

Estudio de identificación de autoría

Ciencias de la Computación e Inteligencia Artificial



12 de marzo de 2019

Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación

Tutores: Luis de Campos Ibáñez y Juan F. Miguel Huete

Índice

[**1.** **Introducción** 2](#_Toc9879716)

[**1.1** **Propósito** 2](#_Toc9879717)

[**1.2** **Partes** 2](#_Toc9879718)

[**1.3** **Justificación** 2](#_Toc9879719)

[**1.4** **Ámbito** 2](#_Toc9879720)

[**1.5** **Relevancia teórica y práctica de la investigación** 3](#_Toc9879721)

[**1.6** **Estado de la cuestión** 3](#_Toc9879722)

[**1.7** **Objetivos, exposición del problema y preguntas de investigación** 4](#_Toc9879723)

[**1.8** **Breve descripción de la metodología de la investigación** 4](#_Toc9879724)

[**1.9** **Esquema** 4](#_Toc9879725)

[2 Profundizar sobre la autoría 5](#_Toc9879726)

[3 Objetivo 5](#_Toc9879727)

[4 Datos 5](#_Toc9879728)

[5 Preprocesamiento 5](#_Toc9879729)

[6 Clasificación 5](#_Toc9879730)

[7 Comparativa 5](#_Toc9879731)

[8 Estudio 5](#_Toc9879732)

[9 Conclusión 5](#_Toc9879733)

[10 Bibliografía 5](#_Toc9879734)

[10.1 Iniciativas parlamentarias 5](#_Toc9879735)

[10.2 Una Encuesta de Métodos Modernos de Atribución de Autoría 5](#_Toc9879736)

[10.3 El efecto del tamaño del conjunto de autores y el tamaño de los datos en la atribución de autoría 5](#_Toc9879737)

# **Introducción**

## **Propósito**

A lo largo de la historia se han producido numerosos debates sobre la autoría de obras transcendentales para el conocimiento humano. El hecho de conocer el autor de un contenido da un peso conceptual extra a sus palabras que, apoyado por su biografía y su circunstancia crean un marco decisivo desde el que poder abarcar cualquier estudio. Esto es debido a que a veces podría ser más conclusivo responder a la pregunta, ¿quién desarrolló el contenido? Que el contenido mismo.

El propósito de este trabajo es automatizar la identificación del autor de un documento sobre de un conjunto de autores previamente definido.

## **Partes**

El trabajo constará de 10 secciones definidas en el índice inicial y fácilmente consultables. El formato electrónico permite navegabilidad sobre el mismo gracias a la inserción de vínculos sobre las entradas. A lo largo del trabajo se pueden encontrar referencias en forma de vínculos al apartado bibliográfico que se encuentra en la última sección de este documento.

## **Justificación**

He decido abordar este trabajo tras la lectura de [una Encuesta de Métodos Modernos de Atribución de Autoría](#_Una_Encuesta_de) y el [efecto del tamaño del conjunto de autores y el tamaño de los datos en la atribución de autoría](#_El_efecto_del). Ambas investigaciones fueron sugeridas por los tutores del proyecto.

También existe un componente previo personal que me ha llevado a aceptar esta temática. Mi afición por la lectura ha ocupado gran parte de mi vida, sobre todo por los distintos estilos literarios, y unido a mi interés por la psicología humana sobre la cual daremos algunas pincelas a lo largo del trabajo.

Desde un punto de vista laboral, mi trabajo actual como Científico de Datos para la Prevención del Fraude y el Crimen Organizado comparte muchas áreas de conocimiento con el tema que nos ocupa. Desde la identificación de nombres sobre listas sancionadas (Watch List Filtering) haciendo uso de emparejamiento por lógica difusa, hasta la identificación y clasificación de alertas sospechosas en los conceptos de las transferencias bancarias.

Desde el punto de vista de las ciencias de la computación y la inteligencia artificial rama del conocimiento que estudio, este problema hace uso de muchos de los conceptos y herramientas que en ellas se explican. Debido a que ha sido una decisión propia el estudio de esta especialidad, es razonable la elección del proyecto.

## **Ámbito**

El problema de la Identificación de Autoría o la Autoría de Documentos es una cuestión interdisciplinar que comparten actualmente lingüística, minería de datos e inteligencia artificial. Siendo estos 2 últimos desde los que se va abordar el trabajo adentrándonos en la subrama aprendizaje automático perteneciente a la inteligencia artificial.

Podemos decir que los dos ámbitos específicos que predominan en este trabajo son la minería de textos y el aprendizaje automático. A continuación, daremos una definición básica de ambos conceptos:

Minería de textos

Conversión de grandes volúmenes de texto en estructuras simplificadas y comprensibles para su uso posterior.

Aprendizaje automático

Estudio de algoritmo de computación que mejoran automáticamente por medio de la experiencia.

Si quisiéramos profundizar más existen números subcategorías dentro del problema al que nos enfrentamos, dependiendo del registro del lenguaje, permisibilidad de faltas de ortografía, tamaño de los textos, número de autores, número de muestras… Más adelante concretaremos estos conceptos en el apartado 2.

## **Relevancia teórica y práctica de la investigación**

[resumen del artículo de la encuesta y algo más]

Relevancia teórica

{Explicar los posibles avances dentro del campo }

A continuación, se presentan algunas aplicaciones que tiene el problema:

* Desde la literatura nos encontramos con debates actuales sobre la identificación de autores de obras anónimas o puesta en duda de obras que fueron atribuidas sin estudio previo. Pero la realidad es que muchas obras continúan anónimas actualmente ya sea por deterioro de la misma o falta de candidatos.
* Un ejemplo de la primera es la conocida obra del Lazarillo de Tormes que tras una lista de candidatos se resolvió su autoría gracias a un estudio lingüista reconociendo a Sebastián de Horozco como escritor.
* Un ejemplo de la segunda es el debate que hubo en torno a las obras de Shakespeare debido a su estilo impropio de la cuna del autor y a las lagunas de su biografía.
* Desde la criminología, derecho, psicología y psiquiatría nos encontramos considerable número de problemas.
* Peritaje de conversaciones electrónicas.
* Autoría de cartas de suicidio.
* Falsificación en las relaciones laborales.
* Falsificación de estudios (TFGs y TFMs).
* Identificación de trastornos.
* Psicología evolutiva y del aprendizaje.
* Desde la automatización de tareas, como es nuestro caso, tenemos algunos ejemplos.
* Documentación automática (nuestro caso)
* Detección de suplantaciones de identidad

## **Estado de la cuestión**

## **Objetivos, exposición del problema y preguntas de investigación**

## **Breve descripción de la metodología de la investigación**

## **Esquema**

Que es el problema de la Autoría de Documentos (Autorships).

La iniciativa parlamentaria permite a un diputado o a un senador presentar al parlamento un proyecto de artículo constitucional, de ley o de decreto. Este proyecto puede ser redactado de modo completo o formulado en términos generales. La comisión de la cámara donde ha sido depositada la iniciativa decide si es necesario darle curso. Por ejemplo, una iniciativa sobre un tema que ya está en discusión en el parlamento no será declarada válida. Si la comisión considera que la iniciativa puede ser acogida, el proyecto sigue el itinerario legislativo clásico (examen por parte de la comisión de la otra cámara, procedimiento de consulta, cámaras del parlamento, etc.)

Desde 1978 se recogen las iniciativas parlamentarias, pero nosotros nos centraremos 2008.

Fuente de las iniciativas parlamentarias:

Lenguaje parlamentario, corpus

Lenguaje oral cuidado

Lenguaje político de controversia, trufado por citas generales (con caracteres atractivos) función de alusión al otro. Estructura: argumentación, muy estructurado.

Extracción de características:

1. Características generales
2. Características lingüísticas
3. Características temáticas
4. Caracteristicas a nivel de palabras

Explicar matriz dispersa

Explicar función logarítmica para la

# Profundizar sobre la autoría

# Objetivo

Estos son los objetivos del proyecto:

1. Realizar un **preprocesamiento** correcto de los **datos** para que puedan ser utilizados como entrada en un modelo de aprendizaje automático.
2. Proponer diversos modelos entrenados de clasificación multietiqueta que presenten buenos resultados.
3. Validar los modelos y compararlos entre sí.
4. Definir un criterio justificado para elegir el mejor modelo y presentarlo en un ambiente de producción.

# Datos

La fuente de datos utilizada son la Iniciativas Parlamentarias del Senado durante el año 2008. Estas iniciativas están recogidas en 5260 ficheros en formato XML. El formato XML presenta la siguiente estructura:

…

…

…

El

El **nombre**

En el esquema explicado se puede observar que

En nuestro caso los datos los hemos extraídos de las [iniciativas parlamentarias del congreso de los diputados del año 2008](#_Iniciativas_parlamentarias) que presentan mayormente un registro formal, una baja permisibilidad de faltas de ortografía, un tamaño medio reducido de los textos, entorno autores y con muestras.

Estrictas reglas

Géneros parlamentarios

Subjetividad (propaganda) VS Información (Debate intrínseco)

Extraccion del tema

Que el texto sea argumentativo significa que usa palabras del contrario en las respuestas. Para demostrar esto dar la matriz de intersección

# Preprocesamiento

Explicar preprocesamiento de los XML

-Eliminar palabras especiales

-Agrupar párrafos en un solo documento

# Clasificación

Explicar 2 modelos:

* SGDClassifier
* Random forest classifier
* Support Vector Machine

Grafica de separación para cada uno generada

# Comparativa

Definir score

Balanceados y no balanceado

Ajuste de hiperparamentros

Graficas sobre hiperparametros

# Estudio

# Conclusión

# Bibliografía

## [Iniciativas parlamentarias](http://www.congreso.es/consti/constitucion/indice/sinopsis/sinopsis.jsp?art=87&tipo=2)

## [Una Encuesta de Métodos Modernos de Atribución de Autoría](http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.440.1634&rep=rep1&type=pdf)

## [El efecto del tamaño del conjunto de autores y el tamaño de los datos en la atribución de autoría](https://www.researchgate.net/publication/220675445_The_effect_of_author_set_size_and_data_size_in_authorship_attribution)