Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias de la Computación

Atribución de autoría combinando información léxico - sintáctica mediante representaciones holográficas reducidas

Tesis presentada por Jovany Marcos Ramírez
para obtener el grado de Maestro en Ciencias de la Computación

Asesores

Maya Carrillo Ruíz, Manuel Montes y Gómez

2015

Índice general

Índice de figuras

Indice de tablas

Agradecimientos

Le agradezco a Dios por haberme acompañado y guiado a lo largo de mi carrera, por ser mi fortaleza en momentos de debilidad y por brindarme una vida llena de aprendizajes, experiencias y sobre todo felicidad.

Mi más grande agradecimiento a mis directores de tesis Dra. Maya Carrillo y Dr. Manuel Móntes , por la confianza, apoyo y dedicación a lo largo de esta investigación. Por haber compartido conmigo sus conocimientos y sobre todo su amistad.

Desde luego gracias al apoyo y al cariño brindado por mis padres y hermana, pilares fundamentales en mi vida. Sin ellos, jamás hubiese podido conseguir lo que hasta ahora. Por los valores que me han inculcado, y por haberme dado la oportunidad de tener una excelente educación en el transcurso de mi vida. Sobre todo por ser un excelente ejemplo de vida a seguir.

A mis amigos de maestría me encantaría agradecerles su amistad, consejos, apoyo, ánimo y compañía en los momentos más dificiles.En general son muchas las personas que han convivido conmigo a lo largo de esta tesis. Algunas están conmigo y otras en mis recuerdos y en mi corazón sin importar donde estén, desde lo ms profundo de mi corazón les agradezco el haberme brindado todo el apoyo, colaboración, ánimo y sobre todo su cariño y amistad.

A todos mi mayor reconocimiento y gratitud. Y que Dios los bendiga.

Jovany

Capítulo 1

Introducción

1.1. Resumen

El presente trabajo, busca determinar si la tarea de atribución de autoría puede beneficiarse, con la combinación de características extraídas de textos, de diferente nivel gramatical. De acuerdo a los trabajos revisados, las características léxicas y en particular el uso de n-gramas de caracteres han producido resultados para la atribución de autoría con un 76 % de precisión en promedio. Se ha visto que los trabajos en los que se utilizan características sintácticas como son n-gramas sintácticos (sn-gramas) han obtenido un 95 % de precisión en corpus de 39 documentos; las gramáticas libres de contexto probabilísticas han producido resultados con precisión de 83 % en corpus de 120 documentos. Estos resultados nos indican que la utilización de aspectos sintácticos, utilizando representaciones textuales novedosas puede conducir a la obtención de resultados aceptables. En particular, en este trabajo se utiliza la representación holográfica reducida para combinar información léxica y sintáctica, representar textos de diferentes autores y comprobar si dicha representación contribuye a mejorar la precisión de la tarea de atribución de autoría.

1.2. Planteamiento del problema

Hoy en día existe una enorme cantidad de información en la web, más del 75 % de ella en texto, misma que se encuentra en redes sociales, foros, códigos fuentes,

etc. Teniendo tanta información disponible no es posible analizarla de forma manual, surge así, la necesidad de encontrar soluciones que nos permitan acceder de manera automática a la información de dichos textos. En particular, la tarea de identificar al autor de algún escrito anónimo, Atribución de Autoría (AA), requiere de la utilización de diversas técnicas, que permitan extraer características estilométricas de los textos, para identificar al autor.

La presente investigacón se sitúa dentro de la tarea de AA, a continuación los objetivos planteados.

1.3. Objetivos Generales y Específicos

Objetivo General

1. Diseñar e implementar un método para la tarea de atribución de autoría que combine atributos de diferente nivel gramatical a fin de evaluar la utilidad de los mismos.

Objetivos Específicos

1. Determinar los atributos que de manera conjunta, permitan caracterizar el estilo de escritura de los autores.
2. Experimentar con representaciones novedosas de documentos que permitan capturar características de diferentes niveles gramaticales.
3. Explorar la utilidad de técnicas de reducción de dimensión vectorial en atribución de autoría.

El resto del documento esta organizado de la siguiente manera: en el capítulo 2 se presentan los trabajos relacionados con la investigación que nos ocupa, así como los conceptos teóricos utilizados durante la misma. En el caítulo 3 se describe el metodo propuesto. En el capítulo 4 se muestran los resultados y experimentos obtenidos por último en el capítulo 5 se encuentran las conclusiones y el trabajo a futuro.

Capítulo 2

Fundamento Teórico

2.1. Antecedentes del Proyecto

La Atribución de Autoría es una tarea, método o mecanismo de trabajo que los expertos utilizan desde el siglo XIX para situar obras anónimas en relación con una época, lugar y un autor [9] , algunos ejemplo de ellos son El burlador de Servilla, Doña Blanca , autorías como obras de Tirso, Francisco de Lyra, Manuel de Sandes, etc. [4], otros acercamientos científicos a la AA fueron hechos por Mendanhall en el sigo XIX quien estudió textos de Bacon, Marlowe y Shakespeare; y Mascol que estudió la autoría de los evangelios del nuevo testamento [4].

De esta manera podremos darnos cuenta que la Atribución de Autoría no es una tarea con poco tiempo de investigación , sino que es una tarea de mucho tiempo atrás. Actualmente algunas de las aplicaciones en las que esta tarea apoya son : acoso sexual, notas suicidas, ataques terroristas, derechos de autor, código fuente, detección de plagio o atribución de algún autor a un documento anónimo [30]. Así las investigaciones para desarrollar nuevos métodos de atribución de autoría son importantes, pues atacan problemáticas vigentes.

2.2. Estado del Campo o del Arte

La Atribución de Autoría (AA), se apoya en modelos estadísticos y computacionales para la identificación de características textuales, que permitan modelar el estilo de diferentes autores a partir de sus textos. Los primeros intentos de medir el estilo de escritura los realizó Mendenhall (1887) sobre obras de Shakespeare, seguido de estudios en la primera mitad del siglo XX por Yule (1938,1944) y Zipf (1932). Más tarde Mosteller y Wallace (1964) realizaron estudios sobre la autoría de artículos federalistas, los cuales contenían una serie de 146 ensayos políticos escritos por John Jay, Alexander Hamilton y James Madison, de estos existen 12 escritos cuyo autor no se ha identificado de manera precisa, pudiendo ser atribuidos a Hamilton o Madison. Mosteller y Wallace, utilizaron el análisis bayesiano de frecuencias de un pequeño conjunto de palabras comunes, como: y, a (preposiciones y conjunciones). Este estudio dio inicio a métodos no tradicionales de AA. Desde entonces y hasta finales de 1990, la investigación en AA se concentró en definir las características para identificar estilos de escritura, esto se conoce como Estilometría [7]. Por lo tanto, una gran variedad de medidas se propuso, incluyendo longitud de oraciones, longitud de palabras, frecuencia de palabras, frecuencia de caracteres y funciones para determinar la riqueza del vocabulario. Durante esta época se desarrollaron métodos para la AA que no eran completamente automáticos, un ejemplo es el método CUSUM (o QSUM) [19], que fue aceptado por tribunales como prueba pericial para determinar la AA, pero rechazada por los especialistas por el hecho de ser poco fiable [25].

Desde finales de 1990, las cosas han cambiado en los estudios de AA. La gran cantidad de los textos disponibles en forma electrónica o escrita, han aumentado la necesidad de manejar de manera eficiente la información. Este hecho, impactó significativamente áreas como: la recuperación de información, aprendizaje de máquina, y el procesamiento del lenguaje natural (NLP). Las investigaciones en esta última han dado origen a herramientas capaces de analizar el texto de manera automática y eficiente, proporcionando nuevas formas de capturar y representar el estilo de escritura [1].

Los estudios sobre AA han propuesto diversos atributos para identificar el estilo de escritura, llamando a estos: marcadores de estilo [**?**], [26]. Actualmente la representación de textos en AA, se centra en características léxicas y de caracteres, considerandolos como una secuencia de palabras o caracteres. Seguido de características sintácticas y semánticas las cuales requieren un análisis lingũístico más profundo[1], algunas de estas características se mencionan a continuación.

2.2.1. Características estilométricas

La Estilometría, campo de la lingũística que estudia el estilo del lenguaje en obras literarias y en la lengua común, utiliza métodos estadísticos para extraer de los textos marcas estilísticas (estilemas)[26]. Ejemplos de estos son:n-gramas de palabras y caracteres, signos de puntuación, palabras funcionales, riqueza del vocabulario, frecuencia de partes de la oración, errores gramaticales, longitudes de palabras, oraciones y párrafos. Stamatatos en [**?**], clasifica estos estilemas como:

Características Léxicas: Son una manera simple de ver un texto como una secuencia de elementos, donde cada elemento corresponde a una palabra, número o símbolo de puntuación. En los primeros métodos de AA se utilizaron medidas simples, tales como tamaño de frases y tamaño de palabras [3]. Una ventaja significativa de estas medidas es que pueden ser aplicadas a cualquier idioma o corpus.

En la tarea de AA los textos, tambien pueden representarse como una bolsa de palabras, es decir, el texto se considera como un conjunto de palabras, cada una de ellas con una frecuencia de ocurrencia sin tener en cuenta su información contextual [2]. En dicho enfoque las palabras más comunes como son: artículos, preposiciones, pronombres, y otros, han demostrado ser características adecuadas para discriminar entre diversos autores [44].

Uno de los métodos más sencillos y eficaces para definir un conjunto de características léxicas para la AA, es el extraer las palabras más frecuentes dentro del corpus que se está estudiando. De esta forma la atribución del autor al texto en disputa se dará por la cantidad de palabras frecuentes. Dentro de los primeros trabajos sobre AA, grupos de como máximo 100 palabras frecuentes se consideraban

adecuadas para representar el estilo del autor [5].

Sin embargo, a pesar de la simplicidad de este enfoque, en ocasiones son necesarias herramientas adiciones, desde rutinas sencillas como la conversión a minúsculas hasta herramientas más complejas como analizadores lingüísticos [20] o lematizadores [8].

Características de Caracteres: Dentro de esta clasificación, el texto puede ser considerado como una secuencia de caracteres, las cuales pueden incluir caracteres alfabéticos sin hacer distinción de mayúsculas o minúsculas, caracteres numéricos, signos de puntuación, etc. [22], [33]. Dichas características son fáciles de extraer en textos de cualquier lenguaje natural y han demostrado ser útiles para cuantificar el estilo de escritura [15].

El objetivo de este tipo de características es extraer frecuencias de n-gramas a nivel carácter. Por ejemplo, los cuatrigramas del inicio de este párrafo pueden ser: *El_o*, *l_ob*, *_obj* y así sucesivamente. Estas características pueden capturar matices de estilo, incluyendo información léxica, información contextual, el uso de puntuación y mayúsculas. Otra de sus ventajas es que suele ser tolerante al ruido, es decir, textos que contengan errores gramaticales o uso extraño de los signos de puntuación, como suele ocurrir en el correo electrónico, foros en línea, entre otros. En estos errores la representación de n-gramas no se ve afectada de forma significativa. Por ejemplo, las palabras **simplistas** o **simplstas** producirán muchos trigramas comunes de caracteres. Por otro lado, estas palabras se consideran diferentes en una representación basadas en palabras. Pero tenido en cuenta la categorización de estilo basado en el texto, este tipo de errores podrían considerarse como rasgos personales del autor [42]. Al igual que las palabras, los n-gramas de caracteres más frecuentes son las características importantes al realizar el perfil de escritura de los autores.

El procedimiento para la extracción de n-gramas es independiente al lenguaje y no requiere herramientas especiales; sin embargo, la dimensionalidad de esta representación se incrementa de forma considerable en comparación con los enfoques

basados en palabras [11], [12].

Kjell (1994) dio inicio al uso de este enfoque en su investigación sobre los artículos Federalistas, extrayendo secuencias de bigramas y trigramas de caracteres; Forsyth y Holmes (1996) se dieron cuenta que los unigramas y bigramas se desempeñaron de mejor forma que las características léxicas en varias tareas de clasificación de textos incluyendo AA. Así mismo, comparaciones recientes de diferentes características léxicas y de caracteres en el mismo corpus de evaluación [14], reportaron mejores resultados con los n-gramas de caracteres.

Un aspecto fundamental a tener en cuenta en el enfoque de n-gramas es la definición de n, es decir, el tamaño de n. Un n de gran tamaño reflejaría una mejor información léxica y contextual, pero también reflejaría mejor información temática. Además, aumentaría de forma considerable la dimensionalidad de su representación, produciendo cientos de miles de características. Por otra parte una n pequeña, de 2 o 3 caracteres, sería capaz de representar una información parcial de las palabras, pero no sería adecuada para la representación de información contextual. Por tanto la selección de la n óptima depende en gran medida del idioma, ya que ciertos lenguajes como griego y alemán tienden a usar palabras largas en comparación con el lenguaje inglés.

Características de Sintácticas: Un método más elaborado para la representación de texto es emplear información sintáctica. Su idea se basa en que cada autor tiende a usar inconscientemente patrones sintácticos similares. Por lo tanto, la información sintáctica es considerada como la huella digital del autor, fiable para su comparación. Considerando previamente que este tipo de información requiere de herramientas robustas de procesamiento de lenguaje natural capaces de realizar un análisis sintáctico de textos. Esto significa que la extracción de las características sintácticas dependa en gran medida del idioma ya que requiere de la disponibilidad de una herramienta capaz de analizar un lenguaje natural particular. Baayen, van Halteren y Tweedie (1996) fueron los primeros en utilizar las medidas de información sintáctica para la atribución de autoría [1].

Otro intento de explorar información sintáctica fue propuesto por Stamatatos et al. [**?**], utilizando una herramienta de Procesamiento de Lenguaje Natural (PLN) capaz de detectar frases del griego moderno sin restricciones. Por ejemplo, la oración **the black dog eat meat**, se analiza de la siguiente manera: **the_DT black_JJ dog_NN eat_VB meat_NN** donde cada palabra está sucedida por un guión bajo seguido de la etiqueta gramátical asignada, apreciando de esta forma que **the** es etiquetado como **DT** (determinante), **black** como **JJ**(adjetivo), **dog** como **NN** (sustantivo) y **eat** como **VB** (verbo), obteniendo la información sintáctica, en base a estas etiquétas.

Características específicas de la aplicación: Las características léxicas, de caracteres y sintácticas son independientes de la aplicación que se esté utilizando, ya que se pueden extraer del texto, puesto que existen herramientas apropiadas y los recursos necesarios para su medición. De igual forma se pueden definir medidas específicas para representar mejor los matices de estilo en un dominio de texto determinado. Las aplicaciones actuales como correo electrónico y mensajes en foros, revelaron la posibilidad de definir medidas estructurales para cuantificar el estilo del autor, dichas medidas incluyeron: el uso de saludos y despedidas en los mensajes, tipos de firma, uso de sangrías, longitud párrafos, entre otros [22], [33], [32]. Distribución de etiquetas, color de fondo y tamaño de letras [18].

En AA una vez extraídos los estilemas, debe decidirse como utilizarlos para llevar a cabo la identificación de autores, existen diversas aproximaciones, a continuación se mencionan algunas de ellas.

2.2.2. Métodos de Atribución

Para llevar a cabo la tarea de AA, se cuenta con un conjunto de autores candidatos, un conjunto muestras de textos, de autores conocidos que cubren a todos los candidatos, y un conjunto de muestras de autores desconocidos; cada uno de ellos debe ser atribuido a un autor candidato. Algunos enfoques concatenan todos los textos de entrenamiento disponibles por autor en sólo un archivo y realizan una

representación con base en características extraídas del estilo del autor, normalmente llamados, enfoques basados en perfil del autor [34]. Mientras que otros enfoques requieren múltiples muestras de texto de entrenamiento por autor para desarrollar un modelo de atribución exacto. Es decir, cada texto de entrenamiento se representa de forma individual como una instancia independiente de cada autor, llamados comúnmente métodos basados en instancias. Están también los enfoques híbridos que convinan características de enfoques basados en el perfil del autor y métodos basados en instancias [1].

Enfoques Basados en perfil del autor

Consisten en modelar el estilo de escritura del autor, basándose en una cantidad considerable de texto escrito por él mismo. Dicho texto es obtenido de la concatenación de todos los documentos de entrenamiento del autor, ignorando pequeñas diferencias entre ellos y extrayendo características que serán de utilidad para descubrir el estilo de escritura del autor. Cabe destacar que no hay representación por separado de cada texto, sino una sola representación por autor. Por tal motivo se ignoran las diferencias entre los textos del mismo autor. Por otra parte, los estilemas extraídos pueden ser muy diferentes en comparación con cada uno de los textos originales. Una arquitectura típica se mnuestra en la figura 1, donde X_A representa el archivo del autor A, y X_u representa el perfil de un autor desconocido [1].

Este enfoque es un proceso relativamente fácil, el cual comprende la construcción de perfiles, para los autores considerados. Seguido de ello, se construye el perfil para el autor de un texto desconocido, entonces el modelo de AA utiliza una función de distancia que calcula las diferencias entre el perfil del texto desconocido y el perfil de los autores considerados. Asi si *PR(x)* es el perfil para texto X, la distancia entre el perfil del texto X y el perfil del texto *Y, d[PR(x),PR(y)]* entonces el autor para el texto x está dado por:

$$autor(x) = arg \quad min \quad d(PR(x), PR(X_a), a \epsilon A \tag{2.1}$$

Donde A es el conjunto de autores candidatos y X_a es la concatenación de los

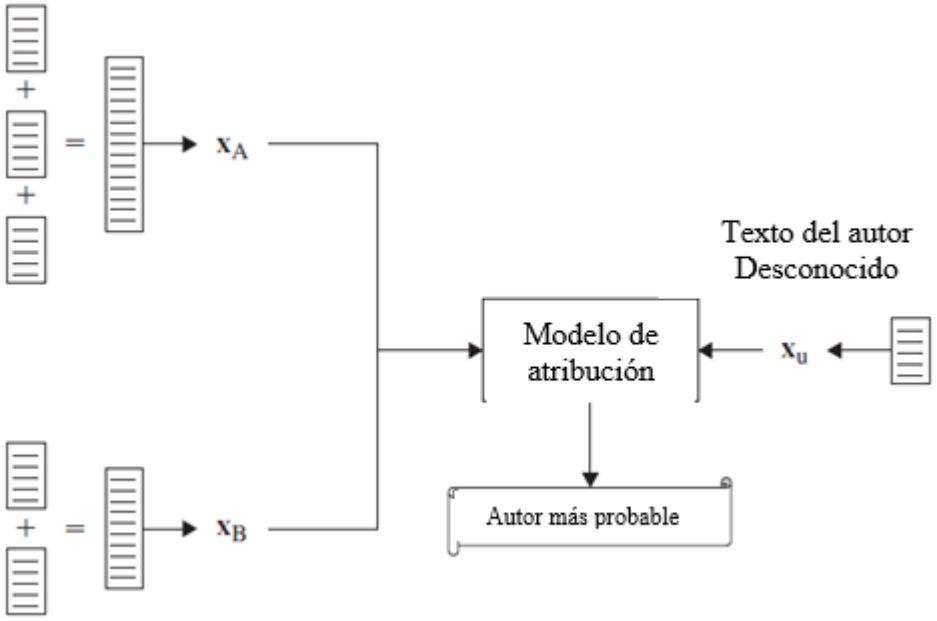

Figura 2.1: Enfoques basados en perfil

textos de entrenamiento para el autor A [1].

Métodos probabilísticos: Estos métodos fueron utilizados desde las primeras aproximaciones para la identificación de los autores, [20]; [34]; [43]; [21]; [35]; [46].

Su objetivo es maximizar la probabilidad *P(x|a)* para un texto x que pertenece al autor candidato a. Entonces el modelo de atribución asignará como autor de un texto a aquel que maximice la similaridad métrica:

$$autor(x) = arg \quad max \quad log_2 \quad \frac{P(x|a)}{p(x|a)}, a \epsilon A \tag{2.2}$$

Donde las probabilidades condicionadas son estimadas por la concatenación x_a de los documentos de entrenamiento para el autor a y la concatenación de los documentos restantes, respectivamente.

N-gramas comunes y variantes. El enfoque de n-gramas comunes (CNG) fue

descrito por Keselj et al [45]. Este método utiliza una representación concreta del perfil del autor. En particular, el perfil *PR(x)* de un texto está compuesto por L los n-gramas más frecuente del texto. Por tanto cualquier medida de distancia es útil para estimar la similitud entre los textos x e y, de acuerdo a la siguiente formula:

$$d(PR(x), PR(y)) = \sum_{g \epsilon P(x) \cup P(y)} \left(\frac{2(f_x(g) - f_y(g))}{f_x(g) + f_y(g)} \right)^2 \tag{2.3}$$

Donde g es un n-grama mientras f_x(g) y f_y(g) son la frecuencia relativa de la ocurrencia de los n-gramas en los textos x e y respectivamente.

La función de distancia CNG funciona bien cuando el corpus de entrenamiento es relativamente equilibrado, pero no en los casos de corpus no balanceados [13].

Enfoques basados en instancias

Estos enfoques consideran cada documento por separado, es decir, cada muestra de texto del corpus está representado por un vector de atributos (x) siguiendo los métodos descritos anteriormente. Esta aproximación es un método clásico de clasificación deonde se define un conjunto de entrenamiento y otro de pruebas, así como un algoritmo de clasificación para generar un modelo de atribución de autoría. Entonces, este modelo es capaz de asignar a un texto de autor desconocido, el autor que con mayor probabilidad lo escribió. La arquitectura típica se muestra en la figura 2.2.

Estos modelos necesitan de varios archivos para la formación de un modelo fiable. Por lo tanto, de acuerdo con los enfoques basados en instancia, en caso de tener sólo un archivo bastante largo, por ejemplo un libro completo, este debe ser segmentado en varias partes, probablemente de la misma longitud. En todos los casos, las muestras de texto deben ser lo suficientemente largas para que las características de los textos puedan representar adecuadamente su estilo. En la literatura se han reportado varias longitudes de textos. Sanderson et al. [20] reportaron trozos de 500 caracteres, Koppel et al. [41] segmentos de 500 palabras. Hirst et al. [38] llevaron a cabo experimentos con bloques de texto de longitud variable (es decir, 200, 500, y

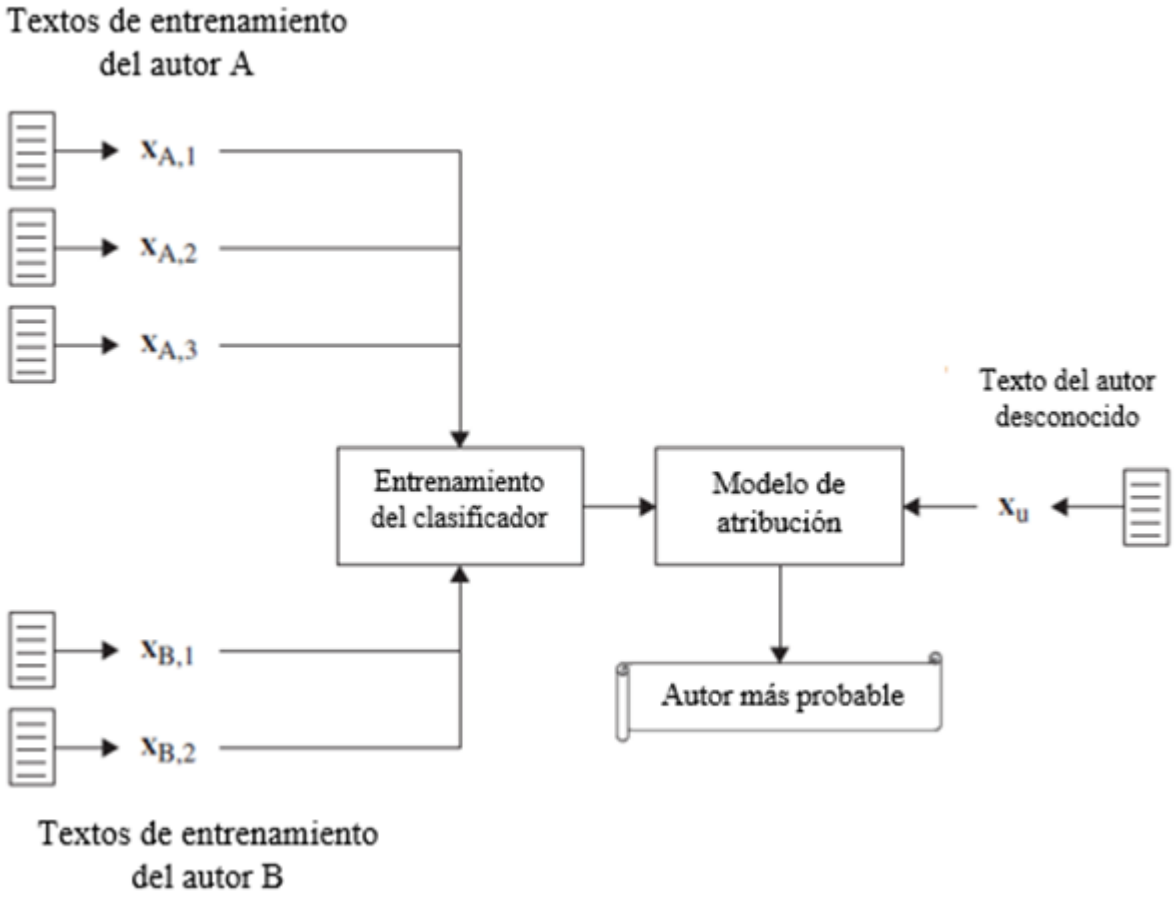

Figura 2.2: Enfoques basados en instancias

1000 palabras) e informaron variación significativa de la precisión en los resultados obtenidos. Por tanto la elección del conjunto de pruebas no es un proceso trivial, afectando directamente el modelo de atribución.

Enfoques híbridos

Este enfoque fue definido por van Halteren [37] el cual toma algunos elementos de los enfoques basados en el perfil del autor y basados en instancias, toma textos por separado, como ocurre en los basados por instancias. Sin embargo los vectores para la representación de textos de cada autor son características promediadas, produciendo un perfil único para cada autor, como ocurre en los enfoques basados en perfil.

2.2.3. Trabajos Actuales

Las investigaciones en AA continúan generando nuevos métodos, como: *Document Author Representation (DAR)*,[30] .

DAR propone hacer uso de la riqueza de los documentos, para encontrar relaciones entre los autores y sus términos, así como la relación entre los autores y sus documentos. Para ello la investigación se dividió en dos fases, a) la construcción de vectores de términos en un espacio de autores, seguido de vectores de documentos en espacio de términos [30] .

Para evaluar la eficacia de DAR, los autores realizan tres experimentos para determinar la precisión. a) DAR usando palabras, b) DAR usando 3-gramas a nivel carácter, c) DAR usando un umbral de frecuencia [30] . Los resultados son comparados con la precisión obtenida para bolsa de palabras

Al evaluar la eficacia de DAR, fueron reportados resultados superiores al 75 % para los experimentos b) y c). Así mismo se observó que aun cuando el conjunto de documentos de autores no está equilibrado, DAR realiza una clasificación adecuada [30] .

Otros trabajos como: *Authorship Attribution Using Probabilistic Context-Free Grammars* planteado por Raghavan et al. [29] Plantea construir una gramática libre de contexto probabilística para cada autor y el uso de esta gramática como un modelo de lenguaje para la clasificación.

Dado el conjunto de entrenamiento de documentos de los autores se construye una gramática libre de contexto probabilística (PCFG) para cada autor en base a los documentos escritos por estos. Una vez construidas, para asignar el autor a un documento del conjunto de prueba, se calculan las probabilidades de generar el documento a partir de las gramáticas generadas para los diferentes autores; el autor del documento será aquel cuya gramática haya obtenido la mayor probabilidad de generarlo.[29]

Durante la investigación se realizaron variaciones a la propuesta inicial, ya que el número de documentos era relativamente bajo, produciendo datos muy dispersos, se añadieron de dos a tres secciones de los corpus Wall Street Journal (WSJ) o corpus Brown de la Penn Treebank para la formación de los conjuntos de entrenamiento. Se realizaron experimentos utilizando máxima entropía y el modelo de n-gramas[29].

De los resultados obtenidos, se observó que el método más exacto para realizar la identificación de autores era PCFG-E (combinación de gramáticas libres de contexto y máxima entropía), obteniendo los mejores resultados en 3 de los 5 corpus utilizados por los Raghavan et al. [29]. Los resultados obtenidos en esta investigación se describen en la Tabla 2.1 Donde se muestran los métodos propuestos, así como clasificaciones tradicionales como Máxima entropía y Naive Bayes, de igual forma los métodos propuestos y algunas de sus combinaciones.

	MaxEnt	NB	Bigram-I	PCFG	PCFG-I	PCFG-E	MaxEnt+Bigram-I
Poetas	56.36	78.18	70.90	78.18	83.63	87.27	76.36
Negocios	83.34	77.78	90.00	77.78	85.56	91.11	92.22
NFL	84.45	86.67	86.67	93.34	80.00	91.11	86.67

Tabla 2.1: Resultados reportados en la bibliografía

Grigori Sidorov[31] propone en su investigación el uso de n-gramas, pero no de la manera tradicional, si no obteniendo los n-gramas en función del orden como se presentan en los árboles sintácticos, es decir, seguir el camino del árbol sintáctico para crear los n-gramas, dando a estos el nombre de n-gramas sintácticos (sn-gramas). Una de sus ventajas principales es que estos sn-gramas se basan en las relaciones sintácticas.

Los sn-gramas permiten ignorar el fenómeno de la lengua ingles que añade un adjetivo antes del sustantivo, minimizando los bigramas en su forma tradicional. Lo mismo sucede en las oraciones subordinadas, y, en general, en cualquier estructura sintáctica. [31].

Se utilizaron sn-gramas con etiquetas de relaciones sintácticas (SR) como elementos de los sn-gramas. Utilizando el analizador de Stanford para determinar las etiquetas de las palabras y para la construcción de los árboles sintácticos.

Con base en los trabajos descritos anteriormente podemos concluir que la tarea de AA no es una tarea reciente, ya que existen investigaciones desde el siglo XIX para el estudio de textos literarios; actualmente esta tarea tiene diversas áreas de aplicación como: detección de plagio de texto o código, identificación de presuntos culpables, criminalística, entre otras. El objetivo de esta tarea es capturar estilemas con los cuales se puede determinar el estilo de escritura del autor. Dentro de las características léxicas se pueden mencionar palabras más frecuentes, tamaño de palabras, tamaño de frases, etc. De igual forma los n-gramas a nivel carácter han demostrado ser de gran utilidad para la tarea de atribución de autoría. Los n-gramas tienen ventajas importantes, entre las cuales podemos mencionar que no hacen distinción entre mayúsculas y minúsculas, suelen ser tolerantes a palabras escritas incorrectamente, desventaja importante de las características léxicas. Así mismo se han utilizado características sintácticas para esta tarea, ya que los autores suelen usar patrones similares, aunque la extracción de estas características es mucho más complicada que las mencionadas anteriormente, además de ser dependientes del lenguaje.

Estudios actuales han demostrado que las características léxicas y n-gramas de caracteres, han generado resultados aceptables. Otras líneas de investigación han demostrado que el uso de características sintácticas puede ser de mucha utilidad para el estudio de esta tarea dando resultados por arriba del 90 % [31]. Investigaciones como la PCFG reportaron resultados superiores al 70 % al utilizar las características sintácticas. Parece entonces que las características sintácticas, pueden apoyar para la creación de nuevos métodos, que generen resultados aceptables y probablemente al combinarse con características léxicas puedan mejorar la precisión de la AA.

De los trabajos analizados, se concluye que factores que impactan directamente en los métodos de clasificación son el tamaño de los archivos de las colecciones, así como que las colecciones sea balanceado o no. De igual forma, se observó que la combinación de diversas características, produce métodos eficaces para la identificación de los autores.

2.3. Medidas de Evaluación

En diversos estudios relacionados con la recuperación de información se han utilizado para la evaluación de los modelos de clasificacion medidas como son presición y recuerdo.

Para definir dichas medidas , se consideran dos clases a y b como se puede observar en la tabla 2.2 [16], donde cada celda representa el número de predicciones positivas y negativas. Así, a + d son los aciertos del sistema , c + b son los errores y la suma de las cuatro celdas (a + b + c + d) equivale al número total de predicciones.

	predicción positiva	predicción negativa	total de predicciones
Clase postiva	a	b	a+b
Clase negativa	c	d	c+d

Tabla 2.2: Aciertos y errores para un sistema de clasificación binario

A continuación se definen las principales métricas utilizadas en la tarea de clasificación, ya que la AA es un caso particular de dicha tarea.

2.3.1. Precisión

La precisión expresa en qué medida el clasificador toma una decisión correcta al ubicar cualquier documento en la clase que le corresponde.

$$precision = \frac{a}{a+c} \tag{2.4}$$

2.3.2. Recuerdo

Indica cuantos de los documentos de pruebas de una clase, son clasificados en la clase correspondiente.

$$recuerdo = \frac{a}{a+b} \tag{2.5}$$

2.3.3. Medida F

Comparar el comportamiento de diferentes clasificadores de textos con dos medidas no es práctico. Para ello es común utilizar la medida F que se define como:

$$F_\beta = \frac{(1+\beta^2) precision * recuerdo}{\beta^2 * precision + recuerdo} \tag{2.6}$$

para β=1, es la medida armónica de la precisión y el recuerdo.

2.3.4. Exactitud

Indica el porcentaje de los documentos que fueron correctamente clasificados.

$$Exactitud = \frac{a+d}{a+b+c+d} \tag{2.7}$$

Capítulo 3

Método propuesto

En este capítulo se propone un método para la tarea de AA; la idea principal es probar que la combinación de características léxicas y sintácticas, utilizando una representación novedosa propuesta por Plate que es la Representación Holográfica Reducida (HRR) puede contribuir a mejorar el desempeõ de la AA [6]. Además buscando optimizar el tiempo de procesamiento, se utiliza un método de reducción de dimensión de espacio vectorial que es la indexación aleatoria propuesta por Magnus Sahlgren [10]. Los resultados obtenidos hasta el momento muestran la viabilidad de la propuesta. A continuación se presentan los conceptos básicos empleados en el método propuesto.

3.1. Categorías Gramaticales

Se denomina gramática a la ciencia que tiene como objetivo de estudio a los componentes de la lengua y sus combinaciones, por tanto se puede definir como el grupo de principios, reglas y preceptos que rigen el empleo de un lenguaje en particular, consta de cuatro niveles, el nivel fonético-fonológico, el nivel sintáctico-morfológico, el nivel léxico-semántico y el nivel pragmático [23].

La fonética: es la rama de la lingüística que estudia la producción y percepción de los sonidos de una lengua. Sus principales ramas son: fonética experimental, fonética articulatoria, fonemática y fonética acústica [40].

La morfología: es la rama de la lingüística que estudia la estructura interna de las palabras para delimitar, definir y clasificar sus unidades, las clases de palabras a las que da lugar (morfología flexiva) y la formación de nuevas palabras (morfología léxica) [40].

De igual forma en este nivel podemos encontrar el concepto de categoría sintáctica el cual se utiliza con diversos sentidos en la literatura y en la lingística. Puede referirse a los conceptos que se expresan mediante los morfemas flexivos (género, número, persona, tiempo, aspecto, etc.), o a las partes de la oración con función sintáctica o sintagmática, en esta última se pueden mencionar ejemplos como: determinante, sustantivo , adjetivo, verbo, adverbio, preposición, conjunción y pronombre, entre otros. [39] .

La sintaxis: es una subdisciplina de la lingüística y parte importante del análisis gramatical que se encarga del estudio de las reglas que gobiernan la combinatoria de constituyentes y la formación de unidades superiores a éstos, como los sintagmas y oraciones [40].

La semántica: La semántica examina el modo en que los significados se atribuyen a las palabras , sus modificaciones a través del tiempo y aún sus cambios por nuevos significados [40].

Por su parte el concepto de léxico encierra varios significados, todos ligados al mundo de la lingüistica. El léxico es el vocabulario de un idioma o de una región (conocidas tambien como unidades léxicas), el diccionario de una lengua o el caudal de modismos. En la gramática se define como categoria o clase léxica a un grupo bien definido de palabras, que tienen la particularidad de hacer referencia a ciertos conceptos, ya sean abstractos o materiales, y que tienen un concepto independientemente de su contexto. Este tipo de palabra puede ser clasificado según su comportamiento a nivel morfolólico o sintáctico [24].

La etimología: Estudia el origen de las palabras, cuándo son incorporadas a un idioma, de qué fuente, y cómo su forma y significado han cambiado [40].

Para la presente investigación se tomaron encuenta las categorias: **léxica y sintáctica.**

Una vez elegidas las categorias a considerar, se buscaron técnicas de reducción de dimensión vectorial, a fin de optimizar el tiempo de ejecución, así como una representación que permitiera combinar los niveles gramaticales seleccionados, llegando a la conclusión de usar técnicas como la que a continuación se presentan.

3.2. Representaciones utlizadas para el método propuesto

En esta sección se describen de manera general los métodos que se utilizarán en esta investigación. La AA es una tarea de clasificación que utiliza diversas características: léxicas, sintácticas o semánticas. La búsqueda para intentan reducir el espacio vectorial de las representaciones textuales utilizadas, a fin de contar con mejores tiempos de procesamiento ha llevado al estudio de diversos métodos. Ejemplos de estos métodos son: la indexación semántica latente (LSI) y recientemente la indexación aleatoria (RI) .

Esta última tiene la ventaja sobre LSI de ser incremental y no necesita de cálculos complicados y tardados como el cálculo de valores singulares, por tal motivo se decidio utilizar RI, misma que se describe a continuación.

3.2.1. Indexación aleatoria

La indexación aleatoria (RI) propuesta por Kanerva et al. [17] consiste de 2 pasos.

1.- Cada contexto (documento o palabra) es asignado en un representación única y generada aleatoriamente llamada vector índice, siendo este de alta dimensión, es

decir, se encuentran en el orden de miles y que consiste en un pequeõ número de 1s y -1s distribuidos al azar, con el resto de los elementos del de vector en 0s [10].

2.- Los vectores de contexto son producidos al recorrer el texto y cada vez que una palabra se produce en un contexto, el vector índice se añade al vector de contexto de la palabra, de esta forma el vector de contexto es la suma de todos los contextos de las palabras encontradas [10].

Algunas de las ventajas de esta metodología son: [10]

- Espacios vectoriales matemáticamente bien definidos y bien entendidos.
- La metodología de creación del espacio de las palabras hace que cierta semántica pueda capturarse.
- Constituyen un enfoque puramente descriptivo para el modelado semántico; no requiere ningún conocimiento lingístico o semántico anterior.

Definida la técnica utilizada para reducir el espacio vectorial, se buscó una manera de combinar la información léxica y sintáctica, tomando la decisón de utilizar la HRR, estas se describe a continuación.

3.2.2. Representaciones Holográficas Reducidas

Tony A. Plate en [6] propone la Representación Holográfica Reducida (HRR) para recuperar analogías. Este trabajo, aunque es la propuesta inicial del empleo de HRRs para representar estructura, no permite evaluar su utilidad para tareas de procesamiento de texto, ya que se platea utilizando sólo seis oraciones. Los HRRs son vectores cuyas entradas siguen una distribución normal *N(0,1/n)*. Plate combina los HRRs utilizando el operador de convolución circular.

La convolución circular ($\otimes$) es un operador asociativo que enlaza dos vectores. x = $(x_o, x_1,,x_{n-1})$ y y = $(y_o,y_1,,y_{n-1})$ para tener un z = $(z_0,z_1,,z_{n-1})$ donde $z = x \otimes y$

se define como:

$$z_i = \sum_{k=0}^{n-1} x_k y_{i-k} \tag{3.1}$$

i = 0 a n-1 (subíndices módulo n)

Con este operador se combino información textual léxica y sintáctica. Para definir las etiquetas sintácticas se utilizó etiquetador de Stanford. Las etiquetas se representaron como HRR y mediante la convolución circular, se combinaron con los vectores índice generados con RI para representar las palabras. Así los vectores resultantes almacenan información sintáctica y léxica de las palabras. Estos vectores se combinaron mediante la suma para representar los documentos.

3.3. Pasos del método propuesta

El método propuesto se ilustra en la figura 3.1.

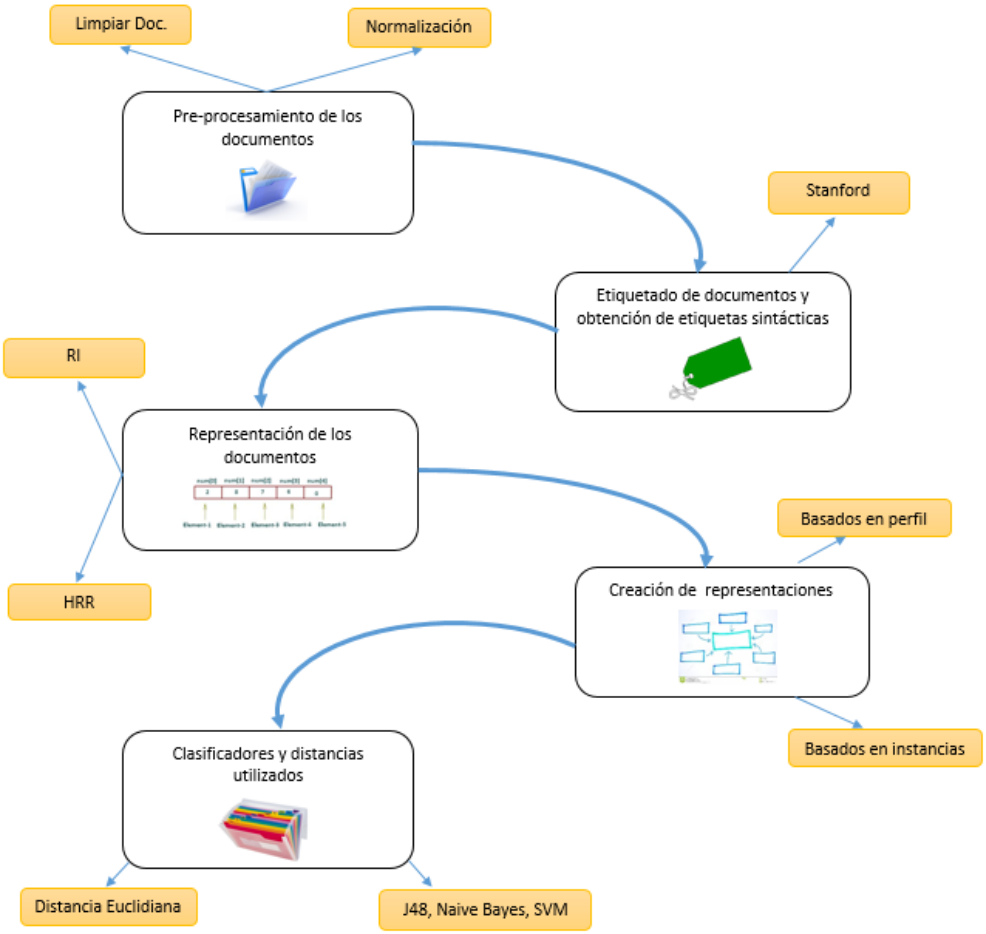

Figura 3.1: Metodo propuesto

A continuación se describen brevemente los pasos anteriores.

3.3.1. Pre-procesamiento de los documentos

Inicialmente se procedío a realizar la limpieza de cada uno de los documentos contenidos en los conjuntos de entrenamiento y pruebas sin excepción alguna, es decir, se eliminaron símbolos no alfanuméricos, así como los valores numéricos.

Ejemplo

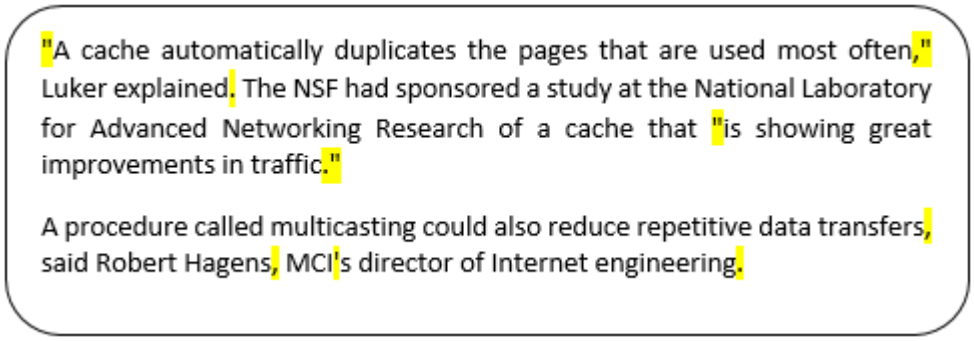

Figura 3.2: Documento sucio

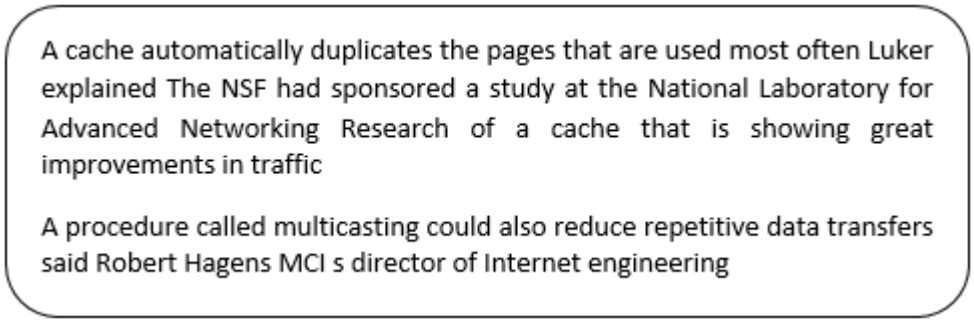

Figura 3.3: Documento limpio

Una vez limpios todos los documentos se procedió a normalizalizarlos, es decir, todas las palabras de los documentos se convertirían a minúsculas.

Ejemplo

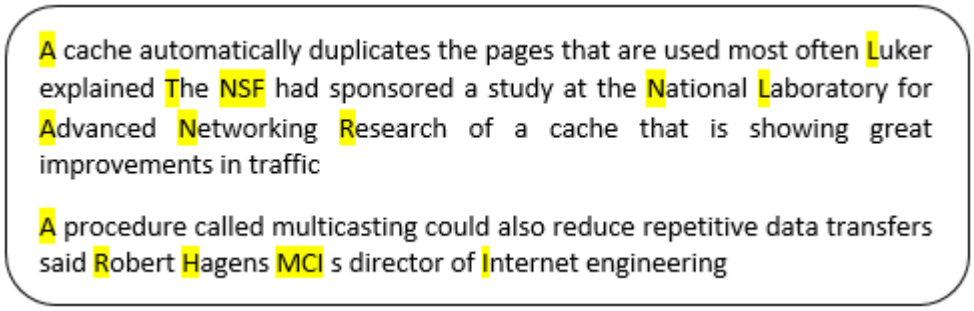

Figura 3.4: Documento no normalizado

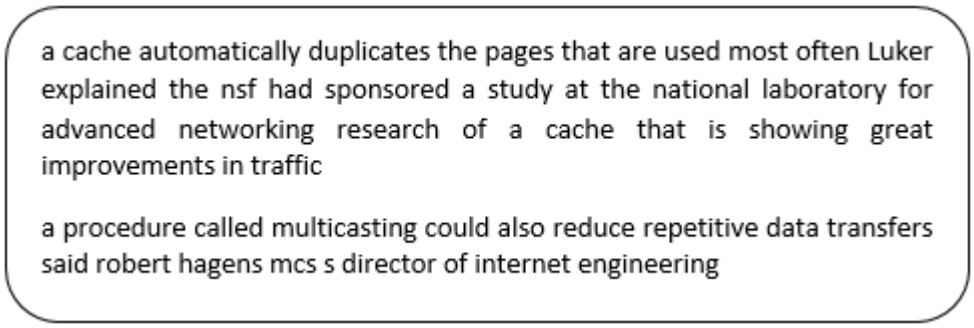

Figura 3.5: Documento normalizado

3.3.2. Etiquetado de los documentos y selección de etiquetas únicas

Preprocesados los documentos, cada uno de los documentos se representó en una línea de un archivo, como se observa en la figura 3.6.

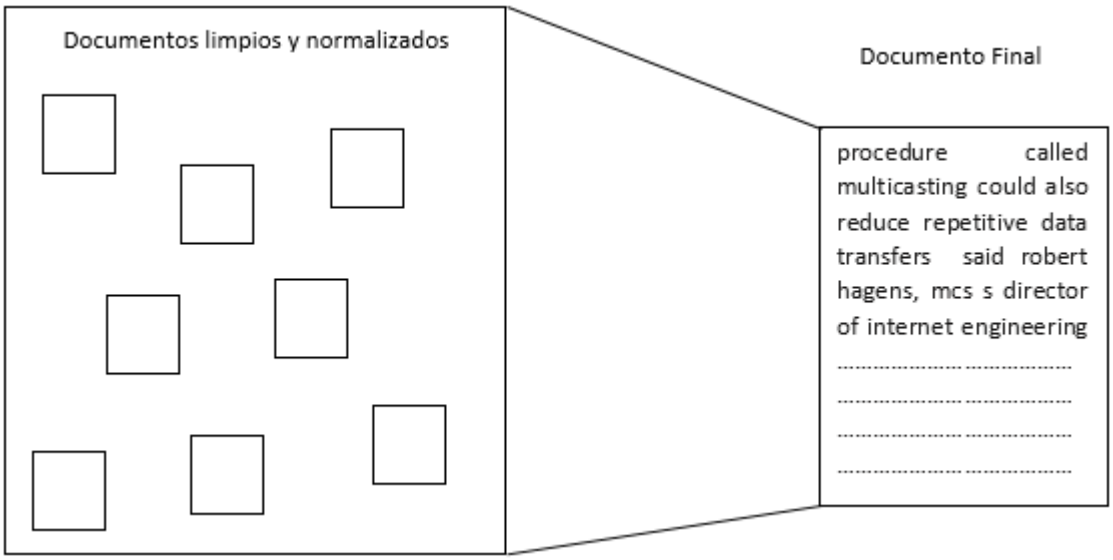

Figura 3.6: Documentos representados en una línea de un archivo

En el archivo generado, utilizando el etiquetador sintáctico Stanford [36], se etiquetaron los documentos como se puede observar en la figura 3.8 .

a cache automatically duplicates the pages that are used most often Luker explained the nsf had sponsored a study at the national laboratory for advanced networking research of a cache that is showing great improvements in traffic

a procedure called multicasting could also reduce repetitive data transfers said robert hagens mcs s director of internet engineering

Figura 3.7: Documento normalizado sin etiquetar

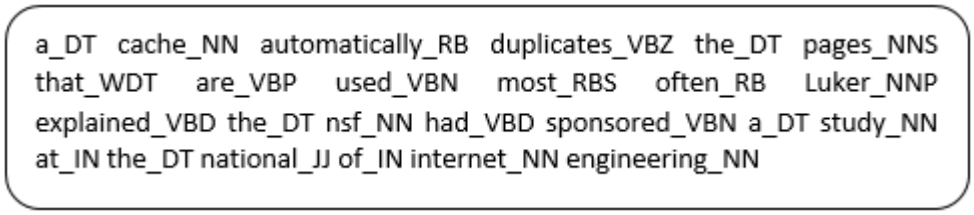

Figura 3.8: Documento normalizado etiquetado

Etiquetados los documentos, se identificaron las etiquetas sintácticas únicas. El número total de etiquetas para los corpus utilizados se presentan en el la tabla 3.1.

Corpus	Etiquetas
Poetas	34
Negocios	34
NFL	34

Tabla 3.1: Etiquetas únicas en los corpus

Se realizó el mismo proceso de etiquetado, para el conjunto de pruebas.

3.3.3. Representación de los documentos

Se utilizó RI para reducir el espacio vectorial, representando todo el vocabulario como vectores de ceros, unos y menos unos. La dimensión del espacio vectorial para todos los experimentos fue de 2048, con un 10 % de 1, 10 % de -1 y el 80 % restante de 0 .

En la Figura 3.9 se ilustra la representación de los documentos. Las etiquetas sintácticas, se representaron como **HRRs** y las palabras como vectores índice (**VI**). Estos se combinaron con la convolución circular. Finalmente los documentos se representaron como la suma de los vectores, que representan a las palabras.

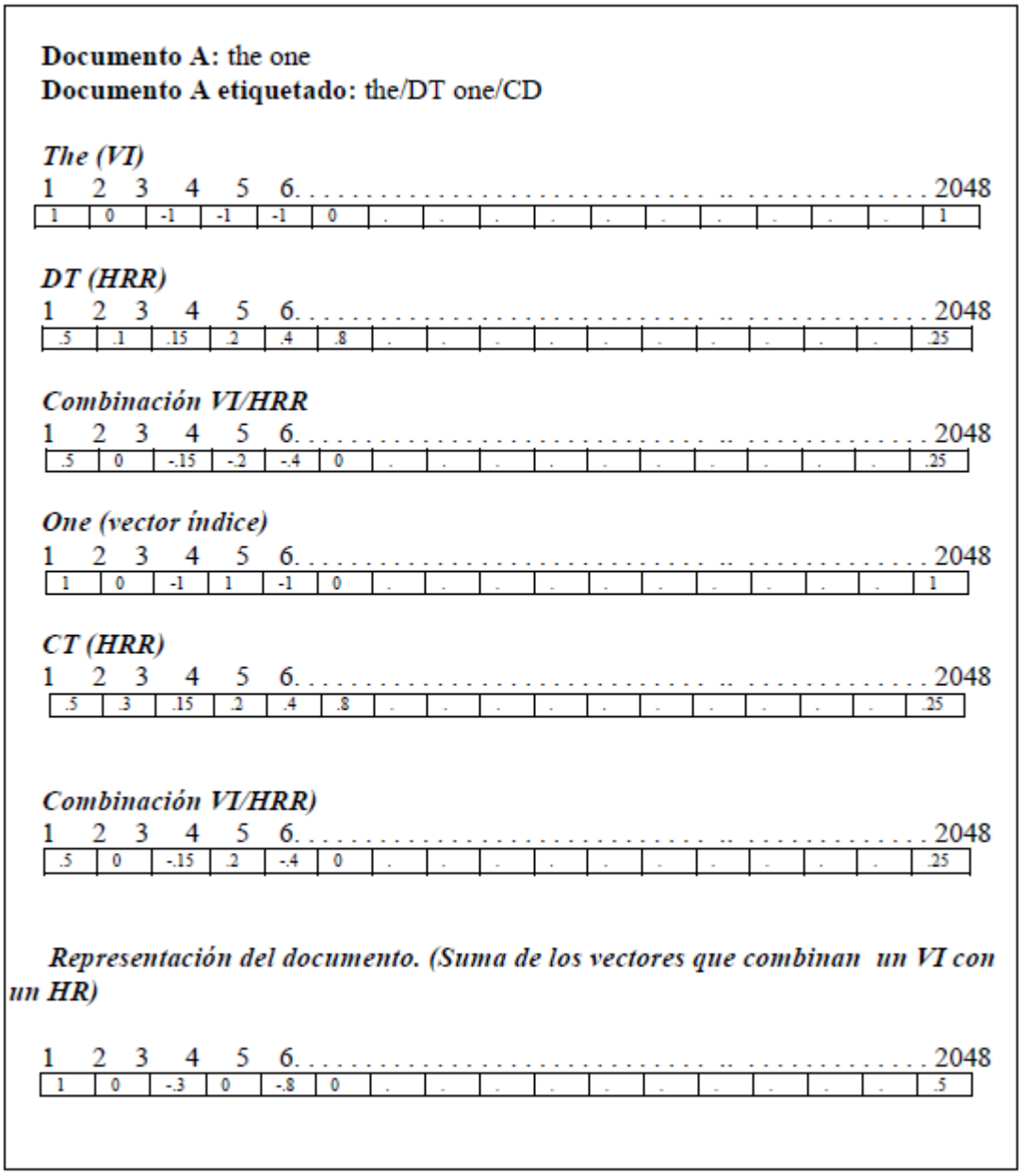

Figura 3.9: Representación de los documentos

3.3.4. Clasificadores y distancias utilizados

Obtenidas las representaciones vectoriales de los documentos tanto para el conjunto de entrenamiento como el de prueba, para evaluar la representación propuesta en AA, utilizando la aproximación basada en instancias, se experimentó con tres clasificadores :

Para el enfoque basado en instancia se experimentó con los siguientes clasificadores:

- ***J48*** : es una implementacin open source en lenguaje de programacin Java del algoritmo C4.5 en la herramienta weka de minera de datos. El algoritmo es usado para generar un árbol de decisión
- ***Naive Bayes***: es un clasificador probabilstico fundamentado en el teorema de Bayes.
- ***SVM***: Son sistemas de aprendizaje que usan un espacio de hipótesis de funciones lineales en un espacio de rasgos de mayor dimensión, entrenadas por un algoritmo proveniente de la teoría de optimización.

Para el enfonque basado en perfil del autor, la semejanza entre un autor desconocido y los perfiles definidos se detrerminó utilizando la distancia euclidiana definida como:

- ***Distancia euclidiana*** : es la distancia "ordinaria" entre dos puntos de un espacio euclídeo, la cual se deduce a partir del teorema de Pitágoras.

Se define formalmente como se presenta en la ecuación 3.2.

$$d_E = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \tag{3.2}$$

Capítulo 4

Corpus y resultados de referencia

Para probar la representación propuesta en AA, se utilizaron 3 cospus: ***Poetas, Negocios y NFL***, con documentos de aproximadamente 300 palabras en promedio. Entre las caracteristicas de estos corpus se puede mencionar que son corpus equilibrados y corpus desequilibrados. En un corpus desequilibrado el número de documentos por autor varia, por ejemplo para el corpus Poetas, hay autores con 17 y otros con 36 o 46 documentos. En contraste en los corpus equilibrados como Negocios y NFL, cada autor contiene el mismo número de documentos. Cabe destacar que los autores hablan del mismo tema en los documenbtos de un mismo corpus. Asi estaremos manejando tres temáticas: poemas, negocios y deportes.

La tabla 4.1 resume la información relevante de los corpus.

	Vocabulario	Etiquetas sintácticas	Documentos entrenamiento	Documentos pruebas	Autores
Poetas	6940	34	146	55	6
Negocios	8492	34	85	90	6
NFL	4982	34	48	45	3

Tabla 4.1: Estadísticas sobre los corpus utilizados en los experimentos.

Capítulo 5

Experimentos y Resultados

En este capítulo se presentan los experimentos y resultados obtenidos con el método propuesto. El conjunto de datos utilizados para las evaluaciones se presentó en el capitulo 4. El método seguido se presentó en la sección 3.3. Los experimentos realizados serán presentados en las secciones 5.1 y 5.2. Posteriormente en la sección 5.3 se mostrarán a detalle los resultados obtenidos de la evaluación del método propuesto.

Se experimentó con enfoques basados en perfil del autor (sección:2.2.2) y con enfoques basados en instancias (sección:2.2.2), cabe destacar que para estos experimentos las colecciones fueron normalizadas por número de documentos y por el tamaño del vocabulario.

5.1. Experimentos con Unigramas palabra⊗etiqueta

Inicialmente se decidió combinar información léxica con información sintáctica para ello se crearon unigramas de palabras con sus etiquetas sintácticas, por ejemplo: **Pedro_NP**, donde el signo (_) indica del lado izquiero la palabra y del lado derecho la etiqueta, para combinar dicha infomación se utilizaron las representaciones descritas en la sección 3.2.

5.2. Experimentos con Bigramas

De igual forma utilizando el enfoque **basado en perfil** y los corpus **Business, NFL, Poetry**, se decidió experimentar con diversas combinaciones de bigramas, esto con el objetivo de observar si combinaciones como: palabras con sus etiquetas, únicamente palabras o etiquetas, podrían mejorar los resultados obtenidos en los experimentos realizados con unigramas.

Experimentos de la forma: PalabraPalabra

Primero se decidió unir bigramas de palabras, por ejemplo en la oración: **El pueblo es bonito**, se formaron los bigramas: **elpueblo puebloes esbonito**, por cada uno de estos bigramas, se creó una representación vectorial de 1's y 0's, seguido de ello se suman los vectores correspondientes a los bigramas encontrados en el documento, generando así una representación vectorial por cada documento de los autores. Las representaciones vectoriales de los documentos se sumaron para obtener el perfil de cada autor.

Experimentos de la forma: palabra $\otimes$ palabra

En el siguiente experimento se calculó la convolución circular de dos palabras, por ejemplo en **El pueblo es bonito**, se consideraron las siguientes combinaciones: **el$\otimes$pueblo pueblo$\otimes$es es$\otimes$bonito**, se creó una representación vectorial por cada convolución y al igual que en el experimento anterior se realizó la suma de los vectores para obtener la representación de cada documento y con la suma de estos poder generar el perfil de cada autor.

Experimentos de la forma: palabra $\otimes$ etiqueta \palabra $\otimes$ etiqueta

Continuando con los experimentos, se formaron bigramas haciendo uso de dos palabras con sus respectivas etiquetas, por ejemplo: **of_in\talk_vb**,donde el signo (\) señala la unión de los unigramas y el simbolo (_) ayuda a identificar del lado

izquierdo la palabra y del lado derecho su etiqueta, Así se generó una representación vectorial calculando la convolución circular de una palabra y su etiqueta (**palabra** $\otimes$ **etiqueta**), después se suman los vectores resultantes, se utilizó el mismo procedimiento que en los experimentos anteriores para generar un único vector por cada documento y en base a estas representaciones generar el perfil de cada autor.

Experimentos de la forma: Etiqueta $\otimes$ Etiqueta

De igual forma se experimentó con etiquetas sintácticas, generando a través del operador de convolución circular, una representación vectorial por cada bigrama, por ejemplo: **in** $\otimes$ **vb**, a continuación se formó una única representación vectorial por cada documento a través de los vectores correspondientes a los bigramas dentro del mismo y con dichas representacione se generó el perfil de cada autor.

5.3. Resultados

A continuación se presentan los resultados obtenidos con los experimentos descritos en la sección anterior.

5.3.1. Resultados de los enfoques basados en instancias

En la tabla 5.1 se muestran los resultados obtenidos empleando el enfoque basado en instancias, como puede observarse son poco favorables y en ocasiones como se aprecia, no se logra identificar de forma correcta ninguno de los documentos que pertenecen a un autor específico, como en el caso del autor 3 y el autor 6.

Debido a los resultados obtenidos se dejó de experimentar con las aproximaciones basadas en instancias.

	Precisión	Recuerdo	Medida F
Autor 1	0.40	**0.90**	0.56
Autor 2	0.28	0.40	0.33
Autor 3	0.00	0.00	0.00
Autor 4	**0.66**	0.40	0.50
Autor 5	0.54	0.60	**0.57**
Autor 6	0.00	0.00	0.00

Tabla 5.1: Corpus poetas, utilizando: J48.

5.3.2. Resultados de los enfoques basados en perfil con unigramas

En las tablas 5.2, 5.3 y 5.4, se muestran los resultados obtenidos para tres corpus empleando unigramas y el enfoque basado en perfil del autor, como puede observarse en la mayoría de los casos los resultados obtenidos son favorables.

Corpus Poetas

	Precisión	Recuerdo	Medida F	Exactitud
Autor 1	0.70	0.70	0.70	**0.89**
Autor 2	0.50	0.40	0.44	**0.81**
Autor 3	0.55	0.50	0.52	**0.83**
Autor 4	0.36	0.40	0.38	**0.76**
Autor 5	0.45	0.50	0.47	**0.80**
Autor 6	0.33	0.40	0.36	**0.87**

Tabla 5.2: Métricas obtenidas para el corpus Poetas

En la tabla 5.2, se aprecia claramente que los resultados obtenidos en cuanto a exactitud son favorables no así considerando la medida F que nos permite evaluar tanto la precisión como el recuerdo.

Corpus NFL Ahora en la tabla 5.3, de igual forma se observa que para el corpus NFL los resultados en exactitud son buenos, y la medida F mejora con respecto al corpus anterior.

	Precisión	Recuerdo	Medida F	Exactitud
Autor 1	0.88	0.90	0.93	**0.95**
Autor 2	0.66	0.80	0.72	**0.80**
Autor 3	**0.90**	0.60	0.72	0.84

Tabla 5.3: Métricas obtenidas para el corpus NFL

Corpus Negocios

	Precisión	Recuerdo	Medida F	Exactitud
Autor 1	0.86	0.86	0.86	**0.95**
Autor 2	0.68	0.86	0.76	**0.91**
Autor 3	0.81	0.86	0.83	**0.94**
Autor 4	0.78	0.73	0.75	**0.92**
Autor 5	0.90	0.66	0.76	**0.93**
Autor 6	0.86	0.86	0.86	**0.95**

Tabla 5.4: Métricas obtenidas para el corpus Negocios

Los resultados obtenidos en el corpus negocios, como se puede apreciar son en su mayoría mejores que los resultados obtenidos en los corpus anteriores. Estos resultados probablemente se debe a que los corpus de entrenamiento y pruebas están balanceados, es decir, contaban con la misma cantidad de documentos para los autores.

En la tabla 5.5 se presentan los promedios obtenidos del enfoque basado en perfil con unigramas, en las secciones subsecuentes se presentan únicamente estos promedios y no los resultados autor por autor.

	Precisión	Recuerdo	Medida F	Exactitud
Negocios	0.86	0.86	0.86	**0.93**
Poetas	0.48	0.48	0.47	**0.82**
NFL	0.81	0.80	0.80	**0.86**

Tabla 5.5: Promedios obtenidos de unigramas

5.3.3. Resultados de los enfoques basados en perfil con bigramas

En las tablas: 5.6, 5.7, 5.8, 5.9 se muiestran los resultados obtenidos con las diversas combinaciones de bigramas descritas en 5.2.

Cabe destacar que los resultados que a continuación se presentan, se obtuvieron promediando el resultado de cinco ejecuciones de cada experimento, debido a que el método propuesto arroja un resultado diferente en cada ejecución, dado que las representaciones vectoriales se forman con valores aleatorios de 1's, 0's o distribución normal, de tal forma que calculando el promedio podemos tener un resultados más preciso que los obtenidos con una única ejecución del método propuesto.

Resultados de bigramas de palabras juntas, de la forma: PalabraPalabra

	Precisión	Recuerdo	Medida F	Exactitud
Negocios	0.79	0.73	0.73	**0.91**
NFL	**0.87**	**0.80**	**0.80**	0.86
Poetas	0.42	0.48	0.47	0.82

Tabla 5.6: Resultados de bigramas formados con palabras continuas

Podemos apreciar que los resultados arrojados por este experimento se comportan de forma similar a los presentados con anterioridad. Para este experimento el corpus que arrojó mejores resultados en cuanto a exactitud fue el corpus Negocios

el cuál a dado los mejores resultados, en la mayoria de los experimentos presentados hasta el momento.

Resultados de convolución de bigramas de la forma Palabra ⊗ Etiqueta

	Precisión	Recuerdo	Medida F	Exactitud
Negocios	0.86	0.83	0.83	**0.94**
NFL	**0.87**	**0.83**	**0.83**	0.89
Poetas	0.50	0.49	0.46	0.82

Tabla 5.7: Resultados de convolucion de palabras

En las tabla: 5.7, podemos apreciar claramente que el corpus Negocios obtiene el mejor resultado en cuanto a exactitud, mientras que para las otras tres medidas es el corpus NFL, el que obtiene los mejores resultados, Sin embargho la medida de mayor importancia para nuestros experimentos, es la exactitud, ya que con ella se compararán los resultados obtenidos con los reportados en la bibliografía, pues ésta únicamente presenta resultados con dicha métrica.

Resultados de bigramas con convolución de palabras y etiquetas de la forma: Palabra⊗ Etiqueta + Palabra⊗Etiqueta

	Precisión	Recuerdo	Medida F	Exactitud
Negocios	**0.82**	**0.81**	**0.81**	**0.93**
NFL	0.79	0.77	0.76	0.83
Poetas	0.47	0.44	0.42	0.80

Tabla 5.8: Resultados de convolucion de palabras con etiquetas

En los resultados del experimento presentado anteriormente podemos observar claramente que para el corpus Negocios se obtienen los mejores resultados en todas

las métricas consideradas.

Resultados convolución de etiquetas

	Precisión	Recuerdo	Medida F	Exactitud
Negocios	**0.62**	0.61	**0.60**	**0.87**
NFL	0.43	**0.66**	0.52	0.67
Poetas	0.37	0.45	0.34	0.77

Tabla 5.9: Resultados convolución de etiquetas

Al igual que los experimentos anteriores, el corpus Negocios genera mejores resultados en tres de las cuatro métricas tomadas en consideración y sobre todo en la métrica de exactitud, ya que se encuentra dos decimas arriba de los demás corpus.

5.4. Evaluación de resultados

Se compararon los resultados obtenidos con los resultados reportados en el artículo **Authorship Attribution Using Probabilistic Context-Free Grammars** [29]. Debe ternerse en cuenta que los corpus utilizados en este artículo fueron complementados con secciones de Treebank. De igual forma se compararon los resultados obtenidos en bigramas con los artículos: **A Weighted Profile Intersection Measure for Profile-Based Authorship Attribution** [27] y **The Use of Orthogonal Similarity Relations in the Prediction of Authorship** [28].En estos artículos se utilizan los corpus Negocios, NLF y Poetas, y se reporta la exactitud obtenida en sus experimentos.

La comparación de nuestros resultados con los reportados en [29] se presenta en las tablas 5.10, 5.11 y 5.12, donde:

- **MaxEnt** y **NB** corresponden a los clasificadores de Máxima Entropía y el clasificador Naive Bayes.

- **Bigram-I** se refiere al modelo de lenguaje de bigramas con suavizado.
- **PCFG** es el método propuesto en [29] que utiliza gramática libre de contexto probabilística para modelar el estilo de cada autor.
- **PCFG-I** corresponde al modelo mencionado anteriormente con interpolación.
- **PCFG-E**, corresponde a la combinación de máxima entropía y PCFG.
- **MaxEnt + Bigram-I**, corresponde a la combinación del clasificador de máxima entropía y bigramas con interpolación.
- La última columna corresponde a la representación empleada en el método propuesto.

Para dar una idea clarade los resultados se utilizan los colores: **Azul**,**Rojo** y **Negro**, donde el color Azul indica el resultado obtenido del método propuesto, el color Rojo índica aquellas propuestas cuyos resultados obtenidos no son superados por nuestra propuesta, en contraste el color Negro indica aquellos resultados propuestos en la literatura los cuales son superados.

5.4.1. Comparando Unigramas

Los experimentos presentados a continuación son comparados como se indicó anteriormente con el artículo: **Authorship Attribution Using Probabilistic Context-Free Grammars** [29]. Los resultados de las métricas se presentran multiplicados por 100.

5.4.1.1. Corpus Poetas

Se puede observar en la tabla 5.10 que los HRR superan al metodo PCFG. Los HRR son superados unicamente al combinar dos de los métodos propuestos en [29].

	Artículo							Propuesta	
	PCFG	PCFG−I	PCFG−E	FLF	RMMF + FLF	MSMF + FLF	CNG − SPI	CMG − WPI	HRR
Poetas	78.18	83.63	87.27	64.27	63.05	78.29	78.18	85.45	82.0

Tabla 5.10: Comparando con el corpus Poetas

5.4.1.2. Corpus NFL

En la tabla 5.11, se puede observar que los HRR son superados por el método propuesto en [29], así como por una de las combianaciones que en el mismo se reporta.

	Artículo							Propuesta	
PCFG	PCFG−I	PCFG−E	FLF	RMMF + FLF	MSMF + FLF	CNG − SPI	CMG − WPI	HRR	
NFL	93.34	80	91.11	89.30	88.35	92.75	91.11	93.34	86.30

Tabla 5.11: Comparando con el corpus NFL

5.4.1.3. Corpus Negocios

En la tabla 5.12, podemos observar que los HRR superan al método PCFG propuesto en [29] así como a cualquier combinacion de mas de uno de los metodos reportados. El corpus negocios es el mas grande y equilibrado de los corpus con los cuales se experimentó.s

	Artículo							Propuesta	
	PCFG	PCFG−I	PCFG−E	FLF	RMMF + FLF	MSMF + FLF	CNG − SPI	CMG − WPI	HRR
Negocios	77.78	85.56	91.11	86.66	83.05	86.29	77.78	80.00	93.30

Tabla 5.12: Comparando con el corpus Negocios

5.4.2. Comparando Bigramas palabras juntas

En las tablas 5.13, 5.14, 5.15 se comparan los resultados obtenidos con lo reportado en [27], [28], [29].

5.4.2.1. Corpus Negocios

Esta forma de bigramas para este corpus supera en su totalidad los resultados de las aproximaciones reportadas en en la literatura.

	Artículo								Propuesta
	PCFG	PCFG−I	PCFG−E	FLF	RMMF + FLF	MSMF + FLF	CNG − SPI	CMG − WPI	HRR
Negocios	77.78	85.56	91.11	86.66	83.05	86.29	77.78	80.00	91.11

Tabla 5.13: Comparando con el corpus Negocios

5.4.2.2. Corpus NFL

Se puede observar que el método propuesto supera únicamente un mtodo de los reportados en la bibliografía.

	Artículo								Propuesta
	PCFG	PCFG−I	PCFG−E	FLF	RMMF + FLF	MSMF + FLF	CNG − SPI	CMG − WPI	HRR
NFL	93.34	80	91.11	89.30	88.35	92.75	91.11	93.34	86.67

Tabla 5.14: Comparando con el corpus NFL

5.4.2.3. Corpus Poetas

Para el corpus Poetas, se observa que el método propuesto supera a cinco de las técnicas reportandas en la bibliografía.

	Artículo								Propuesta
	PCFG	PCFG−I	PCFG−E	FLF	RMMF + FLF	MSMF + FLF	CNG − SPI	CMG − WPI	HRR
Poetas	78.18	83.63	87.27	64.27	63.05	78.29	78.18	85.45	81.2

Tabla 5.15: Comparando con el corpus Poetas

5.4.3. Comparando convolución bigramas de palabras de la forma Palabra $\otimes$ Etiqueta

En las tablas 5.16, 5.17, 5.18 comparamos los resultados obtenidos con los reportados en [27],[28],[29].

5.4.3.1. Corpus Negocios

En la tabla 5.16 se aprecia lo siguiente: El método propuesto supera las diversas técnicas reportadas.

	Artículo								Propuesta
	PCFG	PCFG−I	PCFG−E	FLF	RMMF + FLF	MSMF + FLF	CNG − SPI	CMG − WPI	HRR
Negocios	77.78	85.56	91.11	86.66	83.05	86.29	77.78	80.00	94.83

Tabla 5.16: Comparando con el corpus Negocios

5.4.3.2. Corpus NFL

Para este corpus se aprecia que con HRR únicamente se superan dos de las aproximaciones reportadas en la bibliografía.

	Artículo								Propuesta
	PCFG	PCFG−I	PCFG−E	FLF	RMMF + FLF	MSMF + FLF	CNG − SPI	CMG − WPI	HRR
NFL	93.34	80	91.11	89.30	88.35	92.75	91.11	93.34	89.1

Tabla 5.17: Comparando con el corpus NFL

5.4.3.3. Corpus Poetas

En el corpus Poetas, las HRR's son superados por tres métodos propuestos en [27], [28], [29].

	Artículo								Propuesta
	PCFG	PCFG−I	PCFG−E	FLF	RMMF + FLF	MSMF + FLF	CNG − SPI	CMG − WPI	HRR
Poetas	78.18	83.63	87.27	64.27	63.05	78.29	78.18	85.4	82.1

Tabla 5.18: Comparando con el corpus Poetas

5.4.4. Comparando convolución bigramas de palabras con sus etiquetas: Palabra $\otimes$ Etiqueta + Palabra $\otimes$ Etiqueta

En las tablas 5.19, 5.20, 5.21 se da a conocer el resultado obtenido con esta construcción de bigramas; así como los resultados reportados en [27], [28], [29].

5.4.4.1. Corpus Negocios

En la tabla 5.19 se aprecia que con las HRR's, se superan los diversos métodos propuestos en la literatura.

	Artículo								Propuesta
	PCFG	PCFG−I	PCFG−E	FLF	RMMF + FLF	MSMF + FLF	CNG − SPI	CMG − WPI	HRR
Negocios	77.78	85.56	91.11	86.66	83.05	86.29	77.78	80.00	93.7

Tabla 5.19: Comparando con el corpus Negocios

5.4.4.2. Corpus NFL

Para esta representación de bigramas las HRR's únicamente superan en dos ocaciones a los métodos ya reportados.

	Artículo								Propuesta
	PCFG	PCFG−I	PCFG−E	FLF	RMMF + FLF	MSMF + FLF	CNG − SPI	CMG − WPI	HRR
NFL	93.34	80	91.11	89.30	88.35	92.75	91.11	93.34	83.0

Tabla 5.20: Comparando con el corpus NFL

5.4.4.3. Corpus Poetas

En contraste, para el corpus Poetas, con esta construcción de bigramas, se supera a cinco de las aproximaciones reportadas, siendo superada únicamente por tres de ellas.

	Artículo								Propuesta
	PCFG	PCFG−I	PCFG−E	FLF	RMMF + FLF	MSMF + FLF	CNG − SPI	CMG − WPI	HRR
Poetas	78.18	83.63	87.27	64.27	63.05	78.29	78.18	85.4	80.9

Tabla 5.21: Comparando con el corpus Poetas

5.4.5. Comparando convolución bigramas de etiquetas: Etiqueta ⊗ Etiqueta

Como se mencionó en 5.2, se creó una representación únicamente de etiquetas sintácticas.

En las tablas 5.22, 5.23, 5.24 se compara dicho experimento, con lo reportado en [27],[28],[29]

5.4.5.1. Corpus Negocios

En la tabla 5.22 se puede apreciar claramente que con HRR se superan las aproximaciones propuestas en los artículos mencionados con anterioridad.

	Artículo								Propuesta
	PCFG	PCFG−I	PCFG−E	FLF	RMMF + FLF	MSMF + FLF	CNG − SPI	CMG − WPI	HRR
Negocios	77.78	85.56	91.11	86.66	83.05	86.29	77.78	80.00	92.2

Tabla 5.22: Comparando con el corpus Negocios

5.4.5.2. Corpus NFL

Para el corpus NFL podemos ver una diferencia en comparación al corpus Negocios ya que el método propuesto para dicho corpus sólamente supera uno de los métodos propuestos y es superado en su mayoría.

	Artículo								Propuesta
	PCFG	PCFG−I	PCFG−E	FLF	RMMF + FLF	MSMF + FLF	CNG − SPI	CMG − WPI	HRR
NFL	93.34	80	91.11	89.30	88.35	92.75	91.11	93.34	86.6

Tabla 5.23: Comparando con el corpus NFL

5.4.5.3. Corpus Poetas

Para el corpus Poetas únicamente se supera en dos de los ocho métodos propuestos en los artículos tomados como referencia.

	Artículo								Propuesta
	PCFG	PCFG−I	PCFG−E	FLF	RMMF + FLF	MSMF + FLF	CNG − SPI	CMG − WPI	HRR
Poetas	78.18	83.63	87.27	64.27	63.05	78.29	78.18	85.4	76.3

Tabla 5.24: Comparando con el corpus Poetas

En la tabla 5.25 se muestra los resultados para cada uno de los corpus.

Puede observarse como para el corpus Negocios, que tiene el vocabulario más rico de los corpus considerados, cualquiera de nuestra propuestas aún las que no implican combinación de información léxica y sintáctica sino únicamente la utilización de RI para reducir la dimensión vectorial o RI y convolución circular para formar bigramas, superan a las propuestas reportadas en la bibliografía. Puede observarse como para este corpus los resultados combinando información léxica y sintáctica son mejores con bigrámas que con unigramas. Por lo tanto los HRRs para este corpus, combinan de manera adecuada la información léxica y sintáctica, mejoran los resultados. Para Poetas la aproximación de bigramas simples y unigramas combinando información léxica y sintáctica reportan los mejores resultados. Para NFL que es un corpus con un vocabulario reducido, no equilibrado y poco documentos los mejores resultados se obtuvieron con bigramas de palabras.

	Artículo								Propuesta
	PCFG	PCFG−I	PCFG−E	FLF	RMMF + FLF	MSMF + FLF	CNG − SPI	CMG − WPI	HRR
Unigramas									
Negocios	77.78	85.56	91.11	86.66	83.05	86.29	77.78	80.00	93.30
NFL	93.34	80	91.11	89.30	88.35	92.75	91.11	93.34	86.30
Poetas	78.18	83.63	87.27	64.27	63.05	78.29	78.18	85.45	82.0
PalabraPalabra									
Negocios	77.78	85.56	91.11	86.66	83.05	86.29	77.78	80.00	91.11
NFL	93.34	80	91.11	89.30	88.35	92.75	91.11	93.34	86.67
Poetas	78.18	83.63	87.27	64.27	63.05	78.29	78.18	85.45	81.2
Palabra	$\otimes$	*Palabra*							
Negocios	77.78	85.56	91.11	86.66	83.05	86.29	77.78	80.00	94.83
NFL	93.34	80	91.11	89.30	88.35	92.75	91.11	93.34	89.1
Poetas	78.18	83.63	87.27	64.27	63.05	78.29	78.18	85.4	82.1
Palabra	$\otimes$	*Etiqueta*	$\setminus$	*Palabra*	$\otimes$	*Etiqueta*			
Negocios	77.78	85.56	91.11	86.66	83.05	86.29	77.78	80.00	93.7
NFL	93.34	80	91.11	89.30	88.35	92.75	91.11	93.34	83.0
Poetas	78.18	83.63	87.27	64.27	63.05	78.29	78.18	85.4	80.9
etiqueta	$\otimes$	*etiqueta*							
Negocios	77.78	85.56	91.11	86.66	83.05	86.29	77.78	80.00	92.2
NFL	93.34	80	91.11	89.30	88.35	92.75	91.11	93.34	86.6
Poetas	78.18	83.63	87.27	64.27	63.05	78.29	78.18	85.4	76.3

Tabla 5.25: Resultados generales

Capítulo 6

Conclusiones y trabajo a futuro

Con base en los resultados obtenidos de los experimentos realizados a lo largo de la investigación se puede decir lo siguiente:

Con la combinación de información léxico−sintáctica se obtuvieron resultados aceptables para la tarea de atribución de autoría, los resultados generados por los HRRs son equiparables a los reportados en la bibliografía. Por otra parte pudo observarse que el uso de RI ayuda a mejorar el tiempo de ejecución de la tarea de AA, disminuyendo el tiempo de procesamiento ya que no importa que tan grande sea el vocabulario el tama no de los vectores siempre será de 2048.

Así mismo se pudo observar que los enfoques basados en instancias combinando información léxica y sintáctica, con el uso de los HRR a través del operador de convolución circular, produce resultados poco favorables. En corpus mayores de 50 documentos y equilibrados los HRRs producen resultados adecuados.

Para el corpus NFL los resultados no fueron satisfactorios pienso que esto se debe a que es un corpus desequilibrado y pequeo para generar un modelo útil.

Para el corpus Poetas los resultados obtenidos con los HRRs superan a cinco de las ocho aproximaciones reportadas en la bibliograía, siendo superados únicamente cuando se combinan más de una aproximación.

La propuesta de utilizar HRR para combinar información léxica y sintáctica en los experimentos realizados con el corpus negocios superó a todas las aproximaciones

reportadas en la bibliografía. Este corpus es equilibrado, es decir, que contiene el mismo número de documento para cada uno de sus autores.

Como trabajo futuro se planea experimentar con corpus de mayor dimensión y variar parámetros de RI en cuanto a la definición del tamaño vectorial y concentración de elementos diferentes de cero.

Bibliografía

[1] Efstathios Stamatatos. (March 2009). *A survey of modern authorship attribution methods. The American Society for Information Science and Technology , 60, pp. 538-556*, .

[2] Sebastiani F. . (March 2002). *Machine learning in automated text categorization. ACM Computing Surveys, 34, pp. 1-47*, .

[3] T. C. Mendenhall. (March 1887). *The characteristic curves of composition. Science , 11, pp. 237-246.*

[4] Alfredo Rodríguez López Vázquez. (1987). *Aportaciones críticas a la autoría de El burlador de Sevilla, 40, pp. 5-44.*, .

[5] Burrows J.F. (1987). *Word-Patterns and Story-Shapes: The Statistical Analysis of Narrative Style. Literary and Linguistic Computing, 2,pp. 61-70.*, .

[6] Tony Plate. (1991. *Holographic Reduced Representations: Convolution Algebra for Compositional Distributed Representations. Internacional join conference on artificial intelligence .*

[7] Holmes D.I. (1998). *The evolution of stylometry in humanities scholarship. Literary and Linguistic Computing. Arts - Humanities, 13, pp. 111-117.*

[8] Gamon M. (2004). *Linguistic correlates of style: Authorship classification with deep linguistic analysis features. In Proceedings of the 20th International Conference on Computational Linguistics, pp. 611-617.*

[9] Juan María Apellániz. (2005). *La metodología de la hipótesis de atribución de autor aplicada a las figuras grabadas en los omoplatos de El Castillo. Antropología-Arkeología, 57, pp. 207-216.*, .

[10] Magnus Sahlgren. (2005). *An introduction to random indexing . In Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering.*, .

[11] Stamatatos E. . (2006a). *Authorship attribution based on feature set subspacing ensembles. International Journal on Artificial Intelligence Tools , 20, pp. 1-16.*

[12] Stamatatos E. (2006b). *Ensemble-based author identification using character n-grams.. Workshop on Text-based Information Retrieval, In Proc. 3rd. pp. 41-46.*

[13] Efstathios Stamatatos. (2007). *Author Identification Using Imbalanced and Limited Training Texts. Proceedings of the 18th International Conference on Database and Expert Systems Applications, pp. 237-241.*, .

[14] Grieve J. (2007). *Quantitative Authorship Attribution: An Evaluation of Techniques. Arts - Humanities, 22, pp. 251-270*, .

[15] Jack Grieve. (2007). *Quantitative Authorship Attribution: An Evaluation of Techniques. Arts - Humanities, 22, pp. 251-270*, .

[16] Rosa María Coyotl Morales . (2007). *Clasificación Automática de Textos considerando el Estilo de Redacción.*

[17] Maya Carrillo. (2013). *Representando Estructura y Significado en Procesamiento de Lenguaje Natural, Tratamiento del Lenguaje y del Conocimiento, BUBOK PUBLISHING S.L.,2013.*

[18] Abbasi A. y H. . (September 2005). Chen. *Applying authorship analysis to extremistgroup web forum messages.. Intelligent Systems, IEEE, 20, pp. 67-7.*

[19] Morton A.Q. y Michaelson S. (1990). *The qsum plot. Technical Report. CSR-3-90, University of Edinburgh, UK.*

[20] Sanderson C. y S. (2006). Guenter. *Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation. In Proceedings of the International Conference on Empirical Methods in Natural Language Engineering, pp. 482-491.*

[21] Madigan D. y et al (2005). *Author identification on the large scal. In Proc. of the Meeting of the Classification Society of North America.*

[22] O. de Vel et al. (December 2001). *Mining e-mail content for author identification forensics. ACM SIGMOD Record, 30, pp. 55-64.*

[23] Definición.De. *http://definicion.de/gramatica/, Definición de gramática.*

[24] Definición.DE. *http://definicion.de/lexico/, Definición de Léxico.*

[25] Holmes D.I. y Tweedie F.J. (1995). *Forensic stylometry: A review of the cusum controversy. In Revue Informatique et Statistique dans les Sciences Humaines. Liege, Belgium: University of Liege.*

[26] Fernanda López Escobedo et al. *Exploración de medidas estilométricas para la atribución de autoría. Tercer Seminario de Lingũística Forense. Salón de Actos de la Facultad de Filosofía y Letras, agosto de 2013.*

[27] Hugo Jair Escalante et Al. *A Weighted Profile Intersection Measure for Profile-Based Authorship Attribution*, .

[28] Upendra Sapkota et Al. *The Use of Orthogonal Similarity Relations in the Prediction of Authorship*, .

[29] Sindhu Raghavan et al.. (2010). *Authorship Attribution Using Probabilistic Context-Free Grammars. 48th Annual Meeting of the Association for Computational Linguistics , In Proceedings, pp. 38-42.*

[30] Adrían Pastor et al. (2012). *A New Document Author Representation for Authorship Attribution. Pattern Recognition, 7329, pp. 283-292.*

[31] Grigori Sidorov et al.. (2013). *Syntactic Dependency-Based N-grams as Classification Features. Advances in Computational Intelligence, 7630, pp. 1-11.*

[32] Teng G. et al. (August 2004). *E-mail authorship mining based on SVM for computer forensic. International Conference on Machine Lerning and Cyhemetics, pp. 26-29.*

[33] Zheng R. et al.. (February 2006). *A framework for authorship identification of online messages: Writing-style features and classification techniques. The American Society for Information Science and Technology , 57, pp. 378-393.*

[34] Mosteller F., , y Wallace D.L. (1964). *Inference and disputed authorship: The Federalist. Reading, MA: Addison-Wesley.*, .

[35] Peng F., Shuurmans D., y Wang S. (September-December 2004). *Augmenting Naive Bayes Classifiers with Statistical Language Models. Information Retrieval, 7, pp. 317-345.*, .

[36] The Stanford Natural Language Processing Group. *http://nlp.stanford.edu/software/postagger-faq.shtml.*

[37] Van Halteren y H.. (January 2007). *Author verification by linguistic profiling: An exploration of the parameter space. ACM Transactions on Speech and Language Processing (TSLP), 4.*

[38] Hirst, G., y O. (2007). Feiguina. *Bigrams of syntactic labels for authorship discrimination of short texts. Literary and Linguistic Computing.*

[39] http://es.wikipedia.org/wiki/Categor %C3 %ADa_sint %C3 %A1ctica. *Categoría sintáctica.*

[40] Icarito. *http://www.icarito.cl/enciclopedia/articulo/segundo-ciclo-basico/lenguaje-y-comunicacion/gramatica/2009/12/97-8734-9-la-gramatica-y-sus-partes.shtml,Partes de la gramática.*

[41] Koppel M., Schler J., y E. (2007). Bonchek-Dokow. *Measuring differentiability: Unmasking pseudonymous authors. Journal of Machine Learning Research.*

[42] Koppel M. y J. (2003). Schler. *Exploiting stylistic idiosyncrasies for authorship attribution. In Proceedings of IJCAI Workshop on Computational Approaches to Style Analysis and Synthesis.*

[43] Clement R. y Sharp D. (2003). *N-gram and Bayesian classification of documents for topic and authorship. Literary and Linguistic Computing, 18, pp. 423-447.*

[44] Argamon S. y Levitan S. (2005). *Measuring the usefulness of function words for authorship attribution. In Proceedings of the 2005 ACH/ALLC Conference.*

[45] Keselj V., Peng F., Cercone N., y Thomas C. (2003). *N-gram-based Author Profiles for Authorship Attribution. In Proceedings of the Conference Pacific Association for Computational Linguistics, 3, pp. 255-264.*

[46] Justin Zobel. (2005). Ying Zhao. *Effective and Scalable Authorship Attribution Using Function Words. Information Retrieval Technology, 3689, pp. 174-189.*