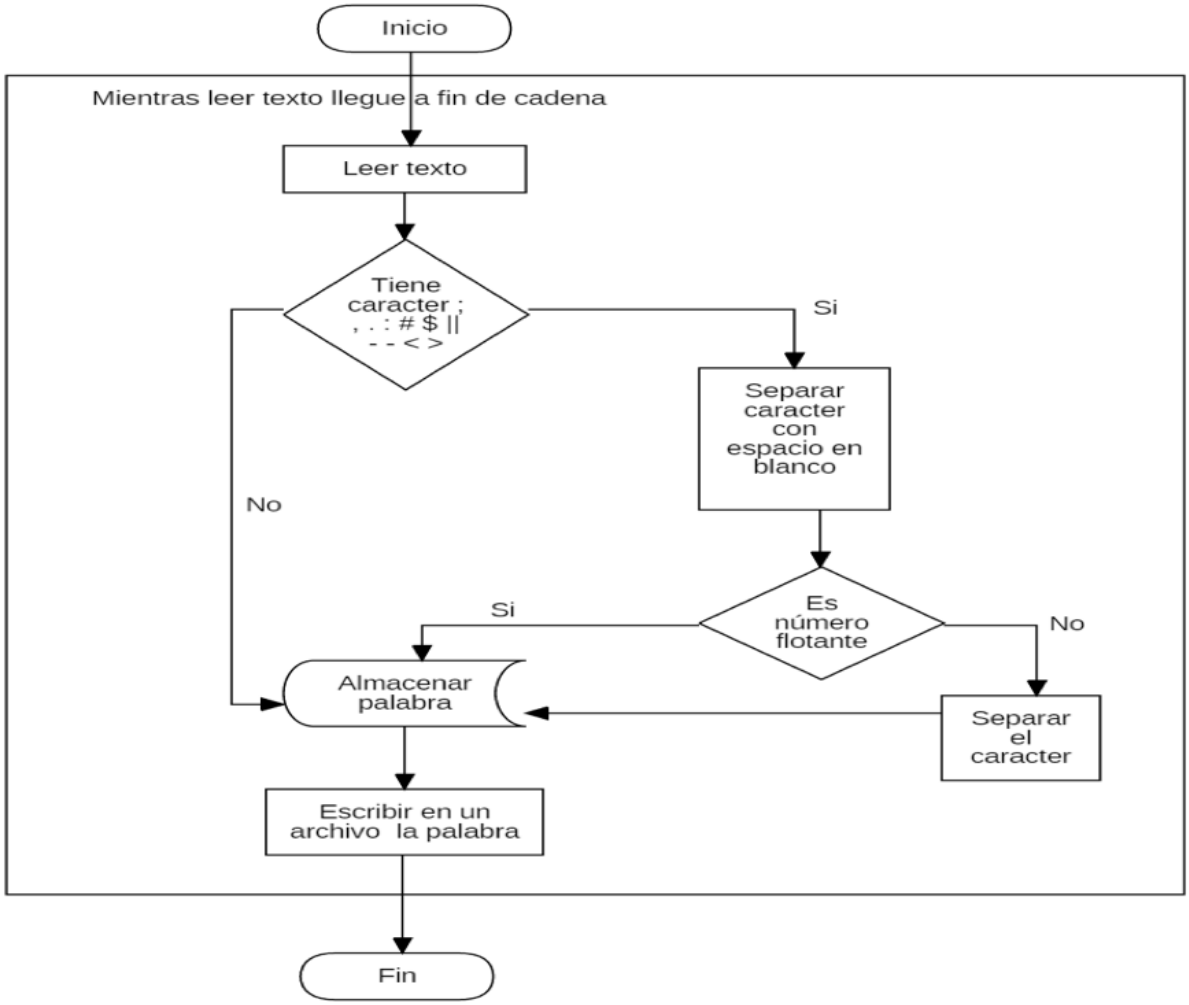

Figura 15: Proceso de tokenización

Proceso de lematización

El proceso de lematización es la segunda etapa del preprocesamiento y tiene como objetivo principal determinar el lema de cada La lematización implica pasar las palabras en su forma base. Para llevar a cabo este proceso se deben reducir las variantes morfológicas de una palabra a sus raíces comunes. Con la ayuda de la herramienta Freeling y su API de Python se obtienen los lemas del texto procesado del resultado del proceso de tokenización.

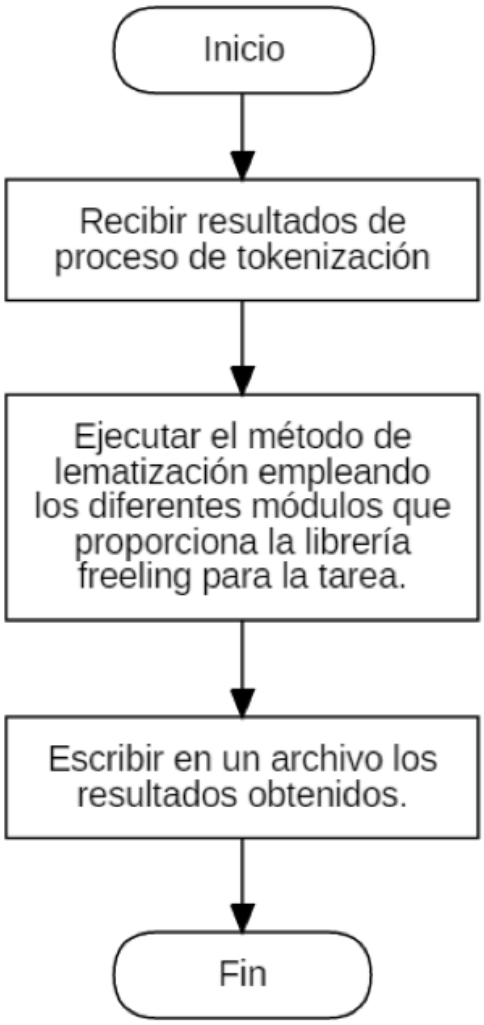

Figura 16: Proceso de lematización

El programa recibe una línea de texto que contiene el resultado de la Tokenización. En ese momento se ejecuta la función de lematización que emplea los distintos módulos que provee la librería Freeling para llevar los lemas de la manera que mejor nos acomode para realizar la clasificación.

En la siguiente se muestra la salida generada del proceso de lematización que recibió como entrada el resultado de la noticia tokenizada mostrada en la Figura 14.

titular : **presentar** el prototipo de el 1 moto volador ruso

cuando radio futuro en su canción enamorar de el moda juvenil **decir** aquel de " el futuro ya estar aquí " justo antes de el estribillo , tener todo el razón . el futuro ser hoy , tal y como ir el velocidad de desarrollo y el nuevo tecnología . hacer poco año , pensar en uno moto volador ser más bien algo de ciencia ficción . hoy , ya no el ser . este ser el prototipo de el que se pretender ser el 1 moto volador ruso . este modelo se llamar hoverbike S3 y el haber construir el empresa hover surf . este vídeo ser grabado en el circuito de carrera de moscú por alexandr atamanov , el director general de el empresa , en el que se ver cómo el aparato se elevar a uno metro de el suelo y comenzar a volar . el mayor banco de rusia , el veb , ya haber invertir varios millón de dólar para el desarrollo de el mismo . " el dispositivo se parecer a uno motocicleta , pero de hacer ser más bien uno prototipo de plataforma en el que se poder construir uno multitud de vehículo . su futuro ser multifacético , existir el posibilidad de crear todo uno familia de medio de transporte " asegurar serguei gorkov , presidente de el banco . hover surf ya haber recibir mucho prereserva de este modelo de asia y de oriente medio , a pesar de que su precio ser aun muy elevar , entre el 50.0 y 80.0 dólar , depender de su configuración .

Figura 17: Noticia lematizada

5.1.4. Recolección de información

La recolección de noticias es un proceso delicado y muy importante en este trabajo. Para obtenerlas hay que considerar varias cosas como por ejemplo el medio por el cual se recolectarán y la cantidad de noticias a recolectar.
Cada uno de los diarios seleccionados cuenta con un conjunto de secciones que lo conforman, algunas veces entre diarios comparten el mismo nombre de sección, pero en otras unicamente se comparten caracteristicas pero el nombre es distinto (por ejemplo, los diarios El Universal y La jornada cuentan con una sección llamada mundo, en cambio el diario Excelsior no cuenta con ella).
En la tabla se muestra el número de noticias recolectadas desde Septiembre 2017 a Marzo 2018.

	Noticias	
	Sección	Cantidad de noticias
El Universal	Cartera	97
	Ciencias	73
	Cultura	42
	De última	44
	Estados	151
	Menú	47
	Metrópoli	196
	Mundo	132
	Nacional	305
	Techbit	62
	Universal Deportes	107
	Espectáculos	223
Total		**1479**
La Jornada	Capital	134
	Ciencias	41
	Cultura	75
	Deportes	110
	Economía	144
	Espectáculos	110
	Estados	158
	Mundo	132
	Opinión	50
	Política	438
	Sociedad	76
Total		**1464**
Excélsior	Comunidad	218
	De la red	125
	Función	117
	Global	160
	Hacker	72
	Nacional	392
Total		**1084**
Total de noticias		**4027**

Tabla 10: Conteo de noticias

Como se muestra en la tabla anterior, cada diario cuenta con una cantidad de noticias y secciones diferentes. Cabe destacar que durante el periodo de enero a marzo de 2018 se hizo un balance en la cantidad de noticias de todos los diarios, es decir, se aumentó el número de noticias para cada una de las secciones con poca cantidad de las mismas, a pesar de ello, se sigue notando un claro desbalanceo entre secciones, esto se debe a que el mismo diario no publica la misma cantidad de noticias de una sección que de otra y eso se refleja en nuestro corpus.

5.2. Extracción de características

Una vez generado el corpus, es necesario transformar toda esa información a una representación vectorial para que computacionalmente pueda ser procesada. Por ello, se seleccionaron 3 técnicas diferentes las cuales formaron parte de la experimentación, con el objetivo de escoger aquella que arroje mejores resultados a la hora de clasificar. Dichas técnicas fueron:

- **Frecuencia**: Es un vector que representa la frecuencia de aparición de un término en un documento.
- **Binarización**: Es un vector que representa si un término se encuentra o no en un documento.
- **TF-IDF**: Es un vector que representa el peso de cada término en un documento.

5.3. Aprendizaje autómatico

Existen varias técnicas de clasificación probabilística que se han utilizado en la clasificación de textos. Los clasificadores seleccionados para la fase de experimentación fueron los siguientes:

- Árboles de decisión
- Máquinas de soporte vectorial
- Naive Bayes Multinomial
- Regresión logística

Esto debido a que son los clasificadores que aparecen con más frecuencia en trabajos de aprendizaje automático supervisado como el que se está presentando y han obtenido los mejores resultados.

5.4. Pruebas y resultados

Una vez dada por concluida la tarea de recolección de noticias, se procedió establecer ciertas medidas a la ejecución de varias pruebas, con el objetivo de comparar eficiencia de los clasificadores establecidos, analizar las diferentes técnicas de representación de datos y extracción de características para poder así obtener una combinación de las mismas, logrando así que nuestro clasificador de noticias arroje la mayor exactitud posible.
Como se está utilizando un aprendizaje supervisado, se debe seleccionar del corpus que porcentaje será utilizado como conjunto de entrenamiento (noticias en las que se basará el clasificador para determinar a qué sección pertenece) y que como conjunto de prueba (noticias que serán clasificadas como prueba para el clasificador). Se determinó que un 90 % del corpus sea utilizado como conjunto de entrenamiento y 10 % de prueba ya que la mayoría de los proyectos de clasificación supervisada utilizan dicho porcentaje.

Todas las pruebas cuentan con preprocesamiento de información (tokenizado y lematizado), la métrica utilizada para comparar los clasificadores es la exactitud.

5.4.1. Representación vectorial

En esta prueba, se buscó aquella representación vectorial que mejor resultado brinde para el objetivo buscado, para ello se comparó la exactitud arrojada de cada clasificador de acuerdo a cada representación utilizada. Los resultados fueron los siguientes:

Árboles de decisión

	Árboles de decisión		
	Frecuencia	Binarización	TF-IDF
El Universal	66.12	63.70	60.48
La Jornada	56.83	56.11	50.00
Excélsior	53.84	53.77	46.23

Tabla 11: Resultados clasificación árboles de decisión por representacion vectorial

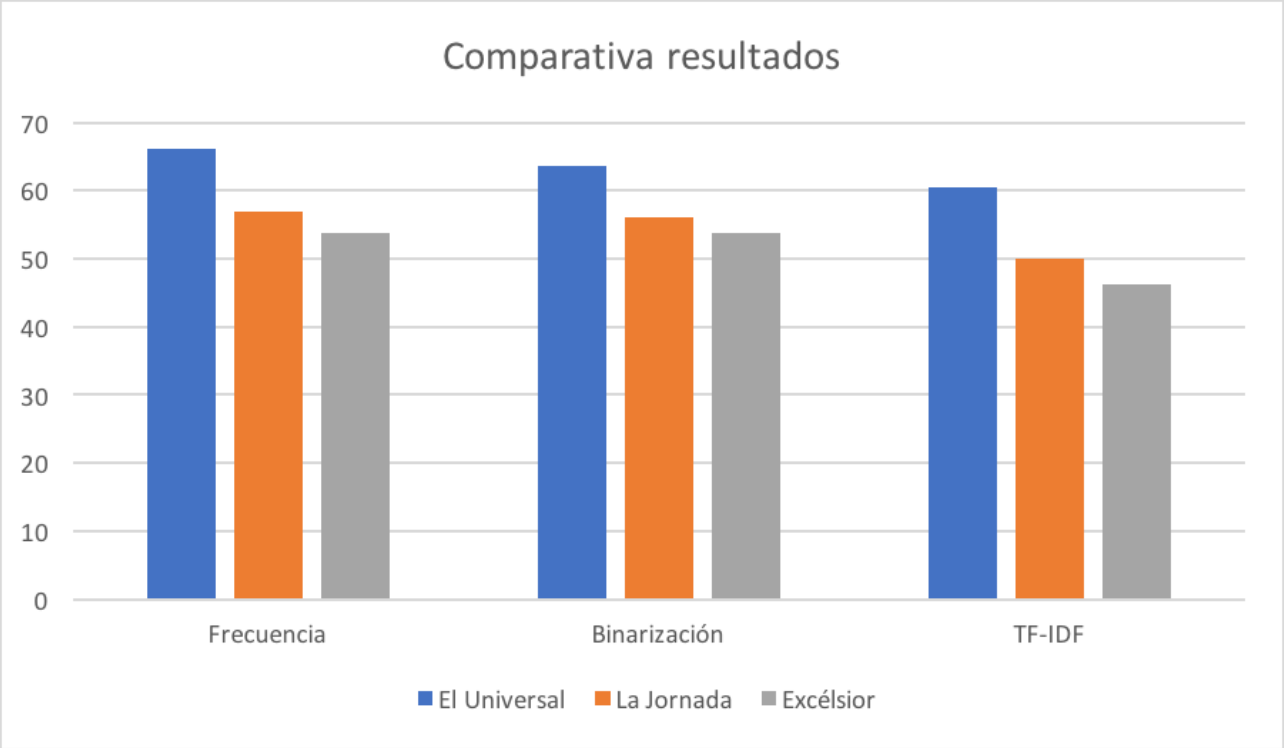

Figura 18: Comparativa de resultados árboles de decisión

Como se puede observar en la gráfica anterior, la exactitud más alta independientemente del diario es cuando se usó una representación de frecuencias. *El Universal* fue el diario que mayor porcentaje de exactitud obtuvo con un 66.12 %, en tanto para los otros diarios no pasaba del 57 % la exactitud en ningún tipo de representación.

Naive Bayes Multinomial

	Naive Bayes Multinomial		
	Frecuencia	Binarización	TF-IDF
El Universal	63.44	63.44	32.41
La Jornada	48.97	61.22	32.65
Excélsior	57.79	63.44	41.28

Tabla 12: Resultados clasificación Naive Bayes por representacion vectorial

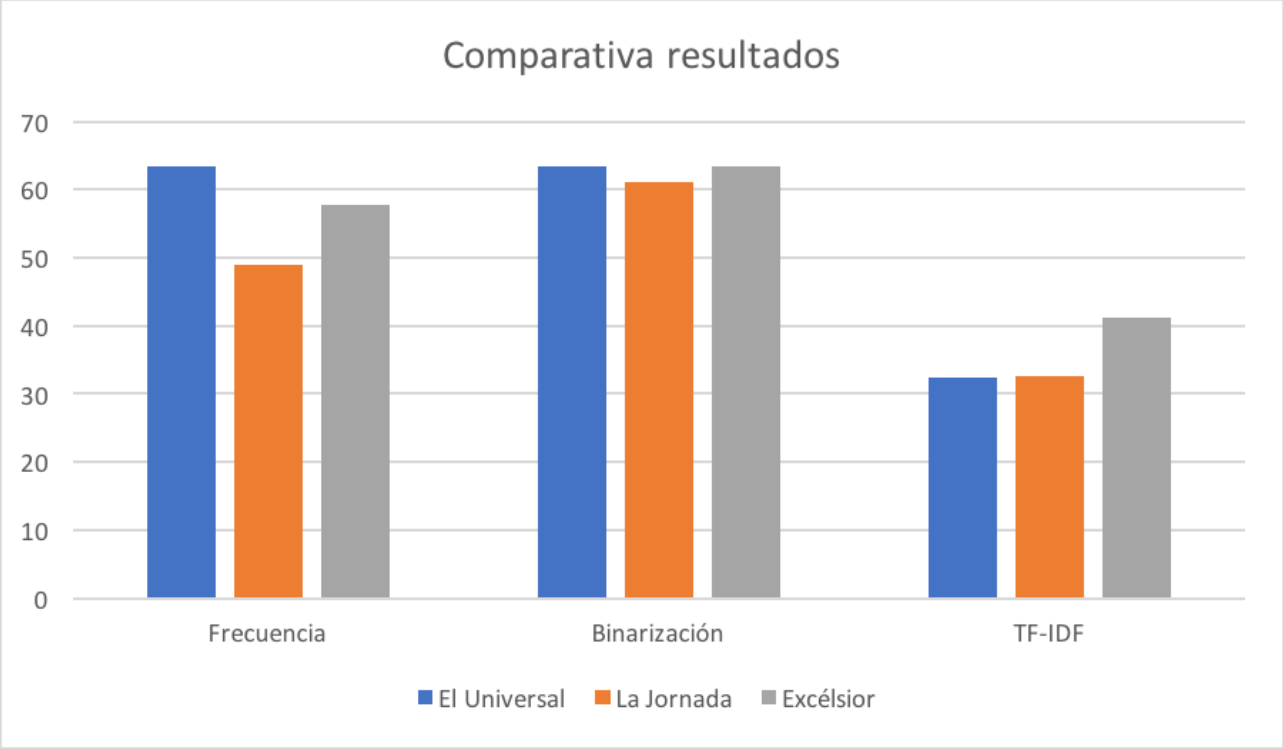

Figura 19: Comparativa de resultados naive bayes multinomial

En este experimento se puso observar que los mejores resultados fueron arrojados por la representación de binarización, ya que la exactitud en cada diario utilizando dicha representación fue mayor a 60 %, pero no supero ni el 70 % siendo así una exactitud mejor que los arboles de decisión, pero con un porcentaje bajo.

Máquina de soporte vectorial lineal

	SVM Lineal		
	Frecuencia	Binarización	TF-IDF
El Universal	81.39	82.75	85.51
La Jornada	80.27	74.82	77.55
Excélsior	77.98	78.89	78.89

Tabla 13: Resultados clasificación SVM Lineal

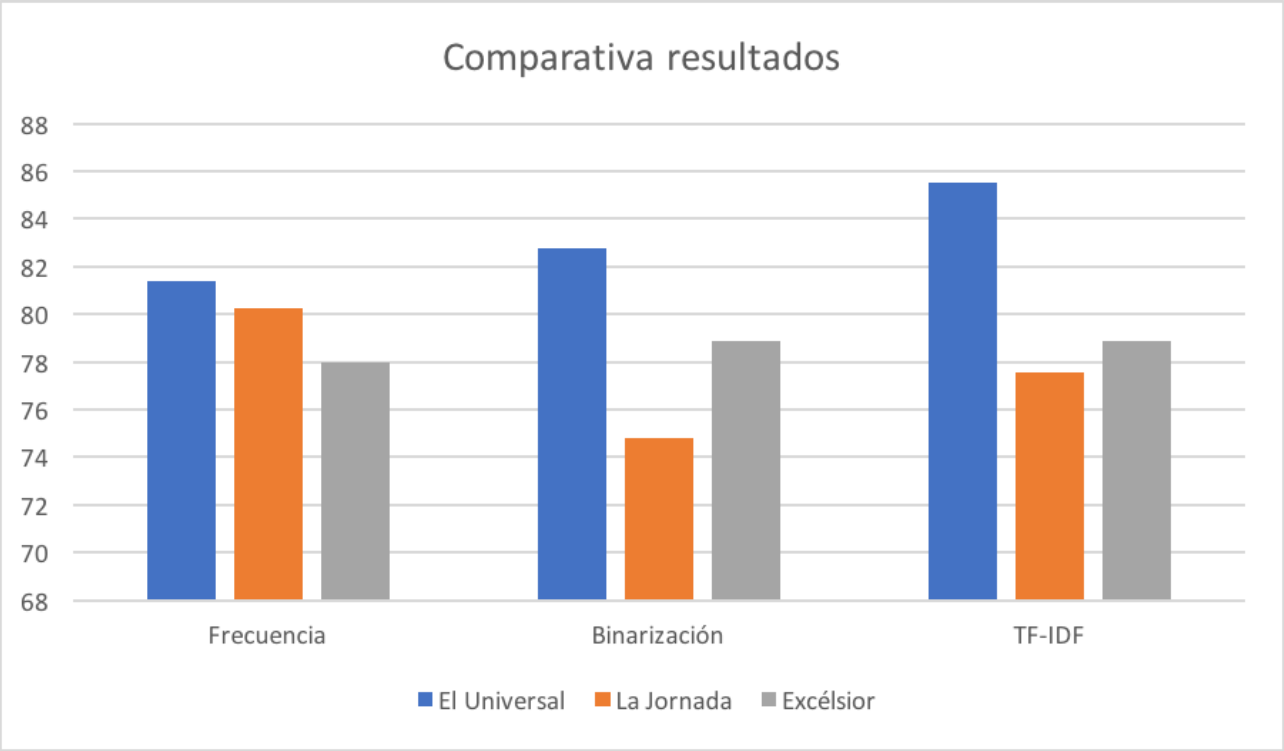

Figura 20: Comparativa de resultados naive bayes multinomial

La máquina de soporte vectorial lineal cuenta con parámetros diferentes a los otros clasificadores, como lo son parámetros de penalización de error (c), se debe de seleccionar un algoritmo de optimización, dicho algoritmo varía dependiendo de si el número de muestras supera al número de características y para este caso, se decidió usar un algoritmo dual, en tanto se utilizó una semilla = 1 la cual se encarga de generar números aleatorios para mezclar los datos, al ser uno siempre se generará la misma mezcla de números aleatorios.
Como se puede observar en los resultados, se llego hasta un 85.51 % de exactitud para el diario el universal utilizando TF-IDF, siendo este resultado uno de los más altos en la primera fase de experimentación, siguiendo la misma representación, los diarios restantes obtuvieron una exactitud mayor a 77 %.

Máquina de soporte vectorial polinomial

	SVM Polinomial		
	Frecuencia	Binarización	TF-IDF
El Universal	73.79	82.06	28.96
La Jornada	76.87	74.14	74.82
Excélsior	75.22	82.62	39.44

Tabla 14: Resultados clasificación SVM polinomial

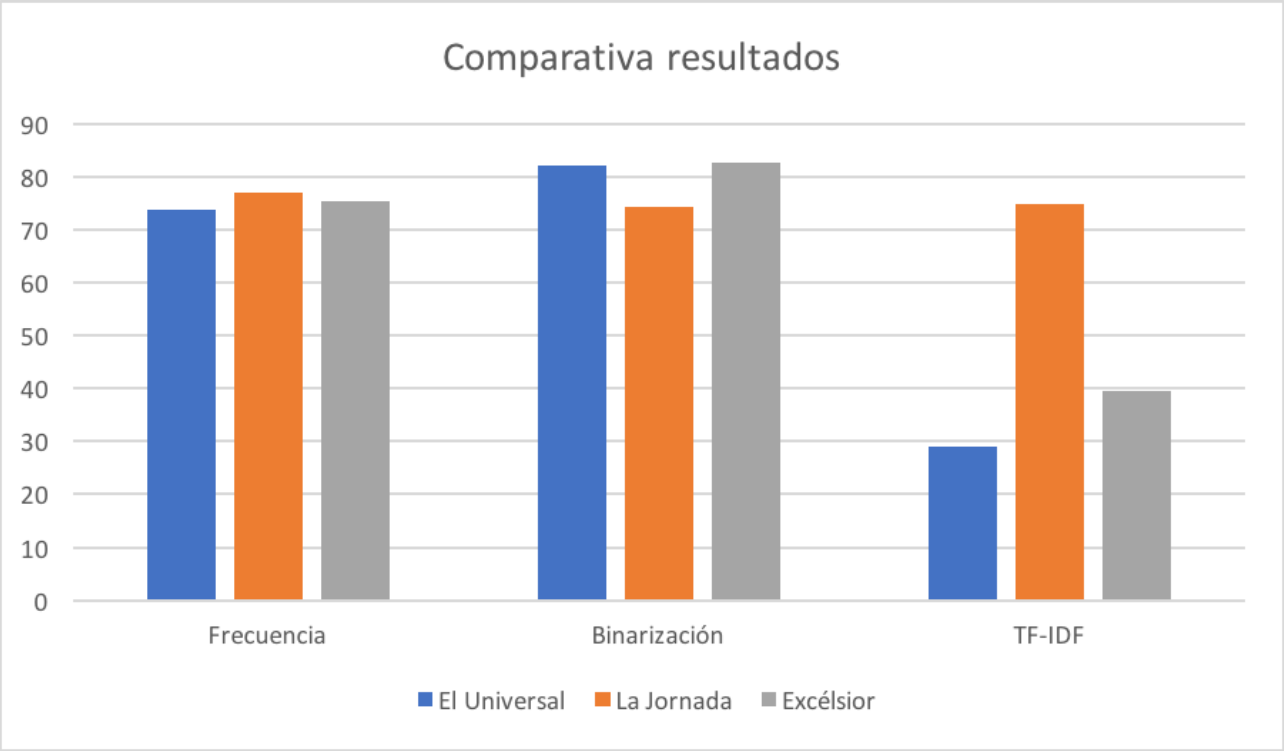

Figura 21: Comparativa de resultados SVM polinomiall

La máquina de soporte polinomial cuenta con diferentes parametros que las máquinas lineales no cuentan, se utilizo un kernel polinomial con un degree de 2, c 1500 , gama 1 x 10 -3 , coeficiente = 10000.
En este caso, la representación tf-idf arrojó un porcentaje muy bajo para los 3 diarios, la binarización arrojo un resultado de 82.62 %, un porcentaje alto. La máquina de soporte vectorial polinomial siendo comparada con la máquina lineal tiene resultados más bajos.

Regresión logística

	Regresión Logística		
	Frecuencia	Binarización	TF-IDF
El Universal	77.51	84.13	85.51
La Jornada	76.35	76.19	78.23
Excélsior	77.98	84.13	77.06

Tabla 15: Resultados clasificación Regresión Logística por representacion vectorial

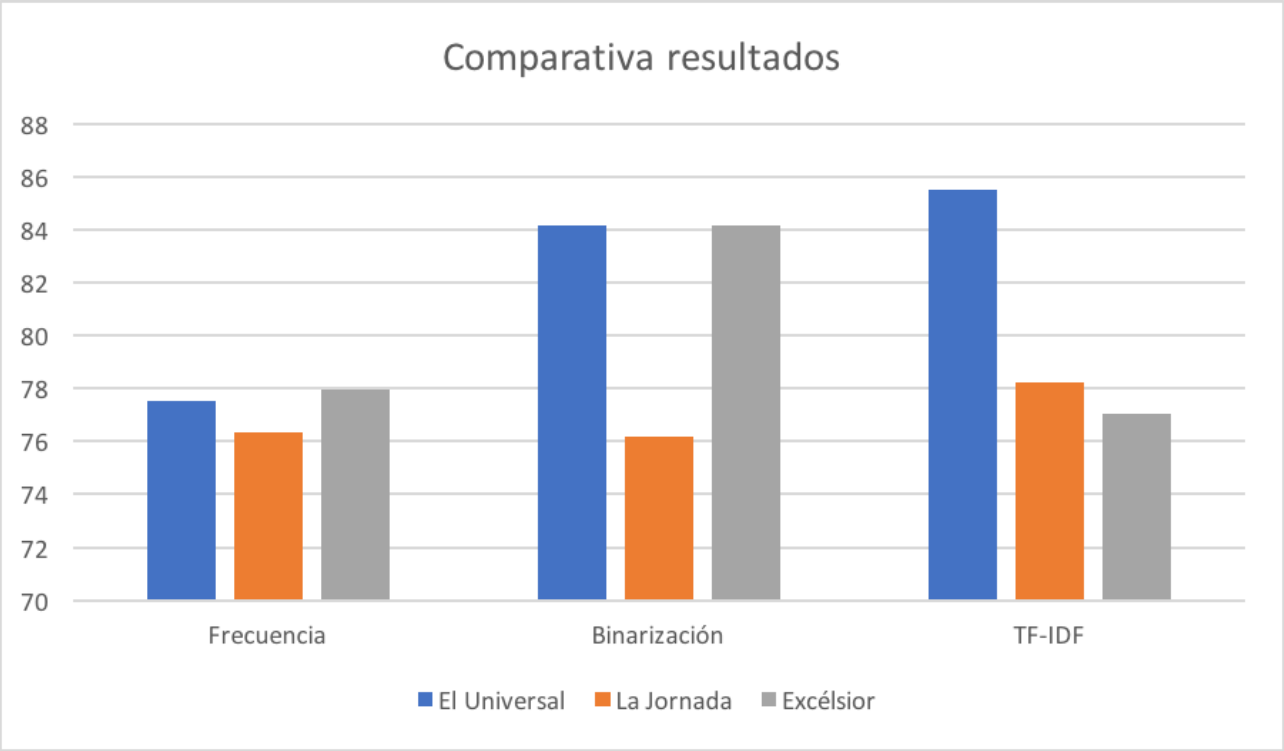

Figura 22: Comparativa de resultados Regresión Logística

Los resultados anteriores detallan que utilizando un TFf-IDF y Regresión logística se tiene un 85.51 % de exactitud para el diario El universal, aquí se dieron los resultados más elevados ya que se tiene un 76.19 % como mínimo de exactitud para el diario La Jornada utilizando binarización.

Conclusión

En los experimentos realizados, se observó que la representación vectorial que arroja mayor grado de exactitud es la representación de TF-IDF, donde la exactitud ha sido más alta y constante independientemente del diario seleccionado.
Hablando de los clasificadores, se observa que los aquellos con un mayor rendimiento son:

- Máquina de Soporte Vectorial: Utilizando kernel lineal arrojó una exactitud mucho más alta a comparación del kernel polinomial, llegando hasta un 85.51 % porciento de exactitud.
- Regresión Logística: la exactitud obtenida en dicho clasificador es de las más altas con respecto a otros clasificadores, arrojando una exactitud alta al igual que la máquina de soporte vectorial lineal.

Por lo tanto, se ha decidido seguir con más pruebas a dichos clasificadores, y en cuanto a la representación vectorial, se seguirá trabajando con TF-IDF que fue la que mayor porcentaje de exactitud brindó.

5.4.2. Stop words

El siguiente experimento fue eliminar del corpus una lista de palabras que pueden o no ayudar a incrementar la exactitud a la hora de clasificar, palabras vacías, irrelevantes también llamadas stop words son aquellas que por su altísima frecuencia de aparición en los textos.
Para ello, se realizaron dos pruebas, una eliminando las stop words del corpus y una manteniéndolas durante la clasificación, los resultados fueron los siguientes:

Manteniendo stop words

	Manteniendo stop words	
	Regresión Logística	Máquina de Soporte Vectorial
El Universal	85.51	85.51
La Jornada	78.23	77.55
Excélsior	77.06	78.89

Tabla 16: Resultados clasificación considerando stop words

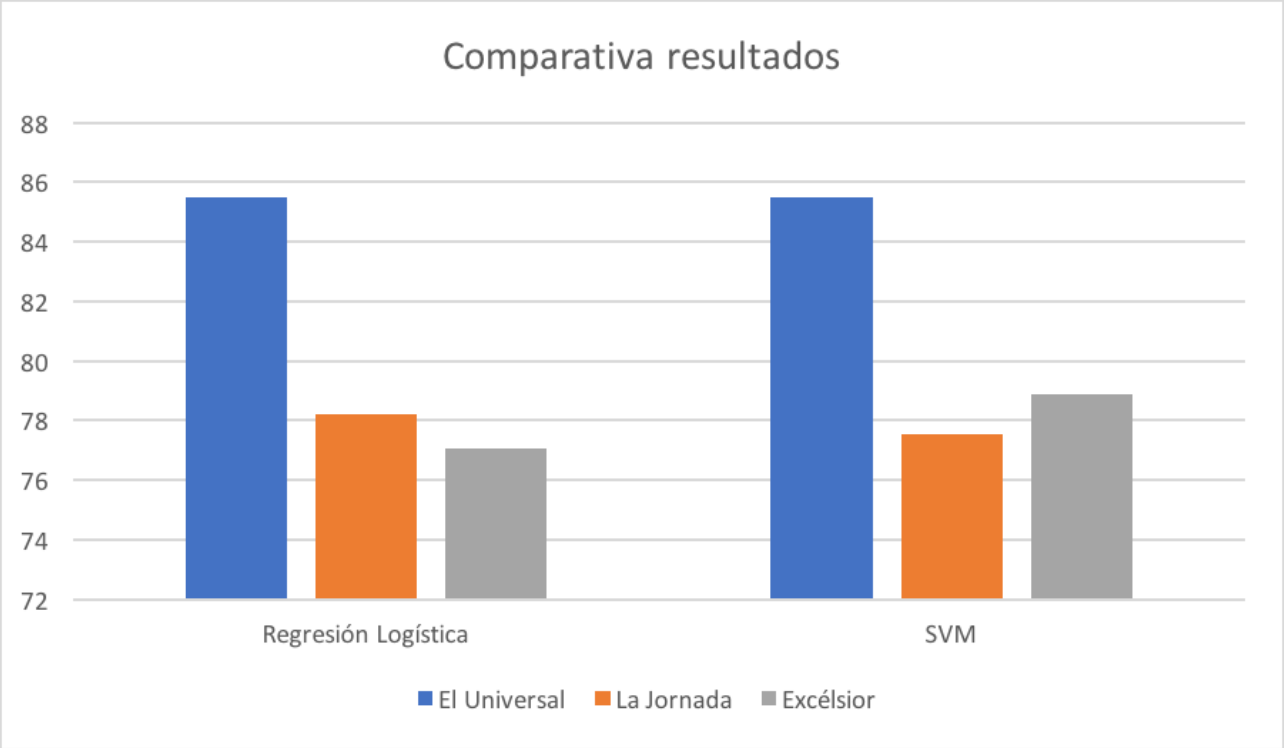

Figura 23: Comparativa de resultados considerando stop words

Como se muestra en la gráfica anterior, los resultados llegan hasta un 85.51% para el caso de las máquinas de soporte vectorial y Regresión logística, siendo muy similar en los resultados obtenidos, se cuenta con un mínimo de 77.06% que arroja la Regresión logística con el diario el Excélsior, la máquina de soporte vectorial tuvo una pequeña ventaja en cuanto a los resultados obtenidos.

Eliminando stop words

	Eliminando stop words	
	Regresión Logística	Máquina de Soporte Vectorial
El Universal	86.20	84.82
La Jornada	78.23	78.91
Excélsior	80.73	79.81

Tabla 17: Resultados clasificación eliminando stop words

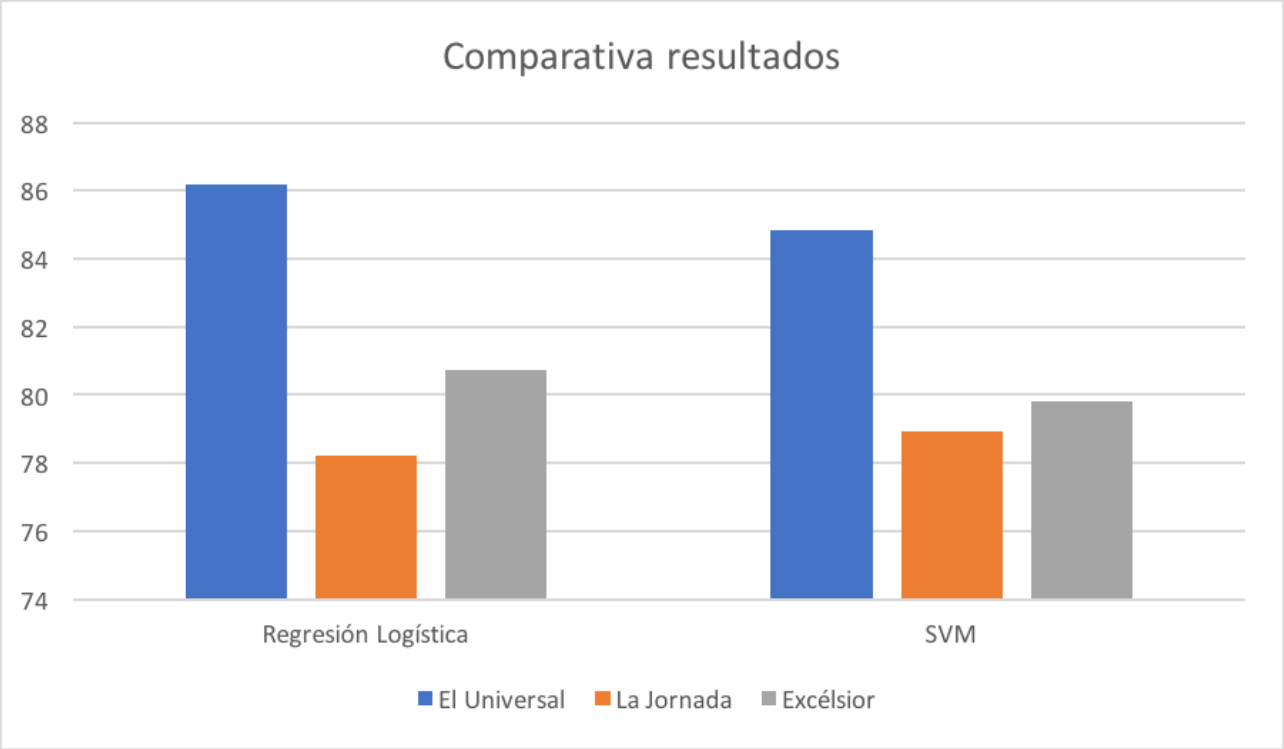

Figura 24: Comparativa de resultados eliminando stop words

En la tabla anterior, se puede observar que se vio levemente favorecida la exactitud para ambos clasificadores, teniendo un nuevo máximo de 86.20 % para el caso de regresión logística en el diario *El Universal*, el resultado mínimo para ambos diarios sobrepasa el 78 %, siendo así un resultado muy bueno a la hora de clasificar.

Conclusión

Se concluye con el experimento anterior, que en ciertas ocasiones el eliminar las stop words (palabras irrelevantes que no aportan nada al texto) puede ocasionar un aumento a la exactitud buscada, ya que, dichas palabras generan ruido a la hora de clasificar ocasionando un leve decremento en la exactitud. Se remarca la importancia de tener un corpus similar en tamaño a Wikipedia u otras fuentes con gran contenido de información, ya que, al tener un corpus reducido si se reduce aún más y teniendo una cantidad de secciones grande, puede afectar incluso de manera negativa a la hora de clasificar. Por lo antes mencionado, se determinó eliminar las stop words para así obtener una mayor exactitud.

5.4.3. Selección de características

En este experimento se realizó la prueba utilizando técnicas de selección de características (proceso de detectar las características relevantes de una muestra de datos), para poder así eliminar palabras irrelevantes que puedan afectar el proceso de clasificación, se estableció quedarse con el 80 % de características más importantes del corpus, en este experimento no se han eliminado las stop words. Los resultados se muestran a continuación:

Selección de características

	Selección de características	
	Regresión Logística	Máquina de Soporte Vectorial
El Universal	84.13	84.82
La Jornada	76.87	75.51
Excélsior	76.14	78.89

Tabla 18: Resultados clasificación empleando selección de características

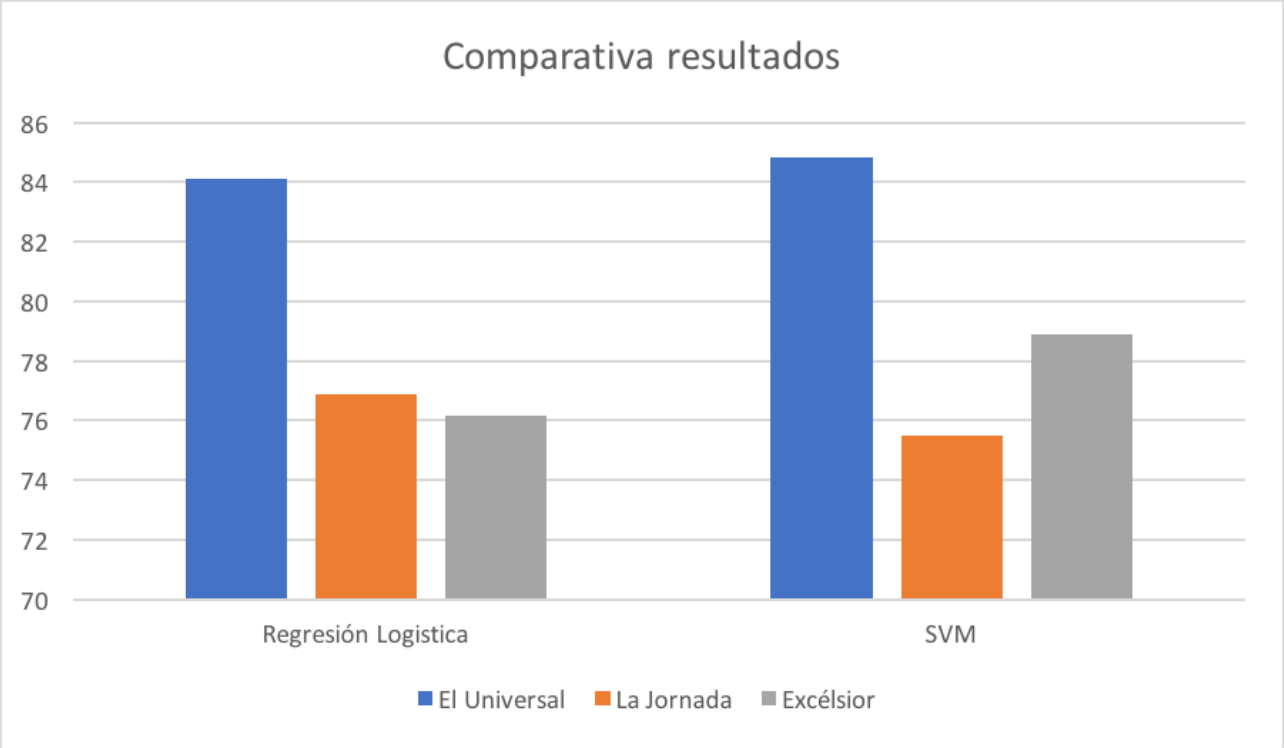

Figura 25: Comparativa de resultados con selección de características

Como se puede observar en la tabla anterior, la clasificación empleando selección de características fue más baja que eliminando stop words, se obtuvo un mínimo más bajo que el anterior ahora con un 76.14 % para el diario Excélsior, debido a que dicho diario cuenta con menos noticias que los demás ocasiona que cuente con menos características, y si se le disminuyen aún más con el proceso de selección de las mismas, la exactitud baja considerablemente.

Selección de características y stop words

Para comprobar lo mencionado en el experimento anterior, se realizó la clasificación empleando selección de características, pero eliminando las stop words, con esto se reduciría aún más el vector de representación vectorial.

	Selección de características y stop words	
	Regresión Logística	Máquina de soporte vectorial
El Universal	84.82	84.82
La Jornada	78.23	77.55
Excélsior	76.14	77.98

Tabla 19: Resultados clasificación empleando selección de características y eliminando stop words

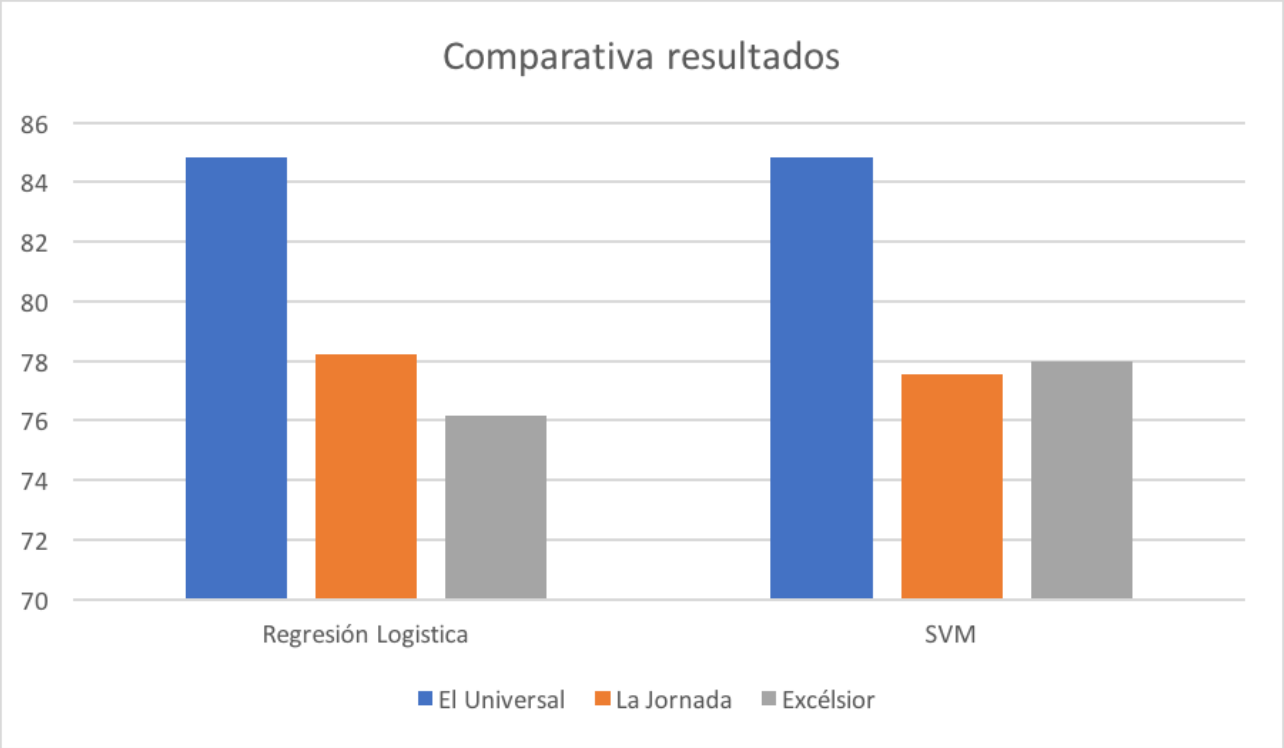

Figura 26: Comparativa de resultados con selección de características y eliminando stop words

Al combinar la selección de características y eliminando las stop words reducen la cantidad de características obtenidas del corpus, y a pesar de ello, aumento un poco la exactitud en cada resultado obtenido independientemente del diario, algunos resultados son muy similares a los obtenidos uncamente empleando selección de características, algunos otros si aumentaron una pequeña cantidad, pero de igual manera, no lograron ser mayores a los que solamente se eliminan las stop words.

Conclusión

El eliminar más características empleando diferentes técnicas como la selección de características, en algunas ocasiones no se obtiene una mejoría en la exactitud, en este trabajo terminal, se tiene una cantidad de características menor a las que un corpus del tamaño de wikipedia podría llegar a obtener, es decir, si se eliminan más características a un corpus con poca cantidad de noticias se estaría perjudicando de manera negativa la exactitud, ya que al contar con menos características,

es más dificíl poder clasificar adecuadamente. Por ende, se nota claramente que si aplicamos este método, se reduce la efectividad a comparación de como se tenia eliminando unícamente las stop words, por lo que se decidió no utilizar selección de características y a la siguiente etapa pasa el clasificador máquina de soporte vectorial aplicando stop words.

5.4.4. Validación Cruzada

Para poder medir de mejor manera la exactitud del clasificador seleccionado, se utilizó la validación cruzada, técnica utilizada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y prueba.
Consiste en repetir y calcular la media aritmética obtenida de las medidas de evaluación sobre diferentes particiones, los datos de muestra se dividen en K subconjuntos. Uno de los subconjuntos se utiliza como datos de prueba y el resto (K-1) como datos de entrenamiento.
Se implementó validación cruzada al clasificador máquina de soporte vectorial con 10 subconjuntos diferentes.

Aplicando Stop words K = 10	
	Máquina de soporte vectorial
El Universal	79.81
La Jornada	77.22
Excélsior	77.94

Tabla 20: Validacion Cruzada SVM

Gracias a la implementación de la validación cruzada podemos observar una exactitud más acertada debido a lo mencionado, obteniendo en promedio resultados cercanos al 80 % de exactitud.

5.4.4.1 Matriz de confusión

La matriz de confusión, es una matriz de n x n donde n es el número de secciones del clasificador, la cual representa el número de predicciones correctas para cada una de las n secciones. De la matriz de confusión se obtuvieron tres indicadores importantes: Precisión, recall y f-measure.

5.4.4.2 Máquina de Soporte Vectorial El Universal aplicando stop words

	Matriz de confusión SVM El Universal implementando stop words											
	Cartera	Cultura	Espectáculos	Estados	Menú	Metrópoli	Mundo	Techbit	Ciencias	Deportes	Nacional	De Última
Cartera	**5.6**	0.0	0.0	0.1	0.1	0.5	0.0	0.6	0.1	0.2	1.1	0.1
Cultura	0.1	**2.3**	0.7	0.3	0.0	0.0	0.0	0.1	0.0	0.0	0.4	0.0
Espectáculos	0.0	0.1	**20.7**	0.0	0.0	0.3	0.0	0.0	0.0	0.1	0.2	0.3
Estados	0.2	0.1	0.4	**6.5**	0.2	0.3	0.0	0.0	0.0	0.1	3.7	0.0
Menú	0.0	0.1	0.0	0.1	**4.3**	0.0	0.1	0.0	0.1	0.2	0.0	0.0
Metrópoli	0.7	0.0	0.5	0.8	0.0	**16.1**	0.0	0.1	0.0	0.2	1.3	0.0
Mundo	0.1	0.0	0.8	0.4	0.0	0.2	**9.8**	0.1	0.2	0.0	0.8	0.0
Techbit	0.4	0.0	0.0	0.0	0.0	0.1	0.4	**4.9**	0.1	0.0	0.3	0.0
Ciencias	0.0	0.1	0.1	0.0	0.3	0.0	0.3	0.1	**5.9**	0.0	0.2	0.0
Deportes	0.0	0.1	0.1	0.2	0.0	0.0	0.2	0.0	0.0	**9.1**	0.1	0.2
Nacional	0.5	0.2	0.4	2.2	0.0	1.8	0.3	0.1	0.8	0.1	**24.7**	0.0
De Última	0.0	0.0	0.6	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	**3.7**

Tabla 21: Matriz de confusión SVM *El Universal* aplicando stop words

	Máquina Soporte Vectorial El Universal aplicando stop words			
	Precision	Recall	F1-Score	Support
Cartera	0.73	0.66	0.69	12
Cultura	0.76	0.58	0.65	1
Espectáculos	0.84	0.95	0.89	28
Estados	0.61	0.56	0.58	9
Menú	0.87	0.87	0.87	6
Metrópoli	0.83	0.81	0.81	23
Mundo	0.88	0.79	0.83	11
Techbit	0.79	0.79	0.79	3
Ciencias	0.81	0.84	0.82	4
Deportes	0.91	0.90	0.90	14
Nacional	0.75	0.79	0.76	32
De Última	0.86	0.82	0.83	2

Tabla 22: Resultados precisión, recall y f-measure Máquina Soporte Vectorial *El Universal* aplicando stop words

Como se puede observar en la tabla anterior, existen secciones con un 91 % de precisión, es decir, del total de noticias que fueron clasificadas a dicha seccion 91 % de ellas son correctas, en tanto la sección de Estados es la que menor precisión y recall obtuvó, esto debido al desbalance de noticias, ocasionando que un 49 % de noticias que fueron clasificadas como Estados sean incorrectas, al igual que recall, con un 56 % de noticias de una clase logra clasificar adecuadamente.

5.4.4.3 Máquina de Soporte Vectorial La Jornada aplicando stop words

	Matriz de confusión SVM La Jornada aplicando stop words										
	Cultura	Deportes	Economía	Espectáculos	Estados	Mundo	Política	Capital	Ciencias	Opinión	Sociedad
Cultura	**6.1**	0.0	0.1	0.5	0.1	0.2	0.3	0.2	0.2	0.1	0.1
Deportes	0.0	**10.7**	0.0	0.0	0.0	0.2	0.5	0.0	0.0	0.1	0.0
Economía	0.0	0.0	**11.1**	0.0	0.3	0.2	2.2	0.2	0.0	0.0	0.2
Espectáculos	0.4	0.0	0.0	**10.4**	0.1	0.6	0.0	0.0	0.0	0.0	0.1
Estados	0.0	0.0	0.2	0.0	**9.1**	0.1	5.1	0.1	0.1	0.0	0.4
Mundo	0.0	0.0	0.2	0.2	0.5	**10.1**	1.3	0.0	0.0	0.0	0.2
Política	0.1	0.1	1.1	0.0	2.1	0.6	**34.8**	1.0	0.1	0.2	1.3
Capital	0.1	0.0	0.1	0.0	0.3	0.0	1.5	**10.4**	0.0	0.0	0.1
Ciencias	0.0	0.0	0.0	0.1	0.1	0.1	0.0	0.0	**3.6**	0.0	0.6
Opinión	0.3	0.1	0.0	0.0	0.1	0.7	1.0	0.1	0.0	**1.1**	0.3
Sociedad	0.1	0.1	0.2	0.1	0.2	0.3	3.0	0.2	0.3	0.2	**2.6**

Tabla 23: Matriz de confusión SVM *La Jornada* aplicando stop words

	Máquina Soporte Vectorial La Jornada aplicando stop words			
	Precision	Recall	F1-Score	Support
Cultura	0.81	0.77	0.78	6
Deportes	0.97	0.93	0.94	11
Economía	0.85	0.78	0.81	16
Espectáculos	0.91	0.89	0.89	6
Estados	0.70	0.59	0.64	21
Mundo	0.77	0.80	0.78	11
Política	0.70	0.84	0.76	48
Capital	0.85	0.85	0.83	13
Ciencias	0.81	0.80	0.80	3
Opinion	0.64	0.29	0.39	6
Sociedad	0.44	0.35	0.38	6

Tabla 24: Resultados precisión, recall y f-measure Máquina Soporte Vectorial *La Jornada* aplicando stop words

Para el caso de La Jornada, observamos que la precisión llegó hasta un máximo de 97 % para Deportes, en el caso de recall llego hasta un 93 % en dicha sección. En tanto, la sección con una precisión y recall más bajo es sociedad con 0.44 % y recall de 35 %.

5.4.4.4 Máquina de Soporte Vectorial Excelsior aplicando stop words

	Matriz de confusión SVM Excélsior aplicando stop words					
	Comunidad	De la red	Función	Global	Hacker	Nacional
Comunidad	**18.5**	0.6	0.0	0.1	0.1	2.5
De la red	0.4	**5.5**	0.9	3.4	0.7	0.2
Función	0.2	0.3	**9.9**	0.2	0.2	0.5
Global	0.3	2.0	0.0	**10.9**	0.3	2.5
Hacker	0.2	0.9	0.8	0.5	**3.9**	0.9
Nacional	1.0	1.3	0.1	0.4	0.1	**35.9**

Tabla 25: Matriz de confusión SVM *Excélsior* SW

	Máquina Soporte Vectorial Excélsior SW			
	Precision	Recall	F1-Score	Support
Comunidad	0.93	0.84	0.88	22
De la red	0.51	0.42	0.46	18
Funcion	0.84	0.87	0.85	8
Global	0.70	0.68	0.68	13
Hacker	0.73	0.54	0.62	5
Nacional	0.75	0.92	0.82	43

Tabla 26: Resultados precisión, recall y f-measure Máquina Soporte Vectorial *Excélsior* aplicando stop words

Para este diario en particular, la precisión no arrojó altibajos muy notorios, es decir, la precisión más baja fue de 51 % para el caso de la sección De La Red, siendo este valor más alto que en otros diarios, de igual manera para el 42 % arrojado por dicha sección para el recall.

Conclusión

Podemos concluir que, al observar el desempeño con base en la matriz de confusión, cada diario cuenta con una sección que arroja un mayor y un menor porcentaje de precisión y recall, notándose de manera inmediata, uno de los factores por lo cuales dichos valores son pequeños se debe al desbalance de noticias, ya que debido a la falta de características de ciertas secciones es posible que otra sección este absorbiendo a otra. También la similitud entre clases es un factor importante, ya que puede confundir a la hora de clasificar.
El clasificador seleccionado fue evaluado con una calificación positiva con respecto a los indicadores mostrados, por ende se utilizará para la creación de nuestro modelo.

5.4.5. Aplicación del modelo

Para poder aplicar el modelo elegido al software, se tienen que generar dos modelos. El primero es el clasificador entrenado con el 100 % de datos, que es el que permitirá predecir nuevas noticias. El segundo es el modelo una matriz de dispersión,que contiene el número total de características utilizadas en la fase de entrenamiento, permite hacer la transformación de las noticias a clasificar de manera que tengan la misma longitud que la matriz original.
Una vez generados los modelos se pueden exportar a la aplicación en un formato .pkl, esto para no repetir el proceso de entreamiento que es tardado.

5.4.6. Conclusiones

A partir de los 144 experimentos aproximadamente realizados en este trabajo terminal en este trabajo terminal y con base en los resultados obtenidos, se puede concluir que el preprocesamiento de información, junto al uso de diferentes representaciones vectoriales, técnicas de extracción de características y clasificador utilizado brindan un aumento o disminución en la exactitud, mucho de ello depende del corpus generado (el cual para este caso, fue recolectado desde cero debido a que públicamente no existen corpus en el idioma español de noticias de diarios de circulación nacional) debido a que los diarios de circulación nacional no muestran la misma cantidad de noticias en cada sección el corpus reflejaba un desbalance en la cantidad de noticias por sección y por ende un desbalance en la cantidad de características de cada una, de igual manera depende de las características que cada representación, técnica de extracción y clasificador tengan. Por ello, el uso de un modelo el cual gracias a un preprocesamiento (tokenización y lematización) de información, usando una representación vectorial TF-IDF(que caracteriza el impacto que tienen las características en los documentos), aplicando Stop Words(palabras que por sí solas carecen de significado y que, por su altísima frecuencia de aparición en los textos, generan un ruido innecesario para la recuperación de información) y utilizando una máquina de soporte vectorial como clasificador se maximizo la exactitud en nuestro objetivo, llegando hasta un 79.81 % de exactitud a la hora de clasificar, se estima que 8 de cada 10 noticias son clasificadas correctamente a la sección que pertenece.

5.4.7. Trabajo futuro

Como trabajo futuro, se propone aumentar el número de noticias con el fin de que aquellas secciones que cuenten con menos noticias aumenten en características, con ello la exactitud tendría un aumento. Adicional a ello, se pretender elaborar un recolector de noticias de diarios de circulación nacional, el cual a través del mismo cualquier usuario que utilice dicho sistema pueda descargar noticias de la sección que el desee (se delimitaran las secciones) de varias fuentes de información(diarios) que el recolector pueda obtener información.

6. Anexos

6.1. Instalación de herramientas

A continuacion, se mostrara el proceso de instalacion de las herramientas utilizadas en este trabajo, las cuales son Freeling y scikit-learn.

6.1.1. Instalación de Freeling

Freeling es una herramienta que provee diversas maneras de instalación, la manera más versátil de instalarla es compilando la biblioteca, debido a que si es necesario hacer algún cambio o existen errores al momento de instalar tenemos la posibilidad de hacer las correcciones necesarias. Es posible realizar la instalación desde su repositorio de Git Hub, que es el que se presenta en este apartado.

Paso 1:
Freeling necesita una serie de herramientas orientadas a C++ ya que es nativa de este lenguaje. A continuación, se muestra el comando para la instalación de un compilador de C++, herramientas de GNU y GIT.

sudo apt-get install build-essential automake autoconf libtool git

Paso 2:
La librería Boost de C++ es una de las piezas fundamentales de la herramienta, la instalación de Boost se realiza con los siguientes comandos.

sudo apt-get install libboost-regex-dev libicu-dev zlib1g-dev

sudo apt-get install libboost-system-dev libboost-program-options-dev

Paso 3:
Para descargar la librería se hace un checkout al repositorio, para la versión más actual se usa una línea de git agregando el directorio donde se desea que se realice la compilación.

git clone https://github.com/TALP-UPC/FreeLing.git directorio

Paso 4:
Ahora se accede al directorio en el que se ha elegido la descarga, y se prepara el repositorio para su compilación.

cd directorio

autoreconf –install

Paso 5: Se configura e instala la librería.

./configure
make
sudo make install

Paso 6: Como la librería es de C++ es necesario compilar un API para que pueda ser usado con Python, es necesario acceder a la carpeta de instalación y ejecutar las líneas que se en listan a continuación.

Es necesario cambiar la ruta del directorio

FREELINGDIR = /usr/local/share

Cambiar la versión de Python con la cual se va a trabajar

PYTHONVER = python3.5

Compilar con el makefile

Make

Lo cual si todo ha sido correcto se generarán los archivos:

_freeling.sofreeling_pythonAPI.cxx

Para probar el que todo sea correcto es posible correr el programa de ejemplo que viene en la misma carpeta.

Python3 sample.py

6.1.2. Instalación scikit-learn

Es una biblioteca especializada para Python que permite el análisis de datos y tiene un enfoque hacia el Aprendizaje Automático. Para hacer la instalación de la librería es necesario contar con las herramientas numpy, scipy por lo que se añadirán a la instalación.

Paso 1:
En caso de no contar con el manejador de paquetes pip se puede obtener con:

apt-get install python3-pip
sudo apt-get update
python get-ip.py

Se puede comprobar la instalación con

Pip3 –version

Paso 2:
Una vez instalado pip se procede a instalar las herramientas, esto se logra con

pip install numpy, scipy, scikit-learn

Paso 2:

Se puede comprobar la instalación generando un pequeño programa donde se importen las librerías, si no genera errores al compilar y en la ejecución es porque se han instalado correctamente.

```
import sklearn
print('La versión de scikit learn es: '.format(sklearn.__version__))
```

6.2. Stop Words

A continuación, se muestran las stop words consideradas para este trabajo.

Stop words				
039	jpe	fbp	dre	*amgl
amgl	hch	x67	x61	x5f
x76	x30	x65	x6f	x63
x68	x74	x79	x40	x32
x6d	x6c	x73	x75	x62
protected	x2e	sc	pmba	lr
x69	tcm	x6a	msl	x64
x7a	ae	radián	nrv	ahc
afcl	cfe	lrs	efe	agv
maf	jpej	x78	irr	rcr
pst	jlcg	nbsp	lcg	kcp
x72	infogram	fb	afp	sjno
x6e	igc	rmlgv	x70	etp
.	,	;	:	”
’	(	)	{	}
[	]	!	¡	¿
?	*	+	/	#
$	%	&	=	—
hch	jci	sarr	*bb	*av
jrr	cva	a	el	los
la	lo	las	contra	ante
bajo	cabe	con	de	desde
en	entre	hacia	hasta	para
por	según	si	so	sobre
tras	y	o	u	e

Tabla 27: Lista de stop words

7. Bibliografia

Referencias

[1] R. Ávila Arguelles, *Clasificación bibliotecaria automática usando identificación simple de términos con métodos lógico-combinatorios a partir de información escasa*. PhD thesis, Centro de Investigación en Cómputo, México, 2008.

[2] A. Téllez-Valero, M. Montes, O. F.-C. Gómez, and L. Villaseñor-Pineda, "Clasificación automática de textos de desastres naturales en méxico,"

[3] D. B. Bracewell, J. Yan, F. Ren, and S. Kuroiwa, "Category classification and topic discovery of japanese and english news articles," *Electronic Notes in Theoretical Computer Science*, vol. 225, pp. 51–65, 2009.

[4] D. B. Bracewell, F. Ren, and S. Kuriowa, "Multilingual single document keyword extraction for information retrieval," in *Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE'05. Proceedings of 2005 IEEE International Conference on*, pp. 517–522, IEEE.

[5] D. Ramdass and S. Seshasai, "Document classification for newspaper articles," *Document classification for newspaper articles*, 2009.

[6] Microsoft, "¿qué es azure?," 2017.

[7] Microsoft, "Documentación de microsoft azure," 2017.

[8] I. B. M. España, "International business machines watson computación cognitiva."

[9] A. C. Vásquez, J. P. Quispe, A. M. Huayna, *et al.*, "Procesamiento de lenguaje natural," *Revista de investigación de Sistemas e Informática*, vol. 6, no. 2, pp. 45–54, 2009.

[10] B. O. Manuel, "Técnicas avanzadas de recuperación de información."

[11] L. R. Yunta, "La lematización en español: una aplicación para la recuperación de información (r. gómez díaz)," *Revista española de Documentación Científica*, vol. 29, no. 1, pp. 175–176, 2006.

[12] B. Luciana, "Lematización material complementario de la clase 5."

[13] F. S. Caparrini, "Introducción al aprendizaje automático."

[14] A. Moreno, "Aprendizaje automático," 1994.

[15] A. McCallum, K. Nigam, *et al.*, "A comparison of event models for naive bayes text classification," in *AAAI-98 workshop on learning for text categorization*, vol. 752, pp. 41–48, Citeseer, 1998.

[16] M. Raedo, M. Jesús, *et al.*, "Combinación de clasificadores: Construcción de características e incremento de la diversidad," 2010.

[17] E. J. C. Suárez, "Tutorial sobre máquinas de vectores soporte (svm)," *Tutorial sobre Máquinas de Vectores Soporte (SVM)*, 2014.

[18] C. Salcedo, "Estimación de la ocurrencia de incidencias en declaraciones de pólizas de importación," *Informe profesional. Universidad Nacional Mayor de San Marcos*, 2002.

[19] M. S. Velasco, "La regresión logística. una aplicación, a la demanda de estudios universitarios," *Estadística Española*, vol. 141, pp. 193–217, 1996.

[20] Á. F. Zazo, C. G-Figuerola, J.-L. Alonso-Berrocal, and R. Gómez-Díaz, "Recuperación de información utilizando el modelo vectorial. participación en el taller clef- 2001," 2002.

[21] V. Kumar and S. Minz, "Feature selection," *SmartCR*, vol. 4, no. 3, pp. 211–229, 2014.

[22] S. Visa, B. Ramsay, A. L. Ralescu, and E. Van Der Knaap, "Confusion matrix-based feature selection.," in *MAICS*, pp. 120–127, 2011.

[23] O. Rodríguez, "Validación cruzada (cross-validation) y remuestreo(bootstrapping)," 2015.

[24] M. Raedo, M. Jesús, *et al.*, "Los periodistas prefieren twitter, según un estudio de la uc3m.," 2010.