Prof. Claudia Celia Díaz Huerta

Clasificación de noticias de diarios de circulación nacional mediante aprendizaje automático

Trabajo Terminal No. 2017-A042

*Alumnos: García Molina José Alejandro, Ramírez Roque Luis Enrique, Sánchez Ramírez Miguel Ángel**

Directores: Juárez Gambino Joel Omar, Consuelo Varinia García Mendoza

e-mail: trabajottredes@gmail.com

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, además muchas de las noticias que aparecen publicadas en Internet no se encuentran clasificadas por secciones, lo cual dificulta su recuperación. 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 en la etapa de entrenamiento. Se probarán diferentes técnicas de extracción de características junto con diferentes algoritmos de aprendizaje automático, y al final se seleccionarán aquellos con la medida de exactitud mayor. Con las técnicas seleccionadas, se podrán clasificar nuevas noticias (etapa de prueba) en la secciones correspondientes de los diarios, de forma automática.

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

1. Introducción

Hoy en día existen muchas fuentes de información disponibles en Internet como blogs, redes sociales, foros, etc. Toda esta información requiere estar estructurada y etiquetada para que los motores de búsqueda como Google puedan acceder a ella y recuperarla. Por lo tanto la tarea de clasificar la información en tópicos es muy importante, ya que 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 qué tópicos abordan. Sin embargo, el costo de esta tarea ha motivado el desarrollo de procedimientos para la clasificación automatizada de texto. Estos algoritmos han permitido automatizar tareas como la de clasificación bibliotecaria [1].

Un caso particular de clasificación de texto es la que realiza el editor de un diario al definir en qué sección de este aparecerá una nota. 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 aparentemente simple para un humano implica un esfuerzo adicional por parte del diario para clasificar correctamente la información antes de publicarla. Por lo tanto, se pueden implementar métodos de clasificación de texto que permitan de forma automática, determinar en cuál sección del diario deberán aparecer las notas, apoyando así al humano en este proceso. Adicionalmente estos métodos pueden ser utilizados sobre noticias que aparecen en Internet y que no han sido previamente clasificadas, para recuperar aquellas relacionadas con alguna sección específica. Debido a lo anterior, en este trabajo terminal se propone realizar un clasificador de noticias que analice el contenido de estas y pueda determinar a qué sección del diario corresponden.

Existen algunos trabajos similares al propuesto. En [2] los autores recolectaron 3000 artículos de un periódico publicado por el Instituto Tecnológico de Massachusetts (MIT por sus siglas en Inglés) y utilizando aprendizaje supervisado probaron diferentes algoritmos de aprendizaje automático para clasificar dichos artículos en 6 categorías diferentes. en [3] se propone un método para clasificar noticias publicadas tanto en inglés como en japonés y además determinan los tópicos de los cuales se habla en las noticias.

También existen algunas herramientas que realizan análisis de texto. Microsoft Cognitive Services [4] a través de su aplicación llamada "Text Analytics" permite determinar los tópicos de los cuales se habla en un conjunto de documentos y extraer las palabras clave de ellos. Otra herramienta es IBM Watson[5] que mediante su aplicación "Alchemy Language" permite clasificar el contenido de páginas web o identificar qué tópicos están siendo más comentados en las

Luz María Sánchez García

Ender Hernández Contreras

noticias.

Una diferencia importante del trabajo propuesto con respecto a las referencias anteriores es el hecho de trabajar con noticias de tres diferentes diarios y además en español. Es importante mencionar que cada idioma tiene sus particularidades y una solución que funciona bien para un idioma no siempre obtiene los mismos resultados en otro, por eso es importante proponer una solución para el caso particular de nuestro idioma. Hasta el momento no se han encontrado referencias de esta tarea específica con diarios de circulación nacional.

2. Objetivo

Clasificar de forma automática noticias de diarios de circulación nacional según su tema y contenido en las diferentes secciones existentes.

Objetivos específicos.

1. Recolectar noticias de los diarios: La Jornada, Excélsior y El Universal (100 de cada uno[1])
2. Extraer características de las noticias, algunas características son: frecuencia de palabras y representación vectorial de palabras.
3. Clasificar noticias en secciones de cada periódico.
4. Seleccionar las técnicas que obtengan los mejores resultados. Algunos de los algoritmos a probar serán Naive Bayes, Máquinas de Soporte Vectorial y Regresión Logística. Para evaluar los resultados del clasificador se utilizará la medida de *exactitud*, la cual determina la razón de instancias clasificadas correctamente con respecto al total de noticias.
5. Crear un modelo para clasificar nuevas noticias a partir de las técnicas seleccionadas.

3. Justificación

La tarea de identificar a qué sección pertenece una noticia dentro de un diario, es una tarea costosa en tiempo y esfuerzo. Por lo tanto, los resultados de este trabajo terminal permitirán:

1) Automatizar el proceso de clasificación de noticias en secciones.
2) Seleccionar los algoritmos de extracción y clasificación con la medida de exactitud mayor.
3) Identificar las características de las noticias que hacen que estas pertenezcan a una sección del diario.

Una vez identificadas las características y algoritmos que mejor realicen la tarea de clasificación, el proceso de determinar a qué sección pertenecen las noticas se podrá realizar de forma automática.

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

Durante el desarrollo de este trabajo terminal se utilizarán los conocimientos obtenidos en la carrera misma como los que se imparten en las asignaturas de artificial intelligence, teoría computacional, análisis de algoritmos y bases de datos, por mencionar algunas.

4. Productos o Resultados esperados

Se pretende tomar las noticias de diarios nacionales como referencia para nuestro trabajo terminal y poder con base en ellos crear una base sólida para poder clasificar noticias futuras quedando los productos a realizar como sigue:

1. Modelo clasificador de noticias
2. Documentación del análisis y diseño
3. Manual de usuario.

[1] Este número puede variar dependiendo de la accesibilidad de cada uno de los diarios

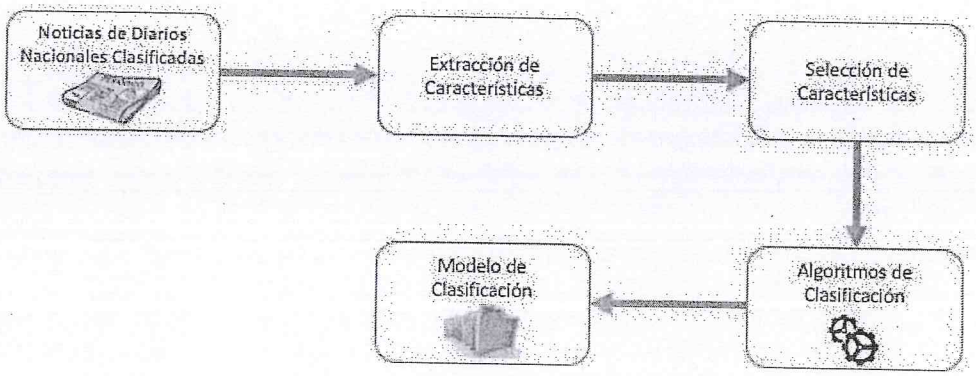


Figura 1. Diagrama de etapa de entrenamiento

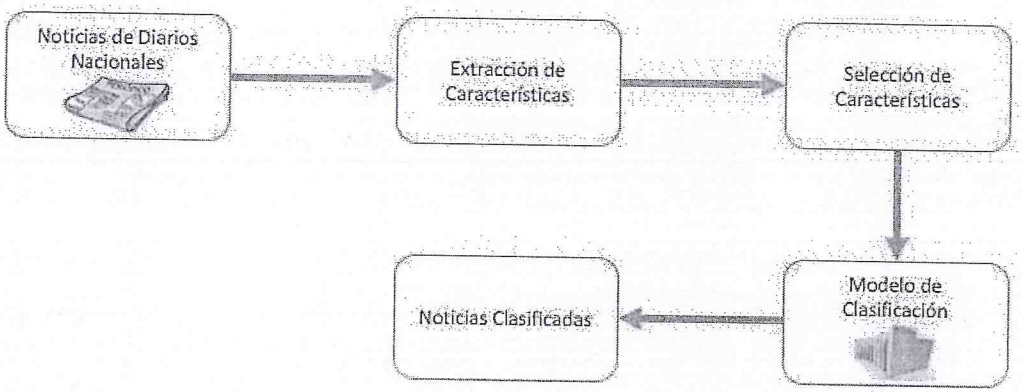


Figura 2. Diagrama de etapa de prueba

5. Metodología

La metodología a emplear será la incremental, una de las bondades de esta metodología es la entrega de resultados en un corto periodo de tiempo, esto con el fin visualizar nuestros avances gradualmente y poder hacer los ajustes pertinentes en cada paso del desarrollo.

Las fases de la metodología incremental son:

- Análisis
- Diseño
- Desarrollo
- Pruebas

Después de cada incremento es posible entregar un producto funcional que será evaluado para ver si cumple los requerimientos necesarios. Se pueden gestionar los riesgos técnicos relacionados con la magnitud del proyecto para realizar correcciones o seguir con otro incremento.

Se crean varias versiones hasta llegar a un producto terminado, en donde las primeras versiones tendrán las funciones más importantes que se requieran.

6. Cronograma

Nombre del alumno(a): José Alejandro García Molina
Título del TT:

Actividad	AGO	SEP	OCT	NOV - DIC	ENE	FEB	MAR	ABR	MAY - JUN
Investigación de trabajos relacionados.									
Recolección de noticias de diarios de circulación nacional.									
Investigación de métodos de selección de características.									
Pruebas iniciales de selección de características.									
Documentación de trabajo terminal.									
Pruebas y selección de métodos de extracción de características.									
Implementación del modelo de clasificación de noticias.									
Pruebas del modelo de clasificación de noticias.									
Documentación de trabajo terminal.									

Nombre del alumno(a): Ramírez Roque Luis Enrique
Título del TT:

Actividad	AGO	SEP	OCT	NOV - DIC	ENE	FEB	MAR	ABR	MAY - JUN
Recolección de noticias de diarios de circulación nacional.									
Investigación de métodos de extracción de características.									
Instalación y pruebas de herramientas para la calificación de texto.									
Pruebas iniciales para la calificación de texto.									
Pruebas y selección de algoritmos de clasificación de texto.									
Implementación del modelo de clasificación de noticias.									
Implementación del modelo de clasificación de noticias.									
Documentación de trabajo terminal.									

Nombre del alumno(a): Sánchez Ramírez Miguel Ángel
Título del TT:

Actividad	AGO	SEP	OCT	NOV - DIC	ENE	FEB	MAR	ABR	MAY - JUN
Investigación de trabajos relacionados.									
Recolección de noticias de diarios de circulación nacional.									
Investigación de algoritmos de clasificación de texto.									
Pruebas iniciales de extracción de características.									
Documentación de trabajo terminal.									
Pruebas y selección de métodos de selección de características.									
Pruebas y selección de herramientas para la clasificación de texto.									
Pruebas del modelo de clasificación de noticias.									
Implementación del modelo de clasificación de noticias.									

7. Referencias

[1] Ávila Argüelles Ricardo (2008), Clasificación Bibliotecaria Automática Usando Identificación Simple de Términos con Métodos Lógico-Combinatorios a Partir de Información Escasa (Tesis de posgrado), Centro de Investigación en Cómputo, México.

[2] Kamal Nigam, Andrew Kachites Mccallum, Sebastian Thrun, Tom Mitchell, Text Classification from Labeled and Unlabeled Documents using EM, Kluwer Academic Publishers,Boston. Manufactured in The Netherlands.

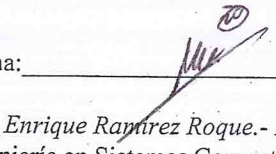
[3] David B. Bracewell, Jiajun Yan, Fuji Ren, Shingo Kuroiwa, Category Classification and Topic Discovery of Japanese and English News Articles, Electronic Notes in Theoretical Computer Science 225 (2009) 51–65.

[4] Microsof Cognitive Services, información recopilada de la página: https://blogs.msdn.microsoft.com/esmsdn/2016/04/01/microsoft-cognitive-services/

[5] IBM Watson AlchemyLanguage, información recopilada de la página: https://www.ibm.com/watson/developercloud/alchemy-language.html

8. Alumnos y Directores

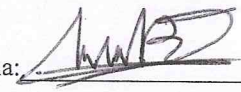
García Molina José Alejandro.- Alumno de la carrera de Ingeniería en Sistemas Computacionales en la Escuela Superior de Cómputo del Instituto Politécnico Nacional, Boleta: 2014630175, Tel. 5551381743, email: alealejandromolina@gmail.com

Firma: 

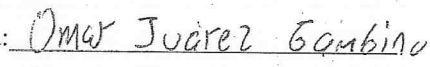
Luis Enrique Ramírez Roque.- Alumno de la carrera de Ingeniería en Sistemas Computacionales en la Escuela Superior de Cómputo del Instituto Politécnico Nacional, Boleta: 2014350598, Tel. 5520147111, email: luis.e.3194@gmail.com

Firma: 

Sánchez Ramírez Miguel Ángel.- Alumno de la carrera de Ingeniería en Sistemas Computacionales en la Escuela Superior de Cómputo del Instituto Politécnico Nacional, Boleta: 2014350699, Tel. 5533526800, email: miguelsanchezr2014@gmail.com

Firma: 

Joel Omar Juárez Gambino.- Licenciado en Informática por la Facultad de Informática, UAS. Maestro en Ciencias de la Computación por el CIC, IPN. Sus áreas de estudio son: Inteligencia Artificial, Lenguaje Natural y Representación de Conocimiento. Departamento de Ciencias e Ingeniería de la Computación, ESCOM, Tel. 57296000 Ext. 52022, email: omarjg@gmail.com

Firma: 

Consuelo Varinia García Mendoza.- Ingeniera en Sistemas Computacionales por la ESCOM, IPN, Maestra en Ciencias en Tecnología Avanzada por el CICATA-Legaria, IPN, Doctora en Tecnología Avanzada por el CICATA-Legaria, IPN. Sus áreas de estudio son: Análisis de algoritmos y Optimización. Departamento de Ciencias e Ingeniería de la Computación, ESCOM, Tel. 57296000 Ext. 52022, email: consuelo.varinia@gmail.com

Firma: