

# Identificación y agrupación de noticias por categorías usando Procesamiento de Lenguaje Natural

## Trabajo Terminal No. \_\_ - \_\_

*Alumna: \*Jiménez López Diana Laura*

*Directores: Juárez Gambino Joel Omar, Calvo Castro Francisco Hiram*

*\*dianaljl.99@gmail.com*

**Resumen** - En este trabajo terminal se propone identificar de forma automática noticias y agruparlas en diferentes categorías a partir del análisis de su contenido. Mediante técnicas de procesamiento de lenguaje natural se analizará un conjunto de documentos para encontrar características que definen a una noticia para desarrollar el método de identificación y así distinguirlas del resto. Posteriormente aquellos documentos identificados como noticias se agruparán en categorías, por ejemplo, aquellas relacionadas a economía, política, deportes, etc. esto en una primera aproximación, con técnicas de agrupamiento. A diferencia de algunos métodos que realizan este trabajo de manera supervisada, en esta propuesta no se requiere tener un conjunto de datos previamente etiquetado para poder distinguir entre documentos que son noticia y asignarles una categoría. Es importante señalar que mucha de la información disponible en Internet carece de estas etiquetas, por lo que el enfoque no supervisado propuesto en este trabajo ayudará a solventar parte de esta problemática. Finalmente, se determinará el grado de precisión de la herramienta comparando los resultados de ambas tareas (identificación de noticia y asignación de categoría) contra un conjunto de documentos donde se tiene identificada esta información.

**Palabras clave** -Análisis de documentos, procesamiento de lenguaje natural, agrupamiento, aprendizaje no supervisado

## 1. Introducción

El periodismo desde hace mucho tiempo ha sido una herramienta sumamente importante para la sociedad ya que permite informarse sobre los acontecimientos importantes o relevantes, esta información llega a la gente en forma de noticias. En el siglo XVIII los diarios eran el principal medio de difusión de estas, mientras que en el siglo XXI medios como la televisión, la radio, y actualmente el Internet son la principal fuente de difusión de noticias [1]. Las noticias son documentos que tienen un impacto social directo, lo que provoca un fuerte interés en su estudio. Debido a la gran cantidad de información disponible hoy en día, la automatización de los procesos de análisis de noticias es un campo de gran interés ya que estas tareas son laboriosas y tardadas.

Existe mucho trabajo relacionado con el análisis automático de noticias. Por ejemplo, algunos autores han propuesto métodos para hacer resúmenes de noticias [2], identificar los tópicos principales de estas [3], su relevancia [4], e incluso se puede determinar si una noticia es falsa [5]. Todos estos trabajos requieren de un conjunto de documentos previamente identificados como noticias y este proceso de identificación generalmente se hace de forma manual lo cual requiere invertir tiempo y esfuerzo. La tarea de identificar de forma automática una noticia, cobra cada día más relevancia debido a la gran cantidad de información que se tiene disponible en Internet. Dada la problemática anterior, en este trabajo terminal se propondrán algunos métodos que permitan identificar de forma automática una noticia mediante el análisis de su contenido. Lo anterior se realizará siguiendo un enfoque no supervisado, para lo cual no se requiere tener un conjunto de documentos previamente identificados como noticias Cabe señalar que el enfoque a seguir aplicado a esta tarea ha sido poco explorado.

Por otro lado, la agrupación de noticias por categorías es muy importante ya que nos facilita el acceder, buscar y filtrar noticias. Este proceso requiere un trabajo arduo y tardado por lo que se ha buscado hacerlo automáticamente, tenemos por ejemplo los Trabajos Terminales realizados en la ESCOM donde se propone clasificar mediante técnicas de aprendizaje automático noticias de diarios de circulación nacional [6] y una aplicación web que recolecta y clasifica noticias por su contenido y fecha de publicación [7]. El trabajo propuesto difiere de los Trabajos Terminales antes mencionados ya que al igual que la identificación de noticias, la tarea de agrupación por categorías se realizará de forma no supervisada mediante el uso de técnicas del Procesamiento de Lenguaje Natural como son etiquetado por parte de oración (*POS tagging*), lematización, uso de características sintácticas, entre otras.

Ambas tareas como ya mencionamos normalmente se abordan desde un enfoque supervisado, este enfoque exige tener datos etiquetados, que no siempre se tienen disponibles por lo que implica una tarea extra, el etiquetado manual, lo cual suele ser muy tardado, tedioso y poco subjetivo, incluso se paga por este servicio, por ejemplo existe *Amazon Mechanical Turk*.

## 2. Objetivo

Identificar de manera automática noticias a partir de un conjunto grande de documentos en inglés, esto se logrará desarrollando un modelo capaz discriminar noticias, con base en el análisis de las características extraídos con técnicas de procesamiento de lenguaje natural y posteriormente agruparlas en diferentes categorías como son deportes, política, etc. basándose en su contenido con técnicas de agrupamiento.

### Objetivos específicos:

- Extraer las características lingüísticas de un documento como léxicas y sintácticas que nos ayuden con la tarea de identificar noticias
- Separar los documentos que cumplan con las características de una noticia
- Agrupar los documentos identificados como noticias en categorías (por ejemplo, deportes, política, etc.)

## 3. Justificación

Los trabajos relacionados con el análisis automático de noticias (mencionados en la sección anterior) tienen un impacto social fuerte, sin embargo, la mayoría de estos parten de la premisa de que se está trabajando con noticias. La identificación automática de noticias es un campo poco explorado lo que hace del trabajo propuesto novedoso y útil para cualquier otra tarea que se desee realizar con este tipo de documentos, por otro lado, la agrupación por categorías se ha abordado en múltiples trabajos, pero siempre desde un enfoque supervisado que requiere de documentos etiquetados manualmente lo cual es costoso en términos de tiempo y esfuerzo. El trabajo propuesto permitirá realizar esta tarea sin un conjunto de documentos previamente etiquetado.

En la siguiente gráfica podemos comprobar que estas tareas abordadas desde un enfoque no supervisado han sido poco exploradas, al ser desde un enfoque no supervisado se ahorran el tiempo, esfuerzo e incluso dinero que se pudiera invertir para el etiquetado de datos que se requiere para hacerlo con enfoque supervisado.

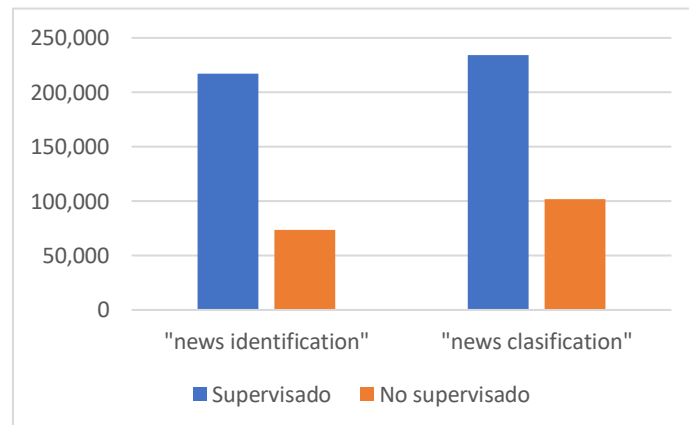


Figura 1. Gráfica que muestra los resultados en una búsqueda sobre los temas a desarrollar en Google Scholar.

#### 4. Productos o resultados esperados

- Método identificador de noticias.
- Método agrupador de noticias.
- Documentación del análisis y diseño.

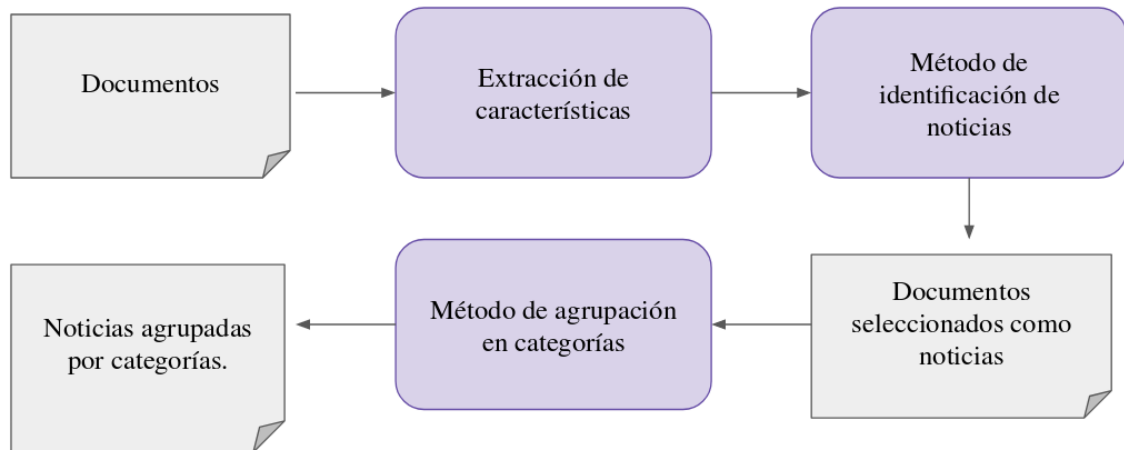


Figura 2. Diagrama del modelo general de la propuesta.

#### 5. Metodología

La metodología que se utilizará será por prototipos ya que permite analizar, diseñar, implementar y probar algunas funcionalidades del sistema, y de esa forma ir cubriendo cada uno de los objetivos específicos. Estamos proponiendo dos tareas diferentes: identificación de noticias y agrupación por categorías de estas, la metodología elegida nos permite desarrollarlas por separado, implementando los prototipos correspondientes para cada tarea [8].

Ambas tareas suponen problemáticas complejas debido que se ha trabajado poco sobre ellas, por ende es necesario ir definiendo diferentes metas y requerimientos en cada iteración de la metodología.

Las etapas de esta metodología son:

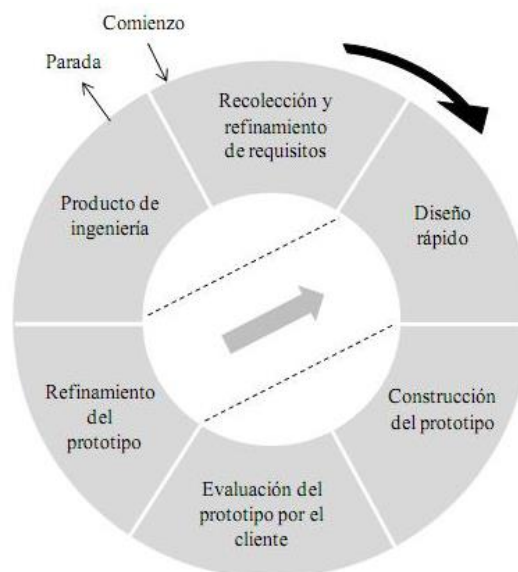


Figura 3. Diagrama de la metodología.

## 8. Cronograma


Actividad	Ago	Sep	Oct	Nov	Dic	Ene	Feb	Mar	Abr	May	Jun
Investigación de trabajos relacionados											
Obtención de documentos											
Análisis y selección de las características de los documentos											
Desarrollo del método identificador de noticias											
Implementación prototipo I											
Evaluación TTI											
Investigación de trabajos relacionados a agrupamiento											
Análisis y selección de características de agrupamiento											
Desarrollo de método de agrupación de noticias											
Implementación del prototipo de agrupación											
Integración de los dos prototipos											
Evaluación de TT 2											
Documentación											

## 7. Referencias

- [1] Copeland, D. A., Martín, S. E., Miller, B. A., & Merrill, J. C. (2003). *The function of newspapers in society: A global perspective*. Greenwood Publishing Group.
- [2] Barzilay, R., & McKeown, K. R. (2005). Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3), 297-328.
- [3] Bracewell, D. B., Yan, J., Ren, F., & Kuroiwa, S. (2009). Category classification and topic discovery of japanese and english news articles. *Electronic Notes in Theoretical Computer Science*, 225, 51-65.
- [4] Phuvipadawat, S., & Murata, T. (2010, August). Breaking news detection and tracking in Twitter. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (Vol. 3, pp. 120-123). IEEE.
- [5] Conroy, N. J., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, 52(1), 1-4.
- [6] García Molina José Alejandro, Ramírez Roque Luis Enrique y Sánchez Ramírez Miguel Ángel (2018). Clasificación de noticias de diarios de circulación nacional mediante aprendizaje automático. Trabajo Terminal de ESCOM-IPN con número 2017-A04, (CDMX).
- [7] Hernández Gómez Carlos Andrés y Meza Martínez Luis Daniel (2019). Recolector y clasificador de noticias, Trabajo Terminal de ESCOM-IPN con número 2018-B013, (CDMX).
- [8] Naumann, J. D., & Jenkins, A. M. (1982). Prototyping: the new paradigm for systems development. *Mis Quarterly*, 29-44.

## 8. Alumna y directores

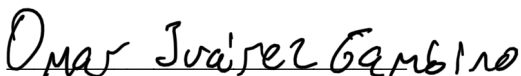
Jiménez López Diana Laura. - Alumna de la carrera de Ing. en Sistemas Computacionales en ESCOM Boleta: 2015010528 Tel. (55) 4485 9869, email [dianaljl99@gmail.com](mailto:dianaljl99@gmail.com)

Firma: 

Calvo Castro Francisco Hiram. - Postdoctorado en Lingüística Computacional por el Instituto en Ciencia y Tecnología de Nara Japón. Profesor Investigador de tiempo completo en el CIC- IPN desde 2006. Áreas de interés: procesamiento del lenguaje natural, aprendizaje automático, inteligencia artificial. Tel. (55) 5729 60 00, ext.: 56516, email [hcalvo@cic.ipn.mx](mailto:hcalvo@cic.ipn.mx)

Firma: 

Juárez Gambino Joel Omar. - Doctor en Ciencias de la Computación por CIC-IPN, Profesor de ESCOM desde 2009, Áreas de interés: Inteligencia Artificial, Lenguaje Natural, Representación de Conocimiento. Tel. (55) 5729 60 00, ext.: 52022, email [jjuaresg@ipn.mx](mailto:jjuaresg@ipn.mx)

Firma: 

CARÁCTER: Confidencial  
FUNDAMENTO LEGAL: Art. 3, frac II, Art. 18, frac II y  
Art. 21, lineamiento 32, frac XVII de la L.F.T.A.I.P.G.  
PARTES CONFIDENCIALES: No. de boleta y Teléfono.