



BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA
FACULTAD DE CIENCIAS DE LA COMPUTACIÓN

Determinación de características en el proceso de detección de autoría

TESIS PROFESIONAL PARA OBTENER EL TÍTULO DE:
MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

Esteban Castillo Juarez

Asesor:

Dra. Darnes Vilaríño Ayala

Co-asesor:

Dra. Maya Carrillo Ruiz

Computers are magnificent tools for the realization of our dreams, but no machine can replace the human spark of spirit, compassion, love, and understanding.

Louis Gerstner

Resumen

En este trabajo de tesis se experimenta con distintos métodos para la correcta representación de un autor dado en el problema de atribución de autoría. El objetivo es determinar un conjunto de características (novedosas) tanto léxicas como sintácticas y semánticas capaces de representar el estilo de escritura de un autor con respecto a otros de la manera más fiel posible.

En la primera parte del trabajo de investigación (Capítulos 1, 2 y 3) se realiza el planteamiento del problema a resolver así como se delimita un conjunto de objetivos a cumplir y se realizan distintas preguntas de investigación con respecto a la forma en que se va a resolver el problema. Por otra parte se realiza una revisión de toda la bibliografía relacionada al problema de atribución de autoría, haciendo una distinción entre los elementos principales que lo componen. Finalmente, se abordan los fundamentos teóricos relacionados a la extracción, representación y clasificación de características.

En la segunda parte (Capítulos 4, 5 y 6) se plantean y describen una serie de técnicas basadas en la extracción y clasificación de características léxicas, sintácticas y semánticas. Además, se propone una metodología para resolver el problema de atribución de autoría basada en la extracción y clasificación de grafos, donde se conserva la secuencia sintáctica de las oraciones para lograr descubrir patrones relevantes sobre los autores por medio de herramientas de minería de datos. Finalmente se presentan los resultados y conclusiones alcanzadas por medio de una serie de pruebas sobre un conjunto de datos preparado para el problema de atribución de autoría.

Los resultados obtenidos en este trabajo de investigación respaldan la idea del uso (de manera individual) y combinación de características, así como las representaciones basadas en grafos como herramientas estables (con respecto al tipo de corpus) para determinar el autor real de un documento de origen desconocido.



Agradecimientos

Este trabajo de tesis se lo dedico a mi madre Maria del Carmen Juarez Corona, a mi padre Esteban Castillo Lopez y a mi hermano Enrique Castillo Juarez. Por todo el apoyo, cariño y comprensión que siempre me han brindado a lo largo de todo este tiempo y por ser el mejor ejemplo de vida a seguir. Este logro es solo de ustedes.

Agradezco a mi asesora Dra Darnes Vilariño Ayala por todo su tiempo, paciencia, apoyo y consejos brindados a lo largo de mi maestría, ya que sin usted no hubiera llegado al final de este camino, **Gracias**.

De igual manera agradezco a:

- Dra. María Josefa Somodevilla García.
- Dr. David Eduardo Pinto Avendaño.
- M.C. Mireya Tovar Vidal.
- Dr. Ivan Olmos Pineda.
- Dra. María de la Concepción Pérez de Celis Herrero.
- Dra. Maya Carrillo Ruiz.

Los cuales dedicaron su tiempo y conocimientos a la revisión y desarrollo de este trabajo de tesis.

Finalmente (pero no menos importante) agradezco a mi compañero de laboratorio I.C.C. Saul León Silverio por su apoyo y ayuda para el desarrollo de mis experimentos.

Índice general

Índice general	I
Índice de figuras	V
Índice de tablas	VII
1. Introducción	1
1.1. Planteamiento de la investigación	2
1.1.1. Problema a resolver	2
1.1.2. Objetivos de la investigación	3
1.1.3. Justificación de la investigación	4
1.1.4. Preguntas de investigación	4
1.2. Aportaciones de la investigación	5
1.3. Organización de la tesis	5
1.4. Conclusión del capítulo	6
2. Estado del arte	7
2.1. Enfoque general	7
2.2. Visión general	8
2.3. Elementos principales de la atribución de autoría	11
2.4. Características en la atribución de autoría	11
2.5. Representación de la información	16
2.6. Clasificación de textos	19
2.7. Conclusión del capítulo	20
3. Atribución de autoría	21
3.1. Autoría del texto	21
3.2. Definición de atribución de autoría	22
3.3. Retos y problemas en la atribución de autoría	23
3.4. Características del texto	26

3.4.1.	Características léxicas	26
3.4.2.	Características sintácticas	27
3.4.3.	Características semánticas	28
3.5.	Corpus	29
3.6.	Clasificación de textos	32
3.6.1.	Árbol de clasificación	33
3.6.2.	Naïve Bayes	33
3.6.3.	Máquina de soporte vectorial	33
3.7.	Conclusión del capítulo	35
4.	Metodología de solución	37
4.1.	Primera aproximación: Palabras	37
4.1.1.	Características escogidas	37
4.1.2.	Representación de las características	38
4.1.3.	Proceso de clasificación	38
4.2.	Segunda aproximación: Extracción y combinación de caracte- rísticas	40
4.2.1.	Características escogidas	40
4.2.2.	Representación de las características	43
4.2.3.	Proceso de clasificación	44
4.3.	Tercera aproximación: Operaciones de conjuntos	46
4.3.1.	Características escogidas	46
4.3.2.	Representación de las características	46
4.3.3.	Proceso de clasificación	47
4.4.	Cuarta aproximación: Estadística	49
4.4.1.	Características escogidas	49
4.4.2.	Representación de las características	50
4.4.3.	Proceso de clasificación	51
4.5.	Quinta aproximación: Extracción de características utilizando representaciones basadas en grafos	53
4.5.1.	Características escogidas	53
4.5.2.	Representación de las características	53
4.5.3.	Proceso de clasificación versión 1	55
4.5.4.	Proceso de clasificación versión 2	56
4.6.	Sexta aproximación: Extracción de características utilizando una expansión basada en grafos	57
4.6.1.	Proceso de clasificación	61
4.7.	Conclusión del capítulo	61

5. Pruebas y resultados	63
5.1. Conjunto de datos	63
5.2. Primera aproximación	64
5.3. Segunda aproximación	67
5.4. Tercera y cuarta aproximación	71
5.5. Quinta aproximación	71
5.6. Sexta aproximación	72
5.7. Participación PAN CLEF 2012	77
5.8. Herramientas utilizadas y desarrolladas	77
5.9. Conclusión del capítulo	79
6. Conclusiones y trabajo a futuro	81
6.1. Conclusiones finales	81
6.1.1. Conclusión PAN CLEF 2012	83
6.2. Contribuciones del trabajo	84
6.3. Trabajo a futuro	84
Apéndice A. Clasificación con WEKA	87
Apéndice B. Búsqueda de patrones con SUBDUE	95
Bibliografía	97

Índice de figuras

3.1.	Representación del proceso de verificación de autoría.	24
3.2.	Representación del proceso de atribución de autoría.	25
3.3.	Clasificación de los tipos de corpus.	30
3.4.	Corpora utilizado en la atribución de autoría.	32
3.5.	Proceso de clasificación de textos.	34
4.1.	Esquema de la primera aproximación.	39
4.2.	Esquema de la segunda aproximación.	45
4.3.	Esquema de la tercera aproximación.	48
4.4.	Esquema de la cuarta aproximación.	52
4.5.	Representación basada en secuencias de palabras categoriza- das gramaticalmente.	55
4.6.	Representación con un modelo de estrella.	56
4.7.	Creación de un grafo por cada oración	57
4.8.	Esquema de la quinta aproximación versión 1.	58
4.9.	Esquema de la quinta aproximación versión 2.	59
4.10.	Esquema de la sexta aproximación.	60
5.1.	Gráfica de resultados de la primera aproximación.	66
5.2.	Gráfica de resultados de la segunda aproximación.	70
5.3.	Proporción de características usadas en la combinación.	70
5.4.	Gráfica de resultados de la tercera aproximación.	73
5.5.	Gráfica de resultados de la cuarta aproximación.	74
5.6.	Gráfica de resultados de la quinta aproximación.	75
5.7.	Gráfica de resultados de la sexta aproximación.	76
6.1.	Gráfica de los mejores resultados de cada aproximación.	83
A1.	Menú inicial de WEKA.	88
A2.	Menú de exploración de WEKA.	89
A3.	Menú de exploración con la opción de un archivo training.	89

A4. Menú de clasificación de WEKA.	90
A5. Menú de tipo de entrenamiento y algoritmo de clasificación. . .	90
A6. Menú de parámetros del algoritmo de clasificación.	91
A7. Menú de inicio de clasificación con training.	91
A8. Menú de selección del archivo test.	92
A9. Menú de prueba del Test.	92
A10. Menú de vizualización de errores del test.	93
A11. Menú de resultados del test.	93

Índice de tablas

3.1. Tabla de las características léxicas más usadas.	27
3.2. Tabla de las características de caracteres más usadas.	28
3.3. Tabla de las características semánticas más usadas.	29
4.1. Descripción de las categorías gramaticales	54
5.1. Número de documentos para training y test asociado al corpus C10	63
5.2. Resultados de la primera aproximación.	66
5.3. Resultados de la segunda aproximación.	69
5.4. Resultados de la tercera aproximación.	73
5.5. Resultados de la cuarta aproximación.	74
5.6. Resultados de la quinta aproximación.	75
5.7. Resultados de la sexta aproximación.	76
5.8. Conjunto de datos usados en PAN'12.	77
5.9. Resultados obtenidos en PAN'12.	77
5.10. Herramientas usadas en la aproximación 1,2,3,4,5 y 6.	78

INTRODUCCIÓN

Never promise more than you
can perform.

Publilius Syrus

Uno de los problemas más comunes en la mayoría de los documentos de texto presentes tanto de manera física como virtual en el mundo es la ausencia de un autor aparente, es decir, la falta de elementos implícitos dentro del texto que ayuden a discernir a quien le pertenece éste. Dicha problemática ha sido llamada de distintas maneras tales como adjudicación de textos, determinación de autoría o como actualmente se le conoce atribución de autoría. En la forma más básica, el problema de atribución de autoría consiste en una serie de muestras de textos de autores conocidos y se pide determinar cuál de ellos escribió un texto de procedencia anónima. Dado las investigaciones hechas sobre esta problemática, esta ha sido enmarcada dentro del modelo o paradigma de problemas de categorización de texto.

Atribución de autoría es un problema que se torna más difícil de resolver cuando se asigna un documento a un autor al cual no le pertenece. Por otra parte la búsqueda de características para representar a un autor específico supone una barrera que depende del tipo de lenguaje utilizado en cada texto. Si bien el lenguaje utilizado en un texto está limitado por el tema tratado o por el público a quién va dirigido (entre otros factores), el autor aun es libre de escribir como desea hacerlo, (lo cual supone un obstáculo más en la detección de dicho autor), es decir, un autor es libre de hacer oraciones o párrafos de la longitud que desee, de usar el vocabulario y los signos de puntuación como lo considere conveniente, o de usar el tiempo verbal que más le agrade. Lo anterior nos lleva al hecho de que a pesar de tener muchas características impredecibles en la detección de autoría (como un conjunto de autores abierto), la búsqueda de características que cree o plasmen fielmente

los principales rasgos de un autor, es una cuestión primordial en el proceso de entrenamiento sobre el mismo.

Teniendo en cuenta la problemática actual para determinar la procedencia de la información, la presente investigación plantea el desarrollo de algoritmos y modelos que permitan recuperar las características distintivas de un cierto autor, utilizando tanto elementos léxicos como semánticos para desarrollar técnicas de aprendizaje, que permitan posteriormente la clasificación de nuevos textos.

1.1. Planteamiento de la investigación

En esta sección se precisa el problema de la investigación a resolver, se definen los objetivos del proyecto, al mismo tiempo se plantea la propuesta de solución y por último se describe la organización de la tesis.

1.1.1. Problema a resolver

En años recientes ha habido un gran interés en la investigación de métodos automáticos para determinar la autoría de documentos anónimos basados en evidencia clara. En este sentido encontrar elementos que digan si cierto autor escribió un documento o no, han llevado a distintas ramas de las ciencias a buscar métodos para refinar y encontrar una solución óptima al problema. En este marco las ciencias computacionales han hecho muchos aportes e innovaciones en la creación de algoritmos capaces de resolver dicho problema. Dichas aportaciones han llevado a los investigadores a realizar numerosos experimentos sobre grandes volúmenes de información, así como a la creación de distintas metodologías para atacar el problema; como es el caso de atribución y verificación de autoría, donde el primero supone una serie de autores de referencia a comparar con un texto desconocido y el segundo es detectar la autoría sin dicha referencia, y sólo usando el estilo del texto a analizar.

Es claro que aunque se han hecho grandes avances existen ciertas áreas, en especial en atribución de autoría que no han sido explotadas del todo, este es el caso de la búsqueda de características semánticas en documentos escritos por diversos autores, las cuales proponen la búsqueda de dependencias de las palabras con respecto a su significado, así como el uso de las palabras

por un cierto autor.

Tomando en cuenta lo anterior, el problema a resolver en este trabajo de tesis se fundamenta en la idea de determinar el conjunto de características óptimas para asignar un documento a un autor dado, para esto se propone el desarrollo de modelos que permitan la detección de características léxicas, sintácticas y semánticas, lo cual nos debe permitir una mejor clasificación de textos desconocidos u anónimos.

1.1.2. Objetivos de la investigación

A continuación se presentan los objetivos generales y específicos relacionados a este trabajo de investigación.

■ Objetivo general

- Desarrollar métodos que permitan extraer características relevantes de un documento para la detección de autoría, utilizando propiedades léxicas, sintácticas y semánticas de los textos.

■ Objetivos específicos

- Estudiar los métodos existentes que extraen características léxicas de los documentos.
- Estudiar los métodos existentes que extraen características sintácticas de los documentos.
- Estudiar los métodos existentes que extraen características semánticas de los documentos.
- Desarrollar métodos para la atribución de autoría que combinen características léxicas y semánticas del texto.
- Determinar la representación de la información adecuada para el conjunto de características escogidas.
- Determinación del corpora a utilizar para la prueba de los métodos propuestos.

1.1.3. Justificación de la investigación

La gran cantidad de texto electrónico en la red ha hecho indispensable el uso de herramientas de recuperación de información y de procesamiento de lenguaje natural. Estas herramientas han desarrollado un conjunto de mecanismos capaces de recuperar, presentar y entender el conocimiento plasmado en los documentos, pero no responden a una pregunta común con respecto a la mayoría de los documentos de la Web **¿a quien le pertenece un documento de procedencia anónima?**. En este sentido aparecieron la atribución y verificación de autoría, los cuales tratan de responder esta pregunta. Para ello estos métodos han utilizado una serie de características para representar el texto.

A pesar del avance en la identificación del autor de un documento existen elementos que no han sido explotados, dichos elementos podrían ayudar a simplificar la identificación de autores, así como ayudarían a representar de mejor manera al estilo de escritura del autor.

Del preámbulo anterior surge el interés de esta investigación, ya que existen características que no han sido utilizadas con la importancia debida. En este sentido la investigación planteada en esta tesis plantea el análisis de las características anteriormente mencionadas, así como propone el uso de éstas en modelos de clasificación para recuperar a un autor dado utilizando un corpus de entrenamiento.

1.1.4. Preguntas de investigación

A continuación se presentan algunas preguntas que surgen en el área de atribución de autoría concerniente a la presente investigación:

- ¿Qué tipo de características aportan una mayor cantidad de información para la representación fiel de un autor dado?
- En el proceso de entrenamiento ¿Cuáles deben ser las principales cualidades con las que debe contar el corpus a utilizar?
- ¿Qué tipo de clasificadores son los más adecuados para detectar a que autor le pertenece un documento dado?

- ¿Qué ventajas y desventajas tiene el manejo de características léxicas, sintácticas y semánticas con respecto a las de otro tipo?

1.2. Aportaciones de la investigación

En esta sección se presentan las aportaciones de la investigación:

- Diseño de una metodología de trabajo para la atribución de autoría basándose en la extracción de características.
- Desarrollo de métodos para la atribución de autoría que combinen características léxicas, sintácticas y semánticas del texto.
- Selección de un corpus ideal para el desarrollo de los métodos.
- Evaluación de las características que mejor se acoplen al estilo de escritura de cada autor.

1.3. Organización de la tesis

Este trabajo de investigación se estructura en 6 capítulos distribuidos de la siguiente forma:

- **Capítulo 1.** Introducción. En esta parte se detalla el problema a resolver, los objetivos de la investigación, la justificación, así como también la propuesta de solución al problema y los productos generados por el trabajo. Al mismo tiempo se presentan las aportaciones que genera la construcción de una metodología para la atribución de autoría.
- **Capítulo 2.** Estado del arte. En esta parte se detallan un número importante de trabajos de investigación relacionados con el ámbito de atribución de autoría, haciendo énfasis en la búsqueda de características para delimitar el estilo de escritura del autor.
- **Capítulo 3.** Atribución de autoría. Se describen conceptos y términos relacionados con las propiedades más importantes de la atribución de autoría que se utilizarán en el desarrollo de este trabajo.

- **Capítulo 4.** Metodología de desarrollo. En este capítulo se detalla el marco metodológico de la presente investigación, que cubre como se va a resolver la atribución de autoría de un documento de origen desconocido.
- **Capítulo 5.** Pruebas y resultados. En este capítulo se comentan las pruebas y resultados experimentales de las diferentes metodologías propuestas en el capítulo 4. Se muestran también los algoritmos usados en cada aproximación y por último se hace una comparación de los resultados obtenidos.
- **Capítulo 6.** Conclusiones y Trabajo a Futuro. Finalmente en este capítulo se muestran las conclusiones alcanzadas en el trabajo de investigación, así como se presentan las contribuciones realizadas en el área(en específico en la extracción de características) y el trabajo a futuro a realizar.

1.4. Conclusión del capítulo

La investigación para determinar el autor de un texto dado, ha experimentado un crecimiento en los últimos años, en específico en la búsqueda de características, las cuales se han transformado en una de las piezas fundamentales para la correcta identificación del estilo de un autor. En este sentido, la mayoría de las metodologías actuales carecen de un conjunto de características léxicas, sintácticas y mayormente semánticas para la correcta representación de un autor, por lo tanto este trabajo de investigación se centra en estudiar diversos tipos de características, para detectar las más adecuadas.

ESTADO DEL ARTE

Never worry about theory as
long as the machinery does what
it's supposed to do.

Robert A. Heinlein

En esta sección se presenta un panorama global del estado del arte en el área de la atribución de autoría. Dicha area, por ser multidisciplinaria y de reciente nacimiento involucra diferentes líneas de investigación y la convierte en un tópico de actualidad.

Adicionalmente, este apartado describe los trabajos más relevantes que se han desarrollado de forma interdisciplinaria para determinar la autoría de un documento de procedencia anónima. Se comentan y se discuten los enfoques, técnicas, procedimientos y herramientas que se han propuesto e implementado. También se señalan las debilidades, aportaciones y areas de oportunidad en atribución de autoría.

2.1. Enfoque general

Se realizó un estudio sobre los artículos publicados en las áreas de Atribución de Autoría, Verificación de Autoría, Recuperación de Información y Procesamiento del Lenguaje Natural. Es con base en este estudio que se realizaron comentarios acerca de los artículos y trabajos; enfatizando sus avances, alcances, enfoques, ventajas y desventajas, así como sus aportaciones científicas encontrando el siguiente panorama general:

- La mayoría de los artículos hablan del uso de características como base para la obtención de elementos distintivos de un autor[1].

- Las características más comúnmente usadas son aquellas que presentan elementos léxicos y sintácticos [2].
- En los enfoques utilizados para resolver el problema de autoría no prevalece la idea del uso del sentido de las palabras para distinguir a un autor [3].
- Dentro de los muchos métodos de clasificación, las máquinas de soporte vectorial (SVM) son las que aportan una mejor clasificación en el área de atribución de autoría [4].
- En la bibliografía revisada los métodos de clasificación usan un aprendizaje supervisado [5, 6] .

2.2. Visión general

La idea principal detrás de la atribución de autoría es la de medir ciertas características del texto, con las cuales se pueda distinguir entre textos escritos por diferentes autores. Dicho campo soportado por métodos computacionales y estadísticos tiene ya una larga historia, empezando desde el siglo XIX y marcado por el estudio de Mosteller y Wallace [7] en la atribución de autoría de documentos federales de Estados Unidos. Este estudio fue seguido con la investigación de Mendenhall [8] sobre las obras de Shakespeare, así como los trabajos desarrollados por Yule [9, 10] y Zipf [11]. De hecho el trabajo de Mosteller y Wallace inició el estudio de la atribución de autoría no tradicional, el cual se fundamenta en la búsqueda de características con respecto al estilo de escritura de los autores.

A partir de los años 90 aparece una línea de investigación lingüística llamada estilometría [12], la cual utiliza una serie de medidas como tamaño de una sentencia, de una palabra así como frecuencias de éstas, para medir la riqueza de un vocabulario asociado a un autor específico. Las metodologías de atribución de autoría propuestas durante los 90 eran del tipo asistidas por computadora más que del tipo basadas en computadora, lo cual significa que el objetivo era raramente el desarrollo de un sistema totalmente automatizado. Ciertamente el problema de ese periodo fue la falta de evaluación objetiva de los métodos propuestos. En la mayoría de los casos, el campo de pruebas fue el de obras literarias de autores desconocidos o en disputa, por lo

que la estimación de la precisión sobre la autoría no era ni siquiera posible. Las limitaciones metodológicas de ese periodo concerniente a la evaluación de procedimientos eran:

- La información textual era demasiado grande (usualmente conteniendo libros enteros) y probablemente dicha información no era homogénea.
- El número de autores candidatos era demasiado pequeño.
- La evaluación del corpus no era controlada por el tipo de tópico.
- La evaluación de los métodos propuestos era simplemente intuitiva.

A partir del 2000 las cosas cambiaron en el estudio de atribución de autoría, dada la gran cantidad de textos electrónicos disponibles en el internet (correo electrónico, blogs, foros, etc), incrementando la necesidad del manejo eficiente de dicha información. Este hecho tuvo un impacto significativo en áreas como lo son Recuperación de Información, Aprendizaje Automático y Procesamiento del Lenguaje Natural, de la misma manera el desarrollo de estas áreas influenció a la atribución de autoría de la siguiente manera:

- La investigación acerca de la recuperación de información desarrolló técnicas efectivas para la representación de grandes volúmenes de información [13].
- Algoritmos de aprendizaje automático permitieron el manejo de datos dispersos y multidimensionales, lo cual proporcionó representaciones más expresivas de la información [14].
- Las herramientas desarrolladas en el área Procesamiento del Lenguaje Natural permitieron el análisis del texto de manera eficiente, así como se crearon nuevas medidas para la representación de estilos (características basadas en la sintaxis) [15].

Más importante aún, la gran cantidad de textos electrónicos disponibles revelan el potencial de la autoría de análisis en diversas aplicaciones [16], así como en diversas áreas incluyendo la de inteligencia [17], ley criminal [18, 19] y en aplicaciones de búsqueda literaria [20, 21]. La década pasada puede ser vista como una nueva etapa en el análisis y creación de nuevas tecnologías dominada por esfuerzos para el desarrollo de aplicaciones prácticas, las cuales

puedan lidiar con textos del mundo real, más que con resolver disputas literarias entre autores. El énfasis sobre la investigación de autoría se da ahora a la evaluación objetiva de las propuestas de métodos, así como la comparación de los diferentes métodos sobre la base de la referencia común de un corpus. Es importante que factores que no tomaban tanta relevancia como son el tamaño del texto de entrenamiento, así como el número de autores candidatos y la distribución de textos de entrenamiento entre estos sean considerados indispensables en la investigación actual.

Más allá del problema central de la atribución de autoría existen otras tareas de análisis de autoría como lo son:

- Detección de plagio, hallar similitud entre 2 textos [22].
- Identificación del perfil de un autor, extracción de información de la edad , estilo de escritura, educación , etc [23].
- Detección de inconsistencias de estilo , encontrar e identificar distintos patrones de diferentes autores sobre un mismo texto [24].

En la actualidad podemos encontrar que la atribución de autoría se ha vuelto un tópico de importancia en muchos ámbitos científicos, el cual es discutido en distintos foros para tratar de encontrar modelos y métricas capaces de resolver el problema de una forma óptima. Este es el caso de uno de los congresos de atribución de autoría más importante a nivel mundial, llamado *CLEF PAN Lab on Uncovering Plagiarism, Authorship, and Social Software Misuse*¹. En dicho congreso se abordan temas relacionados a la atribución de autoría, plagio de textos y vandalismo informático. En el caso de atribución de autoría se proponen nuevos métodos para abordar el problema así como se muestran características novedosas. En este sentido se puede tomar un punto de referencia actual acerca del trabajo realizado en todo el mundo para la obtención de características de calidad que representen el tipo de escritura de cierto autor de la mejor manera.

¹<http://www.webis.de/research/events/pan-11>

2.3. Elementos principales de la atribución de autoría

Es importante notar que para detectar la autoría de un documento de origen desconocido, es necesario obtener un conjunto de elementos que nos ayuden a representar a un autor con respecto a otros. También es indispensable, un esquema de representación para maximizar la representación de las características obtenidas. Posteriormente usar un algoritmo de clasificación, que permita determinar a qué clase(o autor) pertenece un documento. De lo anterior sobresale la importancia de hacer un análisis sobre 3 puntos importantes en la atribución de autoría:

1. Obtención de características
2. Formas de representación de la información
3. Métodos de clasificación de textos

En las siguientes secciones se hará un revisión de los trabajos más importantes basados en los tres puntos anteriores, haciendo énfasis en las propuestas hechas por cada autor para resolver el problema de autoría.

2.4. Características en la atribución de autoría

En esta sección se comentan algunos de los trabajos que se refieren al estudio y obtención de características sobre documentos de texto. Dicho de otra manera, el conjunto de elementos que sirven para representar el estilo con el cual el autor escribe sus documentos. Estos artículos explican los procesos involucrados para realizar la recuperación de dichos elementos así como estos han ido evolucionando dependiendo del contexto con el que se trabajó.

Dentro de las características a analizar en estos trabajos de investigación se encuentran:

- **Características léxicas:** una manera simple de ver un texto es como un conjunto de tokens agrupados en sentencias, cada token corresponde a una palabra, número o marca de puntuación. En este sentido las

características léxicas se pueden ver como la representación de las palabras como tal, en decir, se puede ver como el lenguaje utilizado por el autor para plasmar la información que quiere presentar al lector.

- **Características semánticas:** este tipo de características trabajan sobre el significado y asociación de las palabras, más que con la frecuencia de los textos o uso de los elementos ortográficos sobre estos, es decir, como al autor relaciona las palabras y los conceptos entre sí para formar lo que conocemos como conocimiento.
- **Características sintácticas:** en este tipo de elementos, un texto es visto como una secuencia de caracteres con un significado estadístico.

A continuación se presentan los trabajos más relevantes con respecto a la extracción y uso de las características para representar las propiedades de un autor:

- En el trabajo desarrollado en [25] se propone un método de comparación donde se usan varias colecciones de información considerablemente grandes y se comparan con distintos clasificadores como redes Bayesianas y Naïve Bayes. Las características seleccionadas son: pronombres, conjunciones, artículos, los cuales son incrementados y decrementados para obtener una perspectiva del comportamiento de un autor específico.
- En [3] se propone el uso de una nueva metodología para la prueba de características óptimas sobre un autor dado, en la cual, se mide el grado de precisión de un modelo cada vez que se va eliminando una característica del proceso de entrenamiento, con el objetivo de hacer una estimación del grado de utilidad de dicha característica con respecto a un autor dado. A diferencia de otros artículos en los cuales en base al tipo de tópico usado en el texto por un autor específico o al estilo observado de un autor con respecto a varias obras se escogen las características. En esta investigación se propone que se escoja un conjunto de características base y estas se pueden refinar a medida que se aprende más acerca del autor o del tópico.
- En [26] se propone un método para la atribución de autoría soportado por la idea de que una identificación apropiada sobre un autor debe

considerar medidas estilísticas, así como el tópico tratado por el autor, para esto, este método caracteriza una serie de documentos por medio de un conjunto de secuencias de palabras que combinan palabras funcionales y de contenido.

- La aportación de [1] consiste en la presentación de un estudio acerca de los elementos más importantes presentes en la atribución de autoría desde sus inicios hasta la actualidad, haciendo énfasis en la extracción de características y de cómo estas ayudan en el proceso de detección de un autor. En este trabajo Stamatatos presenta los siguientes tipos de características como la base de los elementos más usados para representar el estilo:
 - Características estilométricas
 - Características léxicas
 - Características semánticas
 - Características de carácter
 - Características de estructura

La importancia de este artículo radica en la presentación de la variedad de tipos de elementos que sirven para conformar la forma en que el autor escribe, así como se muestra las dificultades y los retos aun sin resolver en el área de autoría.

- En [2] al igual que en [1] se hace una revisión del área de atribución de autoría, haciendo énfasis en los inicios de la búsqueda de características sobre los autores, así como en la correcta representación de la información dentro de múltiples dominios en el texto. En este trabajo de investigación se plantean las principales diferencias en los distintos tipos de características presentes en la mayoría de los textos (léxicas, semánticas, sintácticas, entre otras) y se proponen medidas para medir el grado de acierto con respecto a las técnicas usadas para obtener el autor de un documento dado.
- La aportación de [27] consiste en el desarrollo de un sistema basado en la extracción de relaciones semánticas por medio de *Wordnet*². Las

²<http://wordnet.princeton.edu/>

características usadas consideran el contexto de las palabras nominales, así como diferentes tipos de conocimiento extraído de WordNet (tales como sinónimos, hiperónimos, merónimos entre otros). Para la clasificación se usa una red bayesiana usando una aproximación llamada *lightweight* donde las características recopiladas son presentadas en forma de conjuntos binarios.

- En [28] se describe un modelo para la clasificación de relaciones semánticas entre las palabras. En el sistema creado las características que usa son de tipo contextual, esto es características basadas en elementos léxicos y elementos de dependencia del significado de las palabras. Para obtener este tipo de elementos se propone el uso de bases de información léxica como *ProbBank*³ y *FrameNet*⁴. Por otra parte se propone el uso de lo que es llamado *Nominal Role Affiliation Features*, como características que capturen conocimiento acerca de los conceptos presentes en las oraciones, para esto se propone el uso de bases semánticas de información como *NomLex-Plus*⁵ y *VerbNet*⁶, las cuales proporcionan información acerca de las relaciones de una palabra con otras.
- En [29] se presenta una aproximación para la representación de elementos semánticos usando características que no dependen tanto del contexto como de los elementos sintácticos presentes en el texto. Para la creación de este modelo se hace uso de un clasificador de máxima entropía (*Maximum Entropy Classifier*[30]) así como de características que se construyen no por medio del contexto, si no, de las palabras individuales presentes en los documentos y las relaciones que estas guardan con otros elementos. De las características más importantes presentes en el artículo se puede encontrar:
 - El primer y segundo sustantivo presente en una oración.
 - Las primeras 2 palabras a la izquierda y derecha de un sustantivo dado.
 - Todas las palabras entre 2 sustantivos presentes en una oración.
 - La primera y última palabra de una oración.

³<http://verbs.colorado.edu/mpalmer/projects/ace.html>

⁴<https://framenet.icsi.berkeley.edu/fndrupal/>

⁵<http://nlp.cs.nyu.edu/nomlex/index.html>

⁶<http://verbs.colorado.edu/mpalmer/projects/verbnet.html>

- Extracción de sufijos.
 - La versión lematizada de una palabra.
 - Sinónimos de todas las palabras.
- En el artículo [5] se presenta un resumen de la evaluación de la metodología aplicada para la atribución de autoría en el congreso del CLEF PAN 2011⁷, se presentan las distintas soluciones utilizadas por los participantes para atacar este problema. Dentro del análisis hecho con respecto a las características usadas en el artículo se pueden identificar cuatro clases de características fundamentales:
- Características derivadas del uso de las palabras en el texto
 - Palabras
 - N-gramas de palabras
 - Pronombres
 - Palabras del discurso
 - Contracciones y abreviaciones
 - Características a nivel de carácter
 - N-gramas de caracteres
 - Sufijos
 - Puntuación
 - Longitud (sentencia y palabras)
 - Características basadas en el formato del texto
 - Ortografía
 - Longitud (sentencia y palabras)
 - Características basadas en la sintaxis del texto
 - Partes del discurso
 - Sintaxis
- En [31] se propone la combinación de distintas meta características para lograr representar la mayor cantidad de elementos distintivos de un autor. Para la combinación, se toman principalmente en cuenta elementos de tipo estilístico (como es el número de palabras, el número

⁷<http://www.webis.de/research/events/pan-11>

de líneas en el texto, el uso de contracciones, entre otras), así como elementos sintácticos y léxicos. La aportación fundamental de este trabajo es el de plantear la posibilidad de conjuntar distintos niveles de características disjuntos entre sí para aumentar el grado de detalle que se tiene acerca del estilo de un autor, para ello se plantea que los autores guardan similitudes en los patrones de escritura a través de distintos grados lingüísticos del texto.

- En [32] se plantea la idea de usar diferentes clasificadores para establecer un sistema de **voto** donde un conjunto de características es determinado como representativo de un autor sólo si dicho conjunto presenta un comportamiento óptimo en distintos procesos de clasificación. La mayoría de las características usadas para determinar el estilo del autor en este artículo contabilizan el uso de elementos léxicos dentro del texto. Entre las características más importantes se pueden encontrar:

- Número de palabras
- Número de líneas de texto
- Número de caracteres en un documento
- Número de sentencias
- Número de bloques de texto
- Número de palabras cerradas
- Uso de las letras capitales

La importancia de este artículo radica en la combinación de distintos clasificadores para probar el peso y el grado de utilidad de un conjunto de características con respecto a un autor dado.

2.5. Representación de la información

En esta sección se comentan los trabajos relacionados con la correcta representación de la información (o en nuestro caso los documentos asociados a un autor), el cual es un elemento crucial en la atribución y verificación de autoría. La representación de la información permite conjuntar todas las características de un autor en un esquema específico, el cual facilita la construcción de modelos con los cuales se determina la autoría de un documento

dado. La mayoría de trabajos de investigación presentados en esta sección presentan diversas formas para representar la información:

- **Términos:** En esta representación del conjunto de texto original, sólo se eliminan algunos elementos tales como puntuación, palabras cerradas entre otros.
- **Bolsa de palabras:** En esta representación el texto es visto como una colección no ordenada de palabras, donde la frecuencia de ocurrencia de las palabras es más importante que la posición y relación entre éstos.
- **N-gramas:** Esta representación es vista como una secuencia contigua de n-ítems de una sentencia dada, donde cada secuencia guarda elementos tales como orden y relación entre las palabras.
- **LowBow:** Esta representación es llamada bolsa de palabras localmente ponderada, esta toma ventanas no contiguas de elementos. Se crean histogramas que son usados para calcular la relación entre ciertos segmentos de texto.
- **Grafos:** en esta representación se utiliza un esquema de nodos interconectados por medio de aristas, donde cada nodo representa ideas o conocimiento y las aristas son la forma de representar la relación que guardan los nodos entre sí.

A continuación se presentan los trabajos más relevantes con respecto a la forma en que se representa la información:

- En [33] se explica el proceso de descubrimiento de información en largos volúmenes de información por medio de grafos de conocimiento o también llamados de dominio semántico, donde la información se presenta como una unión de conceptos claves que representan un dominio en particular. Además de mostrar los elementos clave de representación del grafo, en este trabajo se hace especial énfasis en la búsqueda de patrones significativos de manera supervisada y no supervisada, haciendo una comparativa entre estas, describiendo sus ventajas y desventajas con respecto a la representación del grafo.
- En [34] se presenta la idea de usar grafos como representación de conocimiento (grafo con significado) para redes semánticas. El grafo está

compuesto por vértices y aristas; los vértices o nodos representan conceptos y las aristas representan las relaciones que tienen dichos conceptos. Los conceptos se construyen considerando el contexto.

- En [35] se propone el uso de la representación de n-gramas de 3 caracteres usando una nueva medida de distancia entre las palabras, la cual es estable para largos conjuntos de documentos; se discute además la dificultad que se presenta en el entrenamiento cuando el corpus es desbalanceado.
- En [36] se describe una metodología para la extracción de información semántica por medio de grafos de texto, donde cada vértice corresponde a una palabra y las aristas son la dependencia semántica asociada entre dos palabras. A través de este método el artículo propone la generación de conocimiento por medio de las relaciones hechas entre las palabras y las dependencias semánticas(o categorías), dicho conocimiento es representado como un nuevo grafo que muestra la información esencial de un autor con respecto a otros. Puesto que el uso y representación de información semántica no es muy habitual en el estudio de atribución de autoría, la importancia del artículo queda justificada para el uso en el ámbito de autoría y la generación de información acerca de un autor dado.
- En [37] se propone un modelo de representación llamado Tensor Space Model, para descubrir las propiedades de estilo de los textos, tratando con ello de disminuir la problemática de contar con un corpus desbalanceado, es decir, que en general no se disponen de documentos distribuidos uniformemente a través de los autores. Se generaliza el modelo de clasificación basado en máquinas de soporte, para que use tensores como soporte. Los experimentos reportados se realizaron sobre varios corpora con dominios muy específicos.
- Las aportaciones de [38] consisten en el uso de histogramas locales sobre n-gramas de caracteres para la atribución de autoría. En particular se muestra que el uso de dichos histogramas permite preservar información secuencial de los documentos, así como es de gran ayuda en situaciones donde el conjunto de entrenamiento no sea muy grande o no esté balanceado. Para lograr la representación anterior en el artículo se comenta que se hace uso de una bolsa de elementos localmente

ponderada (LowBow) y de una máquina de soporte vectorial para la clasificación del texto.

2.6. Clasificación de textos

En esta sección se discuten los trabajos acerca de los métodos de clasificación usados dentro de la atribución de autoría, así como los trabajos en los que se comparan distintas metodologías con respecto a las características usadas para representar el estilo de escritura de un autor.

Es importante destacar que en la definición de clasificación de textos, presentada a continuación, es consistente en todos los artículos presentados en esta sección.

La categorización de textos o clasificación de textos es la tarea de etiquetar o asignar a un nuevo documento a una clase, basado en un conjunto de documentos previamente clasificados. Este tipo de clasificación es llamado un problema supervisado o semi supervisado[39], esto es, para clasificar una nueva entrada el clasificador entrena con un conjunto de documentos previamente etiquetados (training corpora).

A continuación se presentan los trabajos más relevantes con respecto a la categorización de textos usando distintos clasificadores:

- En [4] se explican las bases del algoritmo de clasificación de Roccio aplicado al ámbito de categorización de textos. La aportación de este trabajo de investigación con respecto a otros es el uso o incorporación del TF-IDF[13] dentro del algoritmo, así como también se propone una variación interesante de la comúnmente usada bolsa de palabras. Como referencia especial este trabajo hace una comparación entre Naïve Bayes y Roccio como clasificadores.
- En [40] se explora el uso de las máquinas de soporte vectorial (Support Vector Machine: SVM) para la clasificación de textos, este analiza las características particulares del aprendizaje por medio de información e identifica que máquina de soporte vectorial es adecuada para los distintos tipos de texto. En particular este trabajo de investigación se enfoca en el rendimiento de dos SVMs, el que usa un **kernel polinomial** y el

que usa un **kernel de tipo RBF**. Finalmente hace una comparación con otros métodos de clasificación como: Naïve Bayes , Roccio, C4.5 y K-NN entre otros para mostrar la efectividad de los SVMs.

- En [41] se hace una revisión exhaustiva de los aspectos principales del aprendizaje automático y de como este puede ser usado en la tarea de clasificación de textos. Por otra parte presenta toda la teoría acerca del clasificador de textos probabilístico Naïve Bayes, así como las diferentes variaciones de este y su aplicación a la tarea de clasificación.

2.7. Conclusión del capítulo

En este capítulo se presentan los resultados más relevantes en la tarea de atribución de autoría, obtenidos hasta el momento. Analizando los resultados reportados queda claro que es una tarea aun no resuelta, porque los algoritmos propuestos y las metodologías de solución ofrecen resultados aceptables para corpora con ciertas características , es decir, que dependen mucho de los datos de prueba así como de los de entrenamiento.

ATRIBUCIÓN DE AUTORÍA

I'm not a very good writer, but
I'm an excellent rewriter.

James Michener

En este capítulo se describen conceptos y términos que se utilizan en el desarrollo de esta tesis. En especial, se describen los aspectos relacionados con la atribución de autoría, haciendo énfasis en la determinación de las propiedades de escritura de un autor por medio de la extracción de características léxicas, sintácticas y semánticas.

3.1. Autoría del texto

Escribir nos ayuda a pensar, a descubrir nuevas conceptos, a organizar nuestro pensamiento y plasmar nuestro conocimiento, en otras palabras escribir es convertir las ideas en palabras escritas, de manera ordenada, respetando las reglas de sintaxis y de ortografía, así como de puntuación. Expresar las palabras que se traducirán en ideas originales es responsabilidad del autor, el cual imprime en cada uno de sus escritos un estilo, una forma de pensamiento, una personalidad, capaz de diferenciar su concepción de las ideas con respecto a otros autores, es decir la capacidad del autor de plasmar su autoría sobre un texto.

Distinguir un autor de otro se ha convertido en los últimos años en una tarea difícil de resolver, en especial en el Internet, dado que el crecimiento de la información textual almacenada en la Web ha posibilitado que la mayoría de los textos sean de precedencia desconocida. Como consecuencia la mayoría de los escritos actuales han adoptado una combinación de distintos estilos y

formas de pensamiento, llevando a la autoría de un texto a un nuevo nivel de dificultad.

3.2. Definición de atribución de autoría

Existen numerosas definiciones con respecto a la atribución de autoría. A continuación se muestran algunos extractos de dichas definiciones para lo que se conoce como atribución de autoría clásica.

- Según la definición de [38] la Atribución de autoría es ***la tarea de decidir de un conjunto de autores conocidos, a quien le pertenece un documento dado.***
- En la definición de [2] se dice que la atribución de autoría es ***la ciencia de inferir un conjunto de características de un autor por medio de los documentos que este haya escrito.***
- De acuerdo a [41] es ***el proceso de asignarle autoría a una muestra de texto a su correcto autor, de entre un conjunto finito de autores.***
- Según la definición de [4] la atribución de autoría es ***la correcta identificación del autor de un documento desconocido, tomando como referencia un conjunto de documentos de autores conocidos.***
- En la definición de [42] se plantea a la atribución de autoría como ***la conexión de un documento de precedencia anónima con un específico autor, usando un conjunto de características cuantificables, como indicadores del estilo del autor.***

En resumen se puede decir que la atribución de autoría, es la correcta clasificación de un documento dado, por medio de la captura del estilo de escritura, comparando con un conjunto de autores de los cuales de cada uno se tiene una serie de muestras de texto de su autoría. La atribución de autoría es un área interdisciplinaria, que abarca tantas disciplinas que genera normalmente un conocimiento parcial desde una u otra perspectiva. Algunas de las disciplinas más importantes que participan dentro de estos estudios son la Recuperación de Información, Aprendizaje Automático y Procesamiento del Lenguaje Natural entre muchos otros.

3.3. Retos y problemas en la atribución de autoría

Actualmente existen distintos retos y problemas en cuanto al campo de atribución de autoría se refiere, los cuales se destacan a continuación:

- Problema

- Dado un documento desconocido, encontrar el conjunto de propiedades representativas de un autor determinado, a este tipo de problema se le llama problemas estilométricos.
- Dado un documento de autoría desconocida y documentos de un conjunto de candidatos conocidos, la tarea de la atribución de autoría es la de asignar dicho documento de procedencia anónima, al autor real con respecto al conjunto de autores candidatos. A este tipo de problema se le llama de clase abierta (ver Figura 3.1).
- Dado un documento de autoría desconocida y un texto de un autor específico, la tarea es determinar si dicho texto ha sido escrito por ese autor. A este tipo de problema se le llama de clase cerrada (ver Figura 3.2).

- Retos

- Heterogeneidad del dominio en los textos.
- Metodologías para la reducción eficiente de la dimensionalidad de los datos.
- Búsqueda de características para representar a un autor específico.
 - Variedad de características.
 - ¿Cuál es el mejor tipo de características?
- Delimitación del corpus a usar para un tópico dado.
- Uso eficiente de las herramientas para clasificar la información.
- Análisis eficiente de la interrelación del estilo de escritura de diferentes autores.

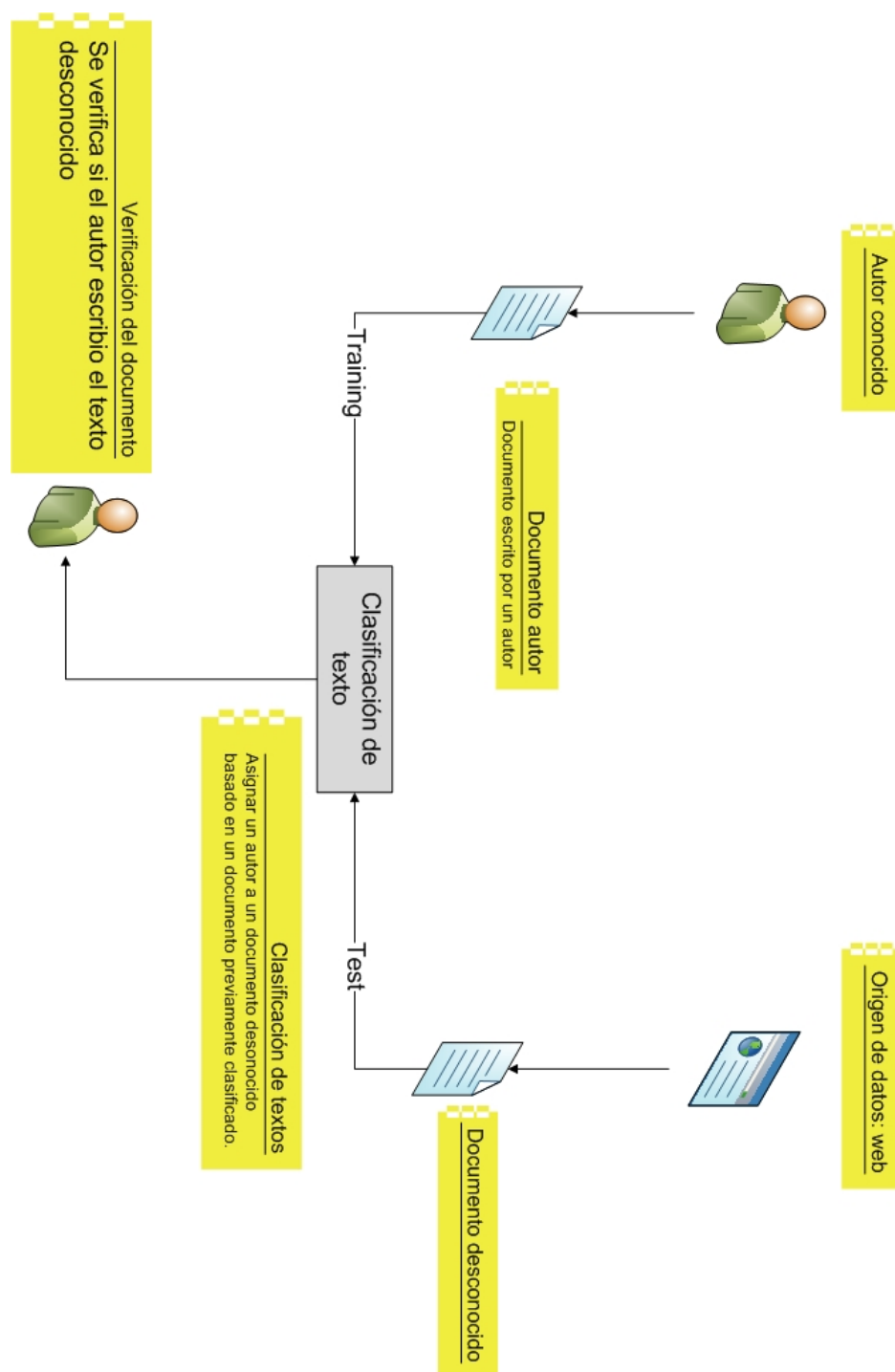


Figura 3.1: Representación del proceso de verificación de autoría.

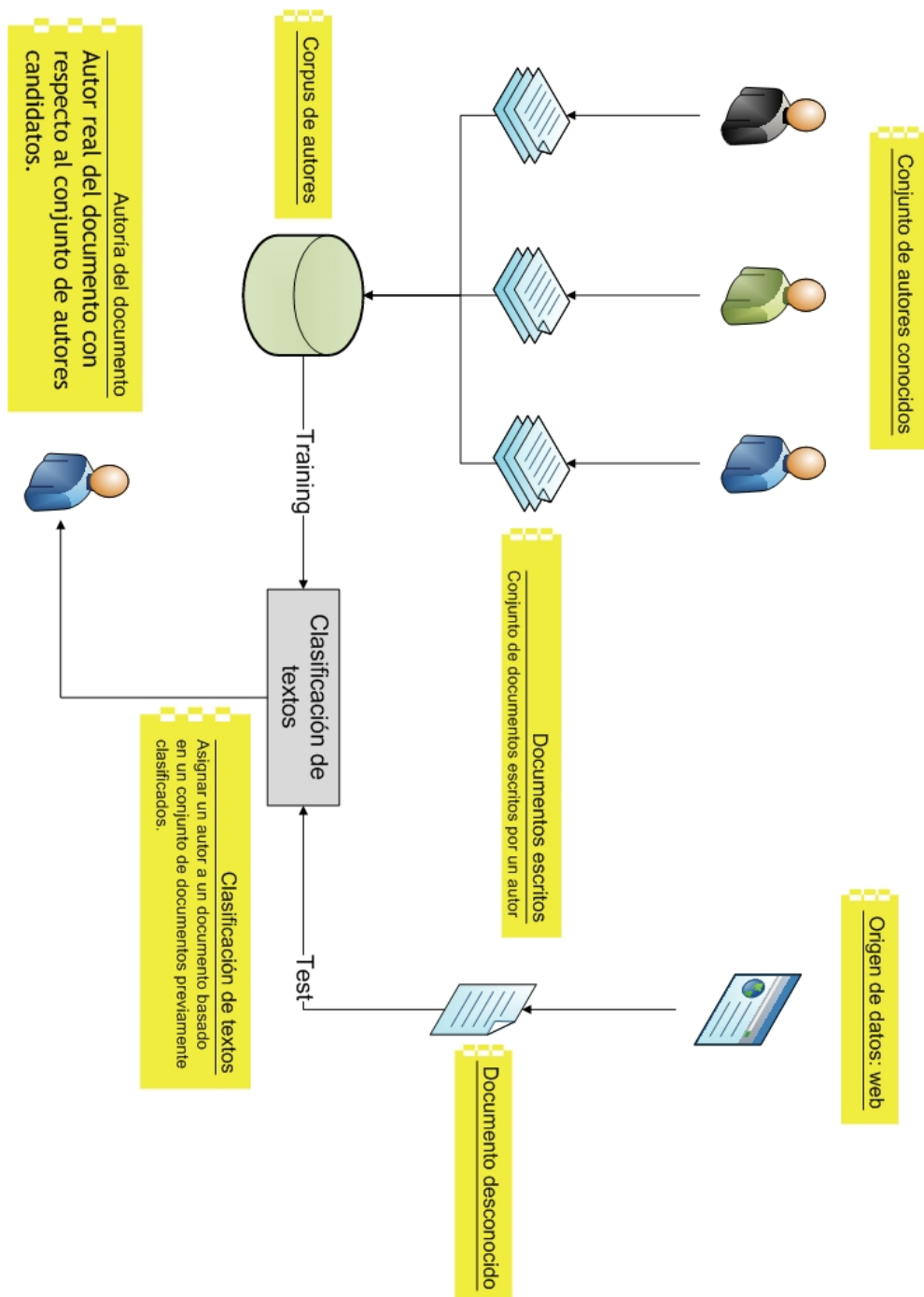


Figura 3.2: Representación del proceso de atribución de autoría.

3.4. Características del texto

Una de las principales preocupaciones en la atribución de autoría es la búsqueda de todas aquellas propiedades cuantificables de un autor capaces de diferenciarlo de otros. A este tipo de elementos presentes en la mayoría de los textos se les llama características, una característica es toda aquella cualidad que determina los rasgos de escritura de un autor y que muy claramente lo distingue del resto, en otras palabras las características que tiene un autor resultan ser sus notas particulares que los diferencian y de alguna manera los hacen ser quien es. La función de las características es la de proporcionarnos información acerca del autor de aquello que queremos conocer.

Existe una gran variedad de características presentes en los textos, cada una de estas representan un aspecto acerca de la idea que el autor quiso dar a conocer. De entre todas las características que existen en los textos, en atribución de autoría podemos encontrar 3 tipos principales, características léxicas, sintácticas y semánticas, cada uno de estas representa un aspecto importante del autor dentro del lenguaje. En las siguientes secciones se habla más a fondo de cada una, haciendo énfasis en la utilidad dentro de autoría.

3.4.1. Características léxicas

El lenguaje se configura como aquella forma que tienen los seres humanos para comunicarse. Se trata de un conjunto de signos, tanto orales como escritos, que a través de su significado y su relación permiten la expresión y la comunicación humana. Ahora bien el léxico es el sistema de unidades lingüísticas del idioma, es decir, el léxico es el vocabulario de un idioma o región, el diccionario de una lengua o el caudal de modismos y voces de un autor. Como tal las palabras son el tema central del léxico y son el mayor caudal del cual se componen, dependiendo del idioma, el léxico cambia y se adapta dando como resultado diferentes tipos de características.

El léxico del autor determina con que frecuencia se usan distintas palabras y en qué orden las usa para formar distintas ideas. Para determinar las propiedades léxicas del autor es necesario comprender el dominio que usa así como el concepto que quiere plasmar, ya que de esto depende el uso del conjunto de palabras a usar, si bien es cierto que aunque varios autores ha-

blen el mismo idioma no es necesariamente una condicional para que usen el mismo conjunto de palabras, por ejemplo el léxico no es el mismo para un libro de matemáticas, que para un libro de poesía. Por lo anterior se puede ver que el léxico tiene que ver en como escribe el autor y lo que escribe.

Dentro de atribución de autoría las características léxicas son representadas muy a menudo como los distintos tipos de palabras que existen en el vocabulario o como todos aquellos elementos que sean distintivos de un tópico en específico.

En la Tabla 3.1 se presentan algunos ejemplos de elementos léxicos tomados en cuenta en la mayoría de los trabajos, así como las herramientas que se usan para encontrarlos y delimitarlos en un texto.

Características	Herramienta requerida
Longitud de palabras	Tokenizador
Riqueza del vocabulario	Tokenizador
Frecuencia de palabras	Tokenizador
Contracciones y abreviaciones	Tokenizador
Verbos modales	Cluster
Uso de prefijos	Tokenizador
Uso de sufijos	Tokenizador
Categorización del texto	Pos Tagger
Uso de las palabras cerradas	Tokenizador
Colocaciones de las palabras	Tokenizador
Puntuación	Tokenizador

Tabla 3.1: Tabla de las características léxicas más usadas.

3.4.2. Características sintácticas

Para escribir una o más palabras no sólo es necesario saber que palabras necesitamos para formar nuestra idea, sino que además debemos de conocer el conjunto de reglas que gobiernan a las palabras, es decir, necesitamos saber la gramática (el estudio de las reglas y principios que regulan el uso de las lenguas y la organización de las palabras) para coordinar y unir las palabras, así como para formar oraciones y expresar conceptos. Ahora bien, en el sentido anterior la sintaxis es el conjunto de pautas que definen las secuencias correctas de los elementos de un lenguaje.

En el ámbito de atribución de autoría, las características sintácticas ayudan a representar el sentido estructurado de los conceptos de un autor, a

mostrar las relaciones estructurales entre las palabras y a integrar de manera correcta a las sentencias para dar sentido a los conceptos dentro de un documento. Varias medidas pueden ser definidas para la identificación de un autor incluyendo conteo de caracteres alfabéticos, de dígitos, de mayúsculas y minúsculas entre otros. Es importante denotar que dada la simpleza de dichas características, estas están presentes en la mayoría de corpus de texto en la actualidad, así como representan fielmente el estilo de escritura de un autor (desde el punto de vista del uso y frecuencia de las palabras en un texto). Otra ventaja de este tipo de características es la posibilidad de ser tolerante al ruido, esto es, textos los cuales contienen una gran cantidad de errores gramaticales así como un extraño uso de los signos de puntuación.

En la Tabla 3.2 se presentan algunos ejemplos de elementos sintácticos tomados en cuenta en la mayoría de trabajos, así como las herramientas que se usan para encontrarlos y delimitarlos en un texto.

Características	Herramienta requerida
Tipos de carácter	Diccionario de caracteres
Sufijos frecuentes	Selector de características
Elementos de Puntuación	Tokenizador
Ortografía	Cluster

Tabla 3.2: Tabla de las características de caracteres más usadas.

3.4.3. Características semánticas

La lengua está formada por un conjunto de signos y normas para ser utilizados dentro de cualquier escrito. El signo lingüístico consta de un significante el cual es la secuencia de fonemas o de letras que percibe el hablante y un significado que es la imagen psíquica que está asociado a un determinado significante. Pues bien, en el sentido anterior la semántica es la disciplina que se ocupa dentro de la lengua del significado de los signos lingüísticos: palabras, oraciones, y textos. Es decir, el estudio de todo aquello perteneciente o relativo a la significación de las palabras en un contexto determinado.

El estudio del significado de las palabras se enfrenta siempre a cierta imprecisión, ya que depende tanto del contexto lingüístico, como del extralingüístico. En el caso del contexto lingüístico de una palabra lo constituyen las demás palabras que lo rodean, por otra parte el contexto extralingüístico

es la situación en la que se pronuncia una palabra.

Dentro de la búsqueda de características semánticas en la atribución de autoría se puede ver que el dominio de las palabras juega un papel muy importante, ya que el significado de estas está formado por un conjunto de semas (cada uno de los rasgos mínimos en los que se puede descomponer una palabra). Sin embargo no todos los semas son igualmente compartidos por todos los textos de diferentes autores, sino que hay algunos de ellos que están presentes, mientras que otros varían. Es decir, el significado de una palabra no es siempre exactamente igual en todos los casos.

En la Tabla 3.3 se presentan algunos ejemplos de elementos semánticos tomados en cuenta en la mayoría de trabajos, así como las herramientas que se usan para encontrarlos y delimitarlos en un texto.

Características	Herramienta requerida
Sinónimos	Tesaurus, Tokenizador
Dependencias semánticas	Tokenizador, Parser semántico, Pos tagger
Hipónimos	Word net
Hiperónimos	Word net

Tabla 3.3: Tabla de las características semánticas más usadas.

3.5. Corpus

Un corpus lingüístico o de textos es un conjunto de ejemplos reales de información que muestran el uso de una lengua. Actualmente encontramos la mayoría de estos almacenados de manera electrónica. Dicho conjunto de información no necesariamente es de una sola lengua (corpus monolingüe) si no que también puede haber corpus que comparen dos lenguas o incluso tres, en cuyo caso se llamarían corpus bilingües o trilingües. Existen diferentes corpus que representan diversos dominios ya sean de arte, literatura, noticias, etc. Cada uno presenta distintos retos, así como cada uno ofrece ventajas y desventajas, pero ¿cuál de ellos es el más adecuado para el proceso de detección de autoría?.

La decisión del uso de un tipo de corpus se hace en base a las necesidades, en el caso de atribución de autoría se necesita que el corpus a usar presente abundantes ejemplos del estilo de escritura del autor, así como un amplio

conjunto de autores. Por otro lado en el caso de verificación de autoría no es necesario ya que sólo se necesita un ejemplo, el cual sea rico en elementos que distingan a un autor. Teniendo lo anterior podemos encontrar la siguiente clasificación general de tipos de corpus [39]:

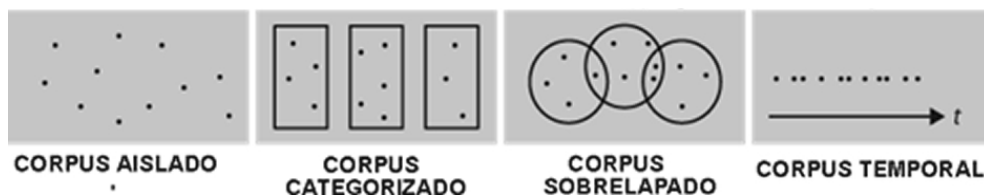


Figura 3.3: Clasificación de los tipos de corpus.

En la Figura 3.3 podemos encontrar 4 tipos fundamentales de corpus presentes en la mayoría de ámbitos:

- **Corpus aislado:** En el cual no existe una categoría definida así como no hay una asociación aparente entre la información que lo conforma.
- **Corpus categorizado:** A diferencia del corpus aislado, este corpus presenta una fuerte categorización de los elementos, esto es, existe una categoría evidente para cada documento que existe.
- **Corpus sobrelapado:** Este corpus es casi igual al corpus categorizado con la salvedad de que existen documentos que pertenecen a mas de una categoría o a más de una clase.
- **Corpus temporal:** En este tipo de corpus existe un dominio claro, el cual se repite a través del tiempo variando los elementos presentes en cada documento.

En la Figura 3.4 se observan los siguientes tipos de corpus:

- **Clef Pan Corpus:** Conjunto de datos en inglés, propuesto en el marco del congreso PAN'11 [43], el cual consta de tres colecciones de textos sobrelapados y desbalanceados acerca de chats de la Empresa Enron, el primero (small) con 26 autores diferentes, el segundo (large) con 72 autores diferentes y el tercero (verify) con sólo un autor para verificación de autoría.

- **Corpus CCAT:** Conjunto de datos en inglés creado por Efstathios Stamatatos a partir del RCV1 [44], consta de noticias corporativas industriales. Este Corpus (ccat) está compuesto de dos colecciones de documentos balanceados y disjuntos, el primero(C50) con 50 autores, cada uno con 50 documentos de su autoría y el segundo(C10) con 10 autores, cada uno con 50 documentos de su autoría.
- **Inagural Corpus:** Corpus [39] con los discursos presidenciales más importantes de los 44 presidentes de los Estados Unidos de América.
- **Brown Corpus:** Corpus[39] categorizado el cual cuenta con 15 categorías temáticas, las cuales son:
 - Reportajes
 - Editorial
 - Reviews
 - Religión
 - Habilidades y hobbies
 - Costumbres
 - Misceláneas
 - Cartas a la prensa
 - Aprendizaje
 - Ficción
 - Misterio
 - Ciencia ficción
 - Aventuras
 - Romance
 - Humor
- **Corpus Gutenberg:** Corpus[39] con alrededor de 1,000 libros de diferentes autores sin ninguna relación aparente.

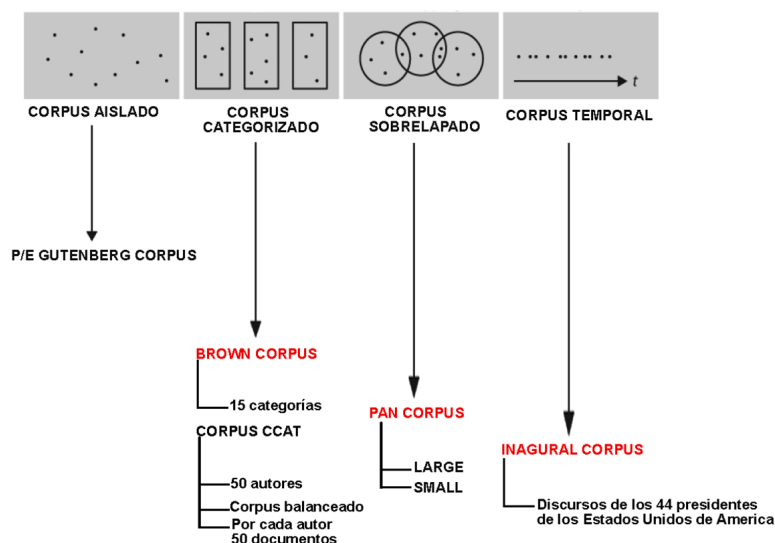


Figura 3.4: Corpora utilizado en la atribución de autoría.

3.6. Clasificación de textos

La categorización de textos o clasificación de textos es la tarea de etiquetar o asignar a un nuevo documento una clase, basado en un conjunto de documentos previamente clasificados. Este tipo de clasificación es llamado un problema supervisado o semi supervisado [39], esto es, para clasificar una nueva entrada el clasificador entrena con un conjunto de documentos previamente etiquetados (training corpora).

En la Figura 3.5 se muestra la tarea de categorización o clasificación de textos aplicada al problema de atribución de autoría, donde en la parte del entrenamiento (training) a) un extractor de características el cual es usado para convertir cada elemento (palabras) en parte del conjunto de entrenamiento, dicho conjunto almacena la información más relevante de cada autor, la cual será usada para clasificar al documento. En el clasificador de textos entra como parámetros el conjunto de entrenamiento y las etiquetas asociadas a cada documento (las etiquetas o clases son previamente asignadas a cada documento), donde la salida será un modelo que represente a cada uno de los autores. En la parte b) llamado paso de predicción(test), el mismo extractor de características extrae las palabras de los nuevos documentos (documentos

de procedencia anónima) a clasificar y las introduce al modelo, dando como resultado una clase asociada al documento.

3.6.1. Árbol de clasificación

De todos los métodos de aprendizaje (o de obtención del conocimiento) los sistemas de aprendizaje basados en árboles de clasificación [45] es quizás el método más fácil de utilizar y de entender. Un árbol de clasificación es un conjunto de condiciones organizadas en una estructura jerárquica, de tal manera que la decisión final se determina siguiendo las condiciones que se cumplen desde la raíz hasta alguna de sus hojas. Una de las grandes ventajas de los árboles de clasificación es que, en su forma más general, las opciones posibles a partir de una determinada condición son excluyentes. Esto permite analizar una situación y, siguiendo el árbol de clasificación apropiadamente, llegar a una sola acción o decisión a tomar.

3.6.2. Naïve Bayes

Es un clasificador probabilístico [41] basado en la aplicación del teorema de Bayes, tomando fuertes asunciones de independencia entre sus elementos. En términos simples el clasificador Naïve Bayes asume que la presencia u ausencia de una característica no está relacionada con la presencia u ausencia de otra característica. Para asignar una clase a un texto de origen desconocido, el clasificador calcula la probabilidad a priori sobre cada característica, la cual se calcula con la frecuencia de cada clase en el entrenamiento. La contribución de cada característica es combinada para obtener una probabilidad de estimación (*likelihood*), la clase con la mayor probabilidad de estimación es asignada al documento desconocido.

3.6.3. Máquina de soporte vectorial

Una máquina de soporte vectorial (Support Vector Machine, por sus siglas en inglés SVM) es un sistema de aprendizaje basado en el uso de un espacio de hipótesis de funciones lineales en un espacio de mayor dimensión inducido por un Kernel, en el cual las hipótesis son entrenadas por un algoritmo tomado de la teoría de optimización, el cual utiliza elementos de la teoría de generalización. Debido a las limitaciones computacionales de las máquinas de aprendizaje lineal estas no pueden ser utilizadas en la mayoría de las

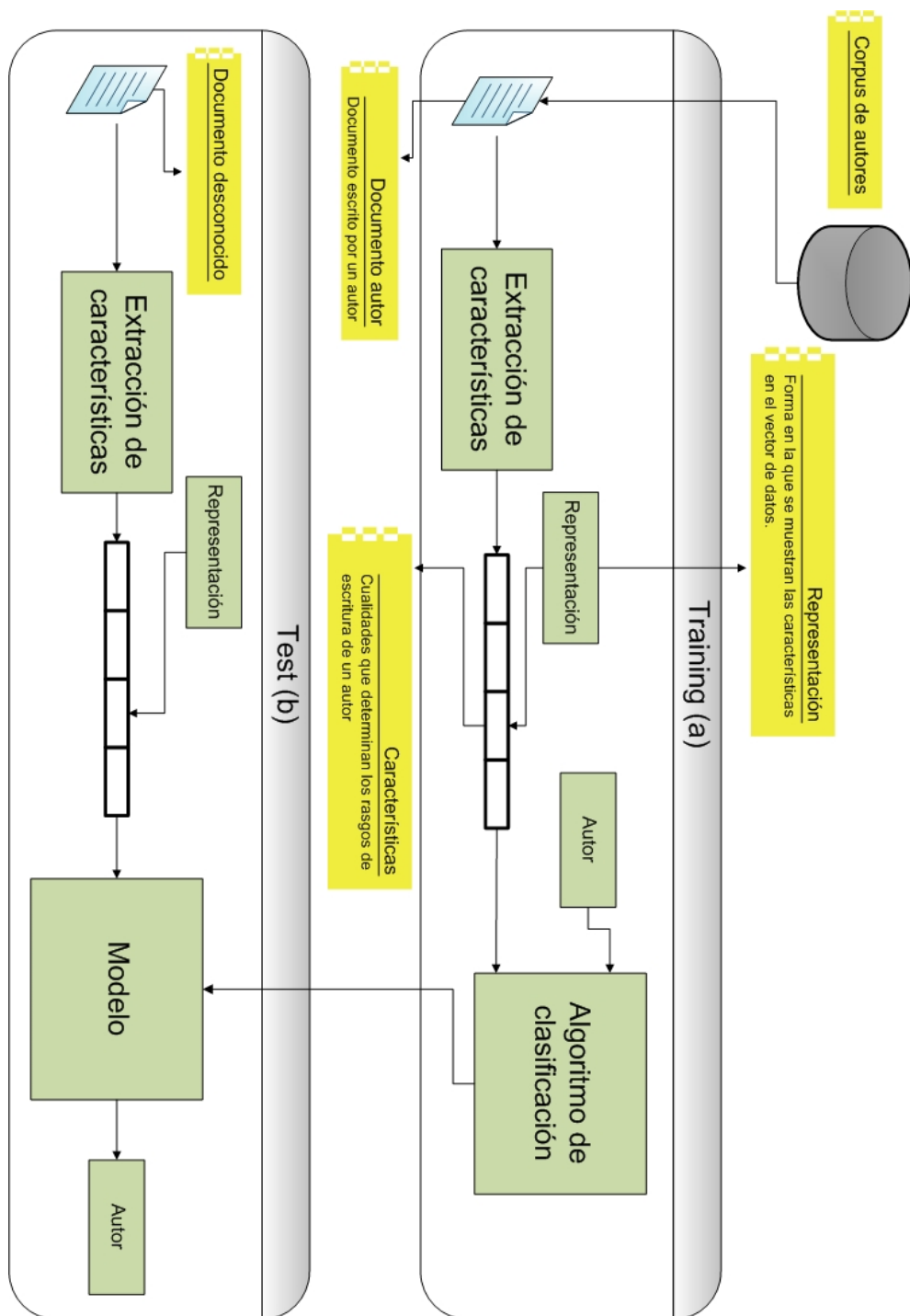


Figura 3.5: Proceso de clasificación de textos.

aplicaciones del mundo real. La representación por medio del Kernel ofrece una solución alternativa a este problema, proyectando la información a un espacio de características de mayor dimensión, el cual aumenta la capacidad computacional de las máquinas de aprendizaje lineal.

3.7. Conclusión del capítulo

En esta sección se describen términos y conceptos relacionados con el desarrollo de esta tesis. En especial aspectos relacionados con la atribución de autoría, la definición formal de ésta, así como la importancia y el uso de uno de los elementos más importantes, las características. Como principal aportación de este capítulo se presentan y se describen los principales tipos de características y como estas se aplican en distintos aspectos del lenguaje para determinar propiedades de un autor.

METODOLOGÍA DE SOLUCIÓN

Computers make it easier to do a lot of things, but most of the things they make it easier to do don't need to be done.

Andy Rooney

En este capítulo se establece el marco metodológico de la presente investigación, que atiende al propósito de encontrar el autor real de un documento de origen desconocido. Para este fin, se siguen varias aproximaciones para resolver el problema, las cuales incluyen el tipo de características a usar y la correcta representación de éstas. Finalmente, en este capítulo se explican los conceptos afines al desarrollo de cada aproximación haciendo énfasis en los pasos utilizados para resolver el problema de forma óptima.

4.1. Primera aproximación: Palabras

La aproximación tiene como objetivo mostrar la metodología usada para detectar los autores reales de un conjunto de documentos, considerando la representación tanto de los documentos de entrenamiento, como los de prueba utilizando todas las palabras presentes en los documentos, así como el uso de un modelo de clasificación supervisado.

4.1.1. Características escogidas

En esta aproximación se toman las palabras originales contenidas en cada documento del conjunto de datos con un mínimo de pre-procesamiento, el cual incluye la eliminación de palabras cerradas, símbolos de puntuación y etiquetas de XML.

4.1.2. Representación de las características

Sea $(x_1, x_2, x_3, \dots, x_n)$ el conjunto de características seleccionadas para representar a los documentos. Cada documento D es representado considerando la frecuencia de aparición de cada una de las palabras (x_n) , es decir, se usa el modelo de bolsa de palabras ponderadas para representar a cada documento [13].

Para la fase de entrenamiento (*training*) se utiliza un vector de características como se muestra en la Ecuación 4.1,

$$D = (\underbrace{x_1, x_2, x_3, \dots, x_n}_{\text{Palabras del documento}}, C) \quad (4.1)$$

donde C es el atributo clasificador del documento, es decir el autor al cual le pertenece dicho documento.

Para la fase de prueba (*test*) se utiliza un vector de características como se muestra en la Ecuación 4.2,

$$D = (\underbrace{x_1, x_2, x_3, \dots, x_n}_{\text{Palabras del documento}}) \quad (4.2)$$

donde no existe un atributo clasificador ya que el documento es de procedencia anónima.

4.1.3. Proceso de clasificación

A continuación se muestran los pasos (ver Figura 4.1) para la correcta clasificación de un conjunto de documentos de procedencia anónima:

1. Extracción de las palabras asociadas a los documentos tanto del *test* como del *training*.
2. Creación del modelo de clasificación a partir del *training* (se escoge un algoritmo de clasificación como los mostrados en el Capítulo 3).
3. Prueba de los documentos del *test* a partir del modelo de clasificación.
4. Validación de los resultados a partir del *gold standard* [13].

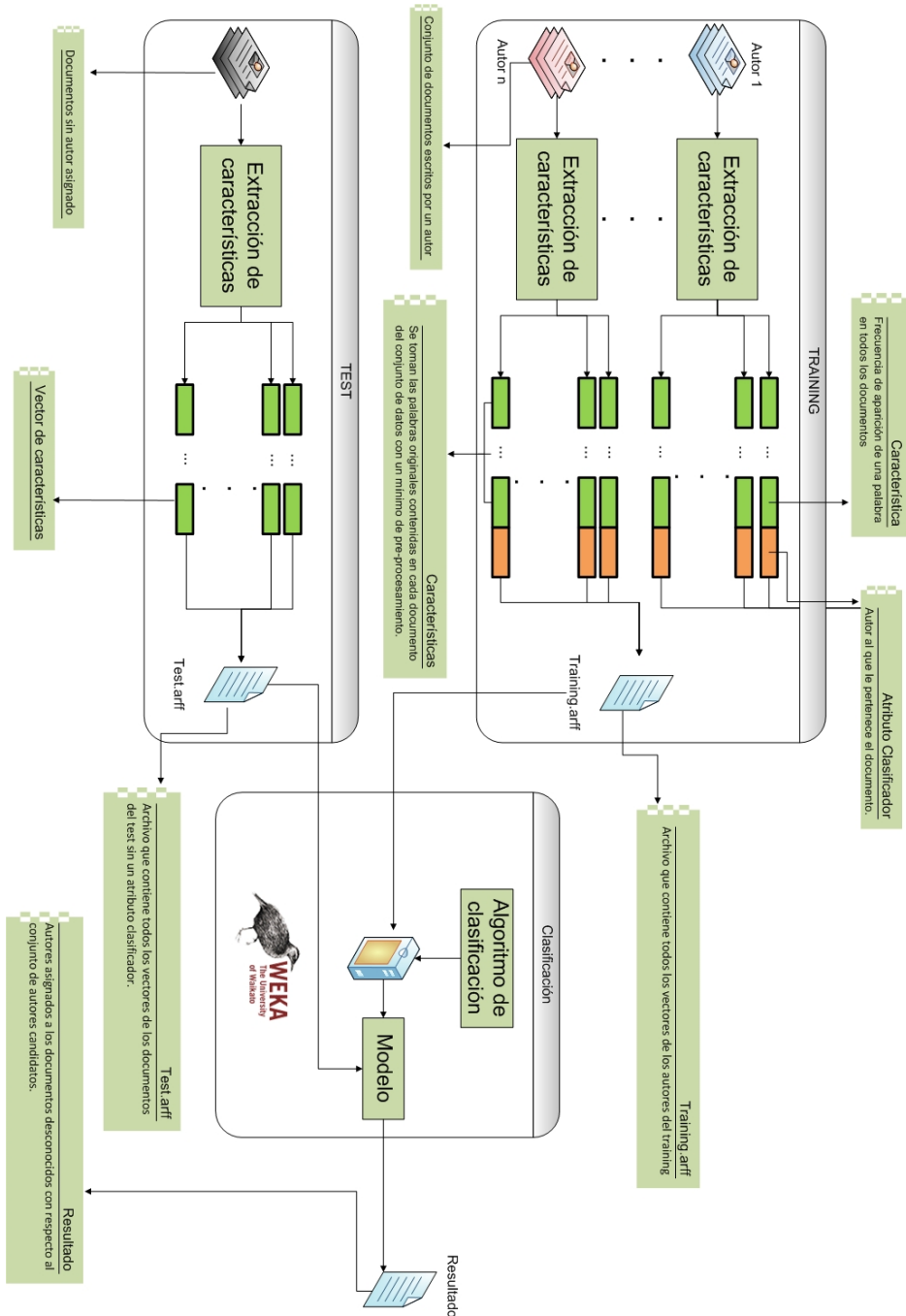


Figura 4.1: Esquema de la primera aproximación.

4.2. Segunda aproximación: Extracción y combinación de características

Esta aproximación tiene como objetivo mostrar la metodología usada para detectar los autores reales de un conjunto de documentos, por medio de la extracción de diferentes tipos de características (lexico-sintáctico-semánticas) tanto a nivel de oración como a nivel de caracter.

4.2.1. Características escogidas

A continuación se presentan y describen las características usadas en esta aproximación.

- Características a nivel de oración:
 - Palabras con frecuencia igual a 1. Son aquellas palabras que aparecen una sola vez en cada documento que pertenece al mismo autor.
 - Palabras con frecuencia igual a 2. Son aquellas palabras que aparecen dos veces en cada documento que pertenece al mismo autor.
 - Palabras más frecuentes. Las 10 palabras que aparecen con mayor frecuencia en los documentos de un autor dado.
 - Colocaciones. Los pares de palabras (bigramas) que siempre aparecen juntos en un documento dado¹.
 - ej. $\{(button, pressed), (federal, court), (south, america)\}$
 - Prefijos de palabras. La subcadena que antecede a la base léxica.
 - ej. $ad \rightarrow \{\mathbf{advance}, \mathbf{adjunct}, \mathbf{adulterate}\}$
 - ej. $de \rightarrow \{\mathbf{descend}, \mathbf{despair}, \mathbf{depend}\}$
 - Sufijos de palabras. La subcadena que precede a la base léxica.
 - ej. $est \rightarrow \{\mathbf{finest}, \mathbf{toughest}, \mathbf{biggest}\}$
 - ej. $tion \rightarrow \{\mathbf{configuration}, \mathbf{consideration}\}$
 - Trigramas de palabras. Todas las subsecuencias de tres palabras de cada documento.

¹Se utilizó el paquete NLTK(<http://nltk.org/>) para obtener estas colocaciones.

- ej. Thanks i'll see you tomorrow
 $\{(thanks, i'll, see), (i'll, see, you), (see, you, tomorrow)\}$
- Palabras cerradas. Las palabras con poco contenido semántico, como artículos, pronombres, preposiciones, etc.
 - ej. $\{and, the, but, did, \dots\}$
- Número de palabras. Se cuantifica para cada documento el número de palabras que posee.
- Número de palabras por oración. Se cuantifica para cada documento el número de palabras que posee cada oración.
- Número de oraciones. Se cuantifica cuantas oraciones posee cada documento.
- Número de palabras capitalizadas. Se cuantifica el número de palabras que tienen como primera letra una letra capital (mayúscula).
- Espacios utilizados en la escritura de documentos. Se cuantifica el número de espacios en blanco dentro de un documento.
- Trigramas de puntuación. Todas las secuencias de 3 signos de puntuación que aparecen a lo largo de todo el documento.
- Trigramas de las categorías gramaticales del documento. Todas las secuencias de 3 categorías² que aparecen a lo largo de todo el documento.
 - ej. she:PRP drove:VBD a:DT silver:NN pt:NN cruiser:NN
 - ◊ PRP: Pronombre posesivo
 - ◊ VBD: Verbo en pasado
 - ◊ DT: Determinante
 - ◊ NN: Sustantivo $\{(PRP, VBD, DT), (VBD, DT, NN), (DT, NN, NN), (NN, NN, NN)\}$
- Palabras al inicio del documento. Frecuencia de aparición de aquellas palabras que aparecen al inicio de los documentos de la colección.
- Palabras al final del documento. Frecuencia de aparición de aquellas palabras que aparecen al final de los documentos de la colección.

²Se utilizó POS-tagger de NLTK(<http://nltk.org/>).

- Sinónimos³. Frecuencia de aparición de todos los sinónimos asociados a cada uno de las palabras de la colección de datos.
 - ej. $palabra = woods \xrightarrow{expansión} \{sentido\ 1, sentido\ 2, sentido\ 3\}$
 $woods(sentido\ 1) \rightarrow \{forest\}$
 $woods(sentido\ 2) \rightarrow \{timber\}$
 $woods(sentido\ 3) \rightarrow \{woodland, timberland\}$
 - Hipónimos³. Frecuencia de aparición de todos los hipónimos asociados a cada uno de las palabras de la colección de datos.
 - ej. $palabra = woods \xrightarrow{expansión} \{sentido\ 1, sentido\ 2, sentido\ 3\}$
 $woods(sentido\ 1) \rightarrow \{rain\ forest\}$
 $woods(sentido\ 2) \rightarrow \{jungle\}$
 $woods(sentido\ 3) \rightarrow \{old\ growth, virgin\ forest\}$
 - Hiperónimos³. Frecuencia de aparición de todos los hiperónimos asociados a cada uno de las palabras de la colección de datos.
 - ej. $palabra = woods \xrightarrow{expansión} \{sentido\ 1, sentido\ 2, sentido\ 3\}$
 $woods(sentido\ 1) \rightarrow \{vegetation, flora, botany\}$
 $woods(sentido\ 2) \rightarrow \{plant\ material, plant\ substance\}$
 $woods(sentido\ 3) \rightarrow \{wind\ instrument, wind\}$
- Características a nivel de caracter:
- Palabras longitud 10. Todas aquellas palabras formadas con 10 letras(con repetición de letras).
 - ej. $\{commercial, difference, comforting, assistance\}$
 - Combinación de vocales. Se elimina de cada palabra las consonantes y se considera la combinación de vocales restantes como una característica (se unen todas las vocales repetidas en cada combinación).
 - ej. $influential \rightarrow iueia \rightarrow iuea$
 - ej. $internet \rightarrow iee \rightarrow ie$
 - ej. $transport \rightarrow ao \rightarrow ao$
 - Permutación de vocales. Se elimina de cada palabra las consonantes y se consideran todas las combinaciones de vocales, cada combinación como una característica.

³Se utilizó WordNet del paquete NLTK(<http://nltk.org/>).

- ej. *influential* → *iueia*
- ej. *internet* → *iee*
- ej. *transport* → *ao*
- Letras del abecedario. Se considera el número de veces que aparece cada letra del abecedario en un documento dado.
- Trigramas de carácter. Se cuantifican todas las subsecuencias de tres caracteres en un documento dado.
 - ej. *we would*
 $\{(w, e, w), (e, w, o), (w, o, u), (o, u, l), (u, l, d)\}$
- Número de caracteres por oración. Se cuantifica para cada documento el número de caracteres que posee cada oración.
- Signos de puntuación. Se cuantifican todos los delimitadores de frases y párrafos que establecen una jerarquía sintáctica.

4.2.2. Representación de las características

Se utiliza el modelo de bolsa de palabras ponderadas (ver la sección 4.1.2) para representar a cada documento D .

Para la combinación de características suponga que se tienen n características como los que se muestran en la Ecuación 4.3,

$$\begin{aligned}
 \text{Característica 1} & \begin{cases} D_1 = (x_1, x_2, x_3, \dots, x_a) \\ \vdots \\ D_m = (y_1, y_2, y_3, \dots, y_a) \end{cases} \\
 \text{Característica 2} & \begin{cases} D_1 = (x_1, x_2, x_3, \dots, x_b) \\ \vdots \\ D_m = (y_1, y_2, y_3, \dots, y_b) \end{cases} \\
 & \vdots \\
 \text{Característica } n & \begin{cases} D_1 = (x_1, x_2, x_3, \dots, x_d) \\ \vdots \\ D_m = (y_1, y_2, y_3, \dots, y_d) \end{cases}
 \end{aligned} \tag{4.3}$$

donde a , b y d pueden tener el mismo tamaño o diferente, dependiendo del tipo de característica.

Para la fase de entrenamiento (*training*) se utiliza un vector formado con características combinadas como se muestra en la Ecuación 4.4,

$$D_{Training} = (\underbrace{x_1, x_2, x_3, \dots, x_s}_{\text{Característica 1}}, \dots, \underbrace{y_1, y_2, y_3, \dots, y_m}_{\text{Característica } n}, C) \quad (4.4)$$

donde C es el mismo autor de las n características. Para la fase de prueba (*test*) se utiliza también un vector formado con características combinadas como se muestra en la Ecuación 4.5,

$$D_{Test} = (\underbrace{x_1, x_2, x_3, \dots, x_s}_{\text{Característica 1}}, \dots, \underbrace{y_1, y_2, y_3, \dots, y_m}_{\text{Característica } n}) \quad (4.5)$$

donde no existe un atributo clasificador C , ya que el documento es de procedencia anónima.

4.2.3. Proceso de clasificación

A continuación se muestran los pasos (ver Figura 4.2) para la correcta clasificación de un conjunto de documentos de procedencia anónima:

1. Extracción de las palabras asociadas a los documentos tanto del *test* como del *training*.
2. Combinación (de ser necesario ya que puede ser una característica única) de una o varias características en un sólo vector representativo del autor
3. Creación del modelo de clasificación a partir del *training* (se escoge un algoritmo de clasificación como los mostrados en el Capítulo 3).
4. Prueba de los documentos del *test* a partir del modelo de clasificación.
5. Validación de los resultados a partir del *gold standard*.

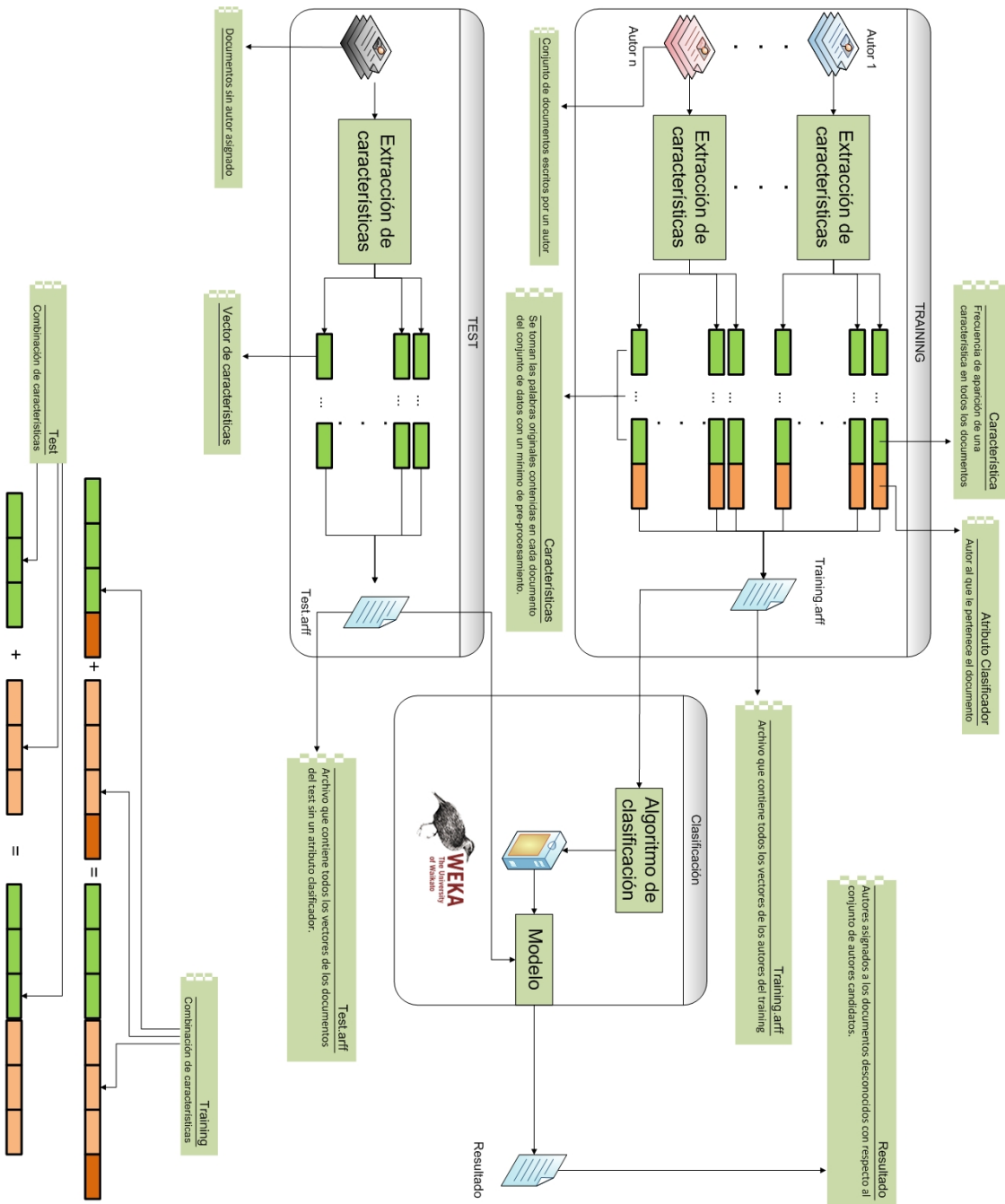


Figura 4.2: Esquema de la segunda aproximación.

4.3. Tercera aproximación: Operaciones de conjuntos

La aproximación tiene como objetivo mostrar la metodología usada para detectar los autores reales de un grupo de documentos, considerando el uso de operaciones sobre conjuntos para combinar y delimitar las características a usar en el proceso de clasificación.

4.3.1. Características escogidas

Se toman las palabras originales contenidas en cada documento del conjunto de datos (tanto del training como del test), eliminando los símbolos de puntuación y las palabras cerradas. Posteriormente se extraen diferentes tipos de características (lexico-sintáctico-semánticas), tanto a nivel de oración como a nivel de carácter (ver la Sección 4.2.1).

4.3.2. Representación de las características

Se utiliza el modelo de bolsa de palabras ponderadas (ver la Sección 4.1.2) para representar a cada documento. Para la combinación de características se utilizan las siguientes operaciones de conjuntos:

- Unión

$$Vector_1 \cup Vector_2 = \{x | x \in Vector_1 \vee x \in Vector_2\} \quad (4.6)$$

- Intersección

$$Vector_1 \cap Vector_2 = \{x | x \in Vector_1 \wedge x \in Vector_2\} \quad (4.7)$$

- Diferencia simétrica

$$Vector_1 \Delta Vector_2 = \{x | x \in Vector_1 \cup Vector_2 \wedge x \notin Vector_1 \cap Vector_2\} \quad (4.8)$$

- Complemento

$$\overline{Vector} = \{x | x \in v \wedge x \notin Vector\} \quad (4.9)$$

Donde v es el conjunto de palabras sin extracción de características

- Diferencia

$$Vector_1 - Vector_2 = \{x | x \in Vector_1 \wedge x \notin Vector_2\} \quad (4.10)$$

Para la fase de entrenamiento (*training*) se utiliza un vector combinado de características como se muestra en la Ecuación 4.11,

$$D_{Training} = (\underbrace{x_1, x_2, x_3, \dots, x_s}_{\text{Característica 1}}, \dots, \underbrace{y_1, y_2, y_3, \dots, y_m}_{\text{Característica n}}, C) \quad (4.11)$$

donde C es el mismo autor con las n características consideradas. Para la fase de prueba (*test*) se utiliza un vector combinado de características como se muestra en la Ecuación 4.12,

$$D_{Test} = (\underbrace{x_1, x_2, x_3, \dots, x_s}_{\text{Característica 1}}, \dots, \underbrace{y_1, y_2, y_3, \dots, y_m}_{\text{Característica n}}) \quad (4.12)$$

donde no existe un atributo clasificador C , ya que el documento es de procedencia anónima.

4.3.3. Proceso de clasificación

A continuación se muestran los pasos (ver Figura 4.3) para la correcta clasificación de un conjunto de documentos de procedencia anónima:

1. Extracción de las características asociadas a los documentos tanto del *test* como del *training*.
2. Combinación de una o varias características en un sólo vector representativo del autor, por medio de las operaciones de conjuntos (union, intersección, diferencia, diferencia simétrica y complemento)
3. Creación del modelo de clasificación a partir del *training* (se escoge un algoritmo de clasificación como los mostrados en el Capítulo 3).
4. Prueba de los documentos del *test* a partir del modelo de clasificación.
5. Validación de los resultados a partir del *gold standard*.

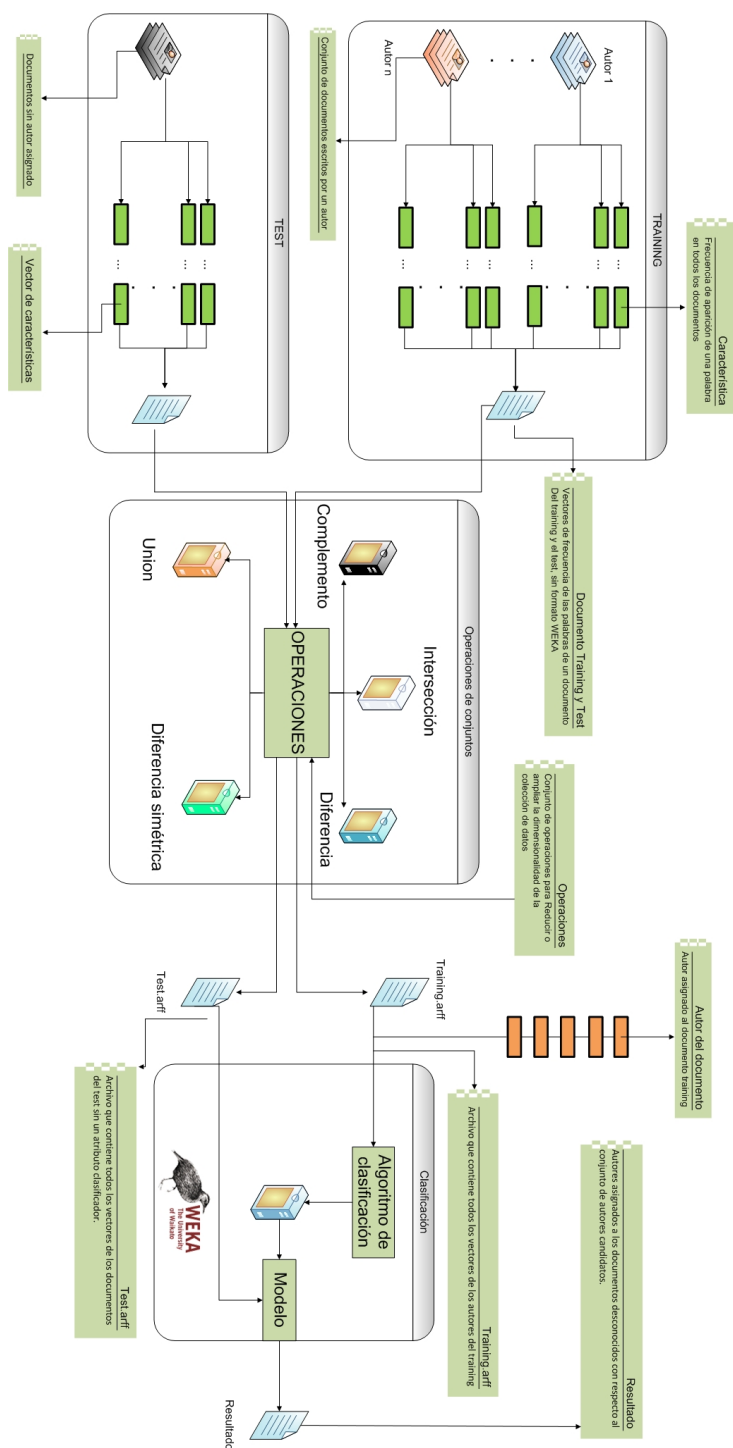


Figura 4.3: Esquema de la tercera aproximación.

4.4. Cuarta aproximación: Estadística

A continuación se muestran las medidas estadísticas utilizadas.

4.4.1. Características escogidas

Para la extracción de las características estadísticas se toman las palabras originales contenidas en cada documento del conjunto de datos (tanto del training como del test), eliminando los símbolos de puntuación. Se representan diferentes características usando una bolsa de palabras ponderada (ver 4.1.2 y 4.2.1), los vectores obtenidos son usados para determinar las siguientes medidas estadísticas:

- Media aritmética($\bar{X}$): La media aritmética o promedio de un conjunto de n mediciones

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} \quad (4.13)$$

- Varianza(S^2): Es la media de las desviaciones cuadráticas de una variable aleatoria, con relación al valor medio de ésta.

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1} \quad (4.14)$$

- Desviación estándar(S): Mide el grado de separación de los datos.

$$S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}} \quad (4.15)$$

- Asimetría o deformación(A_s): Medida que permite identificar si los datos se distribuyen de manera uniforme alrededor del punto central (media aritmética).

$$A_s = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{S} \right)^3 \quad (4.16)$$

- Curtosis o apuntamiento (A_p): Mide la mayor o menor cantidad de datos que se agrupan en torno a la moda.

$$A_p = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{S} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)} \quad (4.17)$$

- Coeficiente de variación(CV): Es la relación entre la desviación estándar típica de una muestra y su media aritmética.

$$CV = \frac{S}{\bar{X}} \quad (4.18)$$

- Moda(M): Es la categoría que ocurre con mas frecuencia, o el valor de x que ocurre más veces.
- Rango(R): Se define como la diferencia entre las mediciones máxima y mínima de un conjunto de mediciones.

Donde n es la longitud del vector de características y X_i es la frecuencia de aparición de una característica en todos los documentos de un autor específico.

4.4.2. Representación de las características

Cada documento D en el conjunto de datos es representado como se muestra en la Ecuación 4.19

$$\begin{aligned} D_1 &= (x_1, x_2, \dots, x_s, C_1) \xrightarrow{\text{Transformación}} (\bar{X}, S^2, S, A_s, A_p, CV, M, R, C_1) \\ &\vdots \\ D_n &= (y_1, y_2, \dots, y_m, C_n) \xrightarrow{\text{Transformación}} (\bar{X}, S^2, S, A_s, A_p, CV, M, R, C_n) \end{aligned} \quad (4.19)$$

Para la fase de entrenamiento(*training*)se utiliza un vector combinado de características estadísticas como se muestra en la Ecuación 4.20,

$$D_{Training} = \underbrace{(\bar{X}, S^2, S, A_s, A_p, CV, M, R)}_{\text{Característica estadística 1}}, \dots, \underbrace{(\bar{X}, S^2, S, A_s, A_p, CV, M, R, C)}_{\text{Característica estadística } n} \quad (4.20)$$

donde C es el mismo autor de las n características. Para la fase de prueba (*test*) se utiliza un vector combinado de características estadísticas como se muestra en la Ecuación 4.21,

$$D_{Test} = (\underbrace{\overline{X}, S^2, S, A_s, A_p, CV, M, R, \dots}_{\text{Característica estadística 1}}, \dots, \underbrace{\overline{X}, S^2, S, A_s, A_p, CV, M, R}_{\text{Característica estadística } n}) \quad (4.21)$$

donde no existe un atributo clasificador C , ya que el documento es de procedencia anónima.

4.4.3. Proceso de clasificación

A continuación se muestran los pasos (ver Figura 4.4) para la correcta clasificación de un conjunto de documentos de procedencia anónima:

1. Extracción de las palabras asociadas a los documentos tanto del *test* como del *training*.
2. Representación de las palabras usando una bolsa de palabras ponderada.
3. Transformación del vector de características (punto anterior) para la obtención de medidas estadísticas.
4. Combinación (de ser necesario ya que puede ser una característica única) de una o varias características estadísticas en un sólo vector representativo del autor.
5. Creación del modelo de clasificación a partir del *training* (se escoge un algoritmo de clasificación como los mostrados en el Capítulo 3).
6. Prueba de los documentos del *test* a partir del modelo de clasificación.
7. Validación de los resultados a partir del *gold standard*.

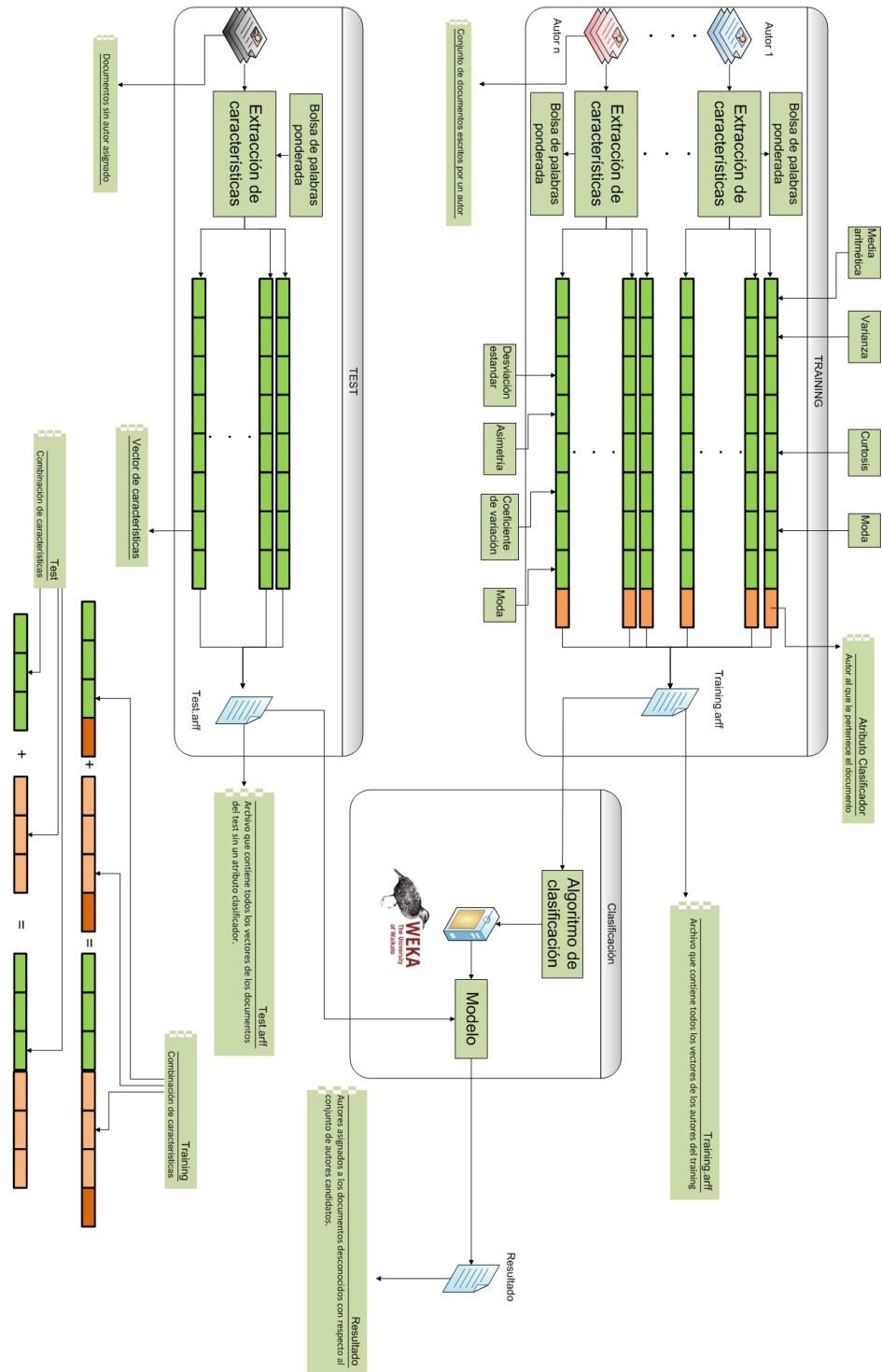


Figura 4.4: Esquema de la cuarta aproximación.

4.5. Quinta aproximación: Extracción de características utilizando representaciones basadas en grafos

La aproximación tiene como objetivo mostrar la metodología usada para detectar los autores reales de un conjunto de documentos, considerando el uso de grafos para representar los textos asociados a ciertos autores. El interés de esta forma de interpretación del conjunto de datos es usar una representación basada en grafos que contemple la secuencia sintáctica de las oraciones de los textos. De manera adicional, se busca considerar las etiquetas de las partes de la oración como características estilométricas.

4.5.1. Características escogidas

En esta aproximación se toman las palabras originales contenidas en cada documento del conjunto de datos con un mínimo de pre-procesamiento, el cual incluye la eliminación de palabras cerradas y símbolos de puntuación. Dichas palabras son representadas usando un grafo para posteriormente obtener las subestructuras comunes (conjunto de palabras representativas del autor que son representadas usando una bolsa de palabras ponderada para obtener su frecuencia de aparición en el conjunto de datos), del conjunto de grafos por medio de la herramienta de minería de datos basada en grafos SUBDUE⁴ [46].

4.5.2. Representación de las características

En esta aproximación se usa una representación basada en la notación gSpan [47] en la cual un grafo es representado como una 4-tupla $G = (V, E, L, l)$ donde V es un conjunto (no vacío) de vértices, E es un conjunto de aristas de la forma $E \subseteq V \times V$, L es un conjunto de etiquetas y l es una función que asigna una etiqueta a un par de vértices asociados.

A continuación se presentan las 2 representaciones usadas en esta aproximación:

⁴<http://ailab.wsu.edu/subdue/>

Representación basada en secuencias de palabras categorizadas gramaticalmente

Cada oración a representar dentro de un documento es etiquetada con su correspondiente parte de la oración utilizando la herramienta TreeTagger⁵. Cada palabra es truncada, usando el Stemmer de Porter⁶, para considerar solamente la raíz de cada término. En esta representación se considera cada vértice como una palabra truncada, y cada arista como la categoría gramatical de la palabra precedente, manteniendo la secuencia de cada una de las oraciones a representar. Las categorías gramaticales contempladas se pueden ver en la Tabla 4.1.

Tabla 4.1: Descripción de las categorías gramaticales

Categoría gramatical	Descripción
PRP	Pronombre posesivo
JJ	Adjetivo
VBN	Verbo pasado participio
WDT	Determinante
NN	Sustantivo
CD	Número
RB	Adverbio
NNS	Sustantivo singular
CC	Conjunción
RBR	Adverbio comparativo
MD	Modal
JJR	Adjetivo comparativo
VBG	Verbo presente participio
VBD	Verbo pasado
VBP	Verbo presente, no es 3era persona del singular
VBZ	Verbo presente, 3era persona del singular
FW	Palabra desconocida
PRP	Pronombre personal
VB	Verbo en forma base
NNP	Sustantivo plural
IN	Preposición y conjuncion
JJS	Adjetivo superlativo
PDT	Determinante

Por ejemplo, considere la oración: **second qualifier long road leading 1998 world cup**, entonces el grafo asociado con la representación propuesta quedaría tal y como se muestra en la Figura 4.5.

⁵<http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>

⁶<http://tartarus.org/~martin/PorterStemmer/>

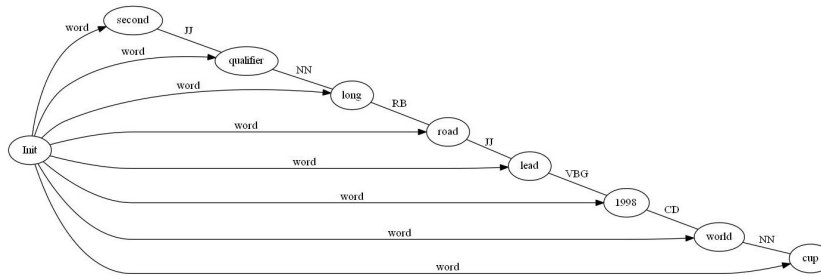


Figura 4.5: Representación basada en secuencias de palabras categorizadas gramaticalmente.

Representación utilizando un modelo de estrella

Al igual que para la representación anterior se etiqueta cada palabra de la oración y posteriormente se trunca, utilizando las mismas herramientas comentadas con anterioridad. Se separa el documento por oraciones. Se asocia un vértice a cada palabra de la oración y a su respectiva categoría gramatical. Estos vértices quedan conectados por medio de una arista. A continuación se conectan todas las palabras que conforman la oración, así como todas las categorías gramaticales, respetando el mismo orden que poseen las palabras dentro de la oración. Se crea un vértice inicial (*init*), que se conecta tanto con cada palabra de la respectiva oración, como con cada categoría gramatical, por lo tanto, por cada oración de cada documento se tiene un grafo.

Por ejemplo, si nuevamente se considera la oración: **second qualifier long road leading 1998 world cup**, entonces el grafo asociado con la representación propuesta quedaría tal y como se muestra en la Figura 4.6

4.5.3. Proceso de clasificación versión 1

A continuación se muestran los pasos (ver Figura 4.8) para la correcta clasificación de un conjunto de documentos de procedencia anónima:

1. Pre-procesamiento de las palabras asociadas a los documentos tanto del *test* como del *training*.
2. Creación de un grafo representativo por documento, a partir del conjunto de datos del *training*.

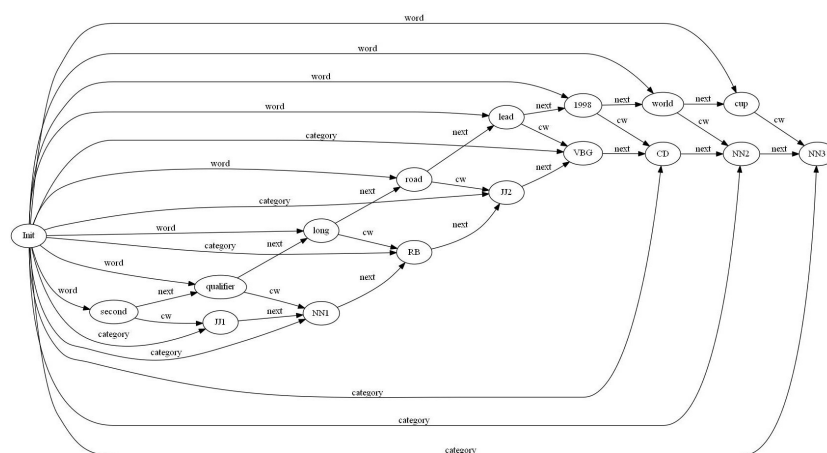


Figura 4.6: Representación con un modelo de estrella.

3. Uso de la herramienta SUBDUE para extraer los patrones (palabras) relevantes.
4. Creación del modelo de clasificación a partir de las palabras obtenidas en SUBDUE (se escoge un algoritmo de clasificación como los mostrados en el Capítulo 3).
5. Prueba de los documentos del *test* a partir del modelo de clasificación.
6. Validación de los resultados a partir del *gold standard*.

4.5.4. Proceso de clasificación versión 2

En los siguientes pasos se muestra (ver Figura 4.9) la forma de clasificar un conjunto de documentos de procedencia anónima:

1. Pre-procesamiento de las palabras asociadas a los documentos tanto del *test* como del *training*.
2. Creación de un grafo representativo por cada oración presente en los documentos tanto del *test* como del *training* (ver Figura 4.7).
3. Uso de la herramienta SUBDUE, combinando un grafo del *training* (muestra positiva) y un grafo del *test* (muestra negativa) para todos los grafos del conjunto de datos.

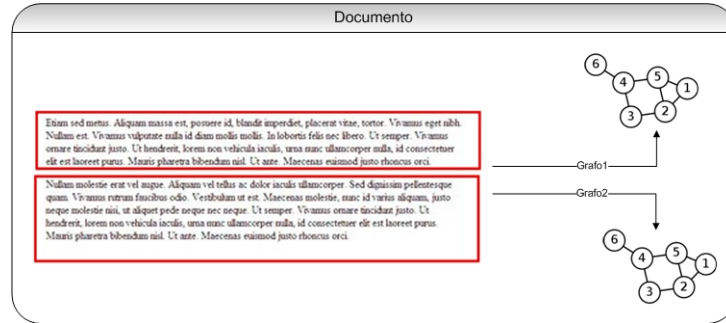


Figura 4.7: Creación de un grafo por cada oración

4. Los grafos que posean patrones comunes (isomorfismo parcial) tanto en el *test* como en el *training* son cuantificados de acuerdo a los siguientes parámetros:
 - Número de vértices
 - Número de aristas
 - Longitud total del grafo
5. Dado los resultados estadísticos obtenidos en el punto anterior se escoge un algoritmo de clasificación (como los mostrados en el Capítulo 3).
6. Prueba de los documentos del *test* a partir del modelo de clasificación.
7. Validación de los resultados a partir del *gold standard*.

4.6. Sexta aproximación: Extracción de características utilizando una expansión basada en grafos

En esta aproximación se realiza una expansión semántica sobre las 2 representaciones basadas en grafos que contemplan la secuencia sintáctica de las oraciones de los textos asociados a un autor específico (ver la Sección 4.5). La expansión consiste en la extracción, ya sea por separado o en conjunto de todos los sinónimos, hipónimos e hiperónimos (ver la Sección 4.2 en la

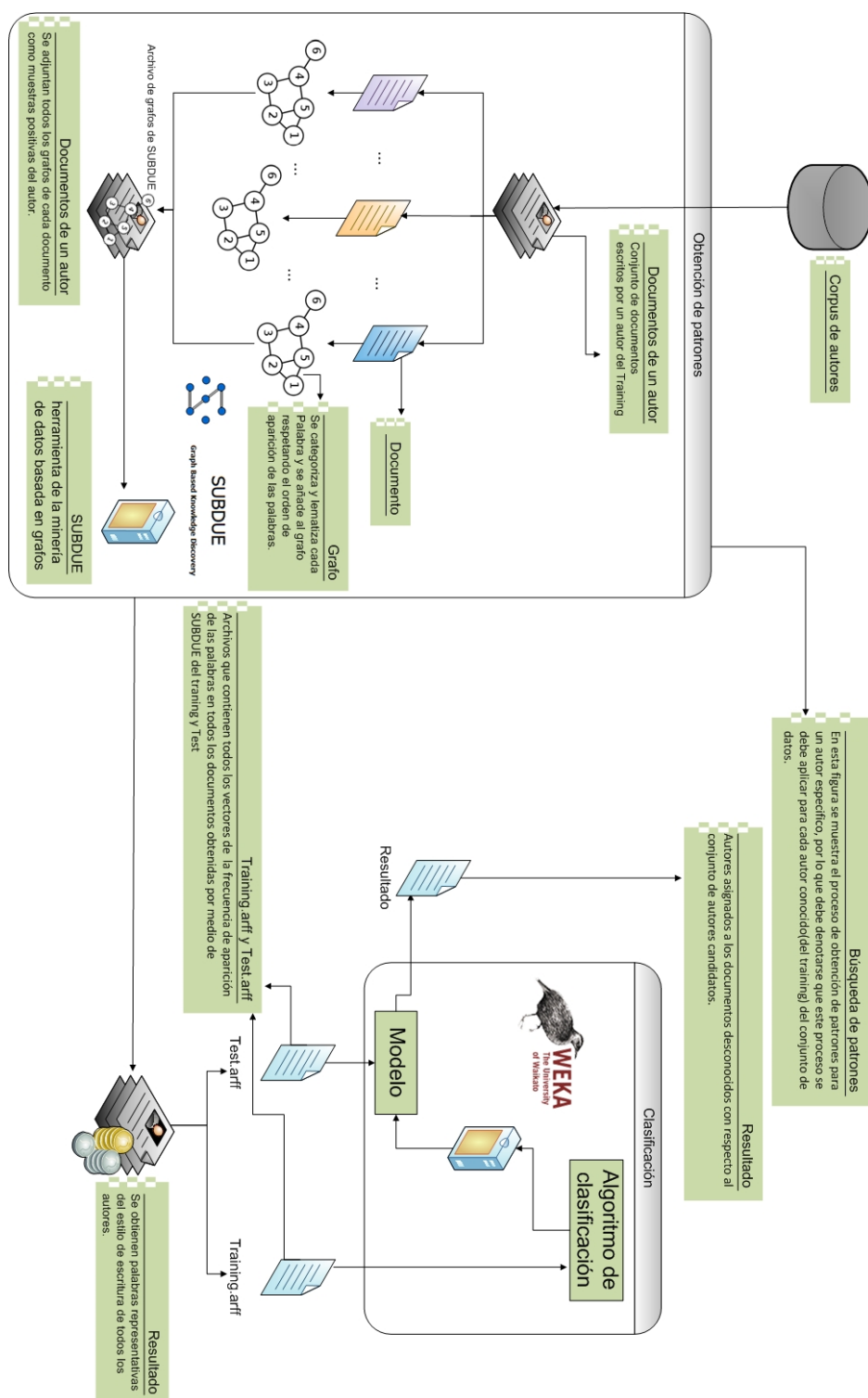


Figura 4.8: Esquema de la quinta aproximación versión 1.

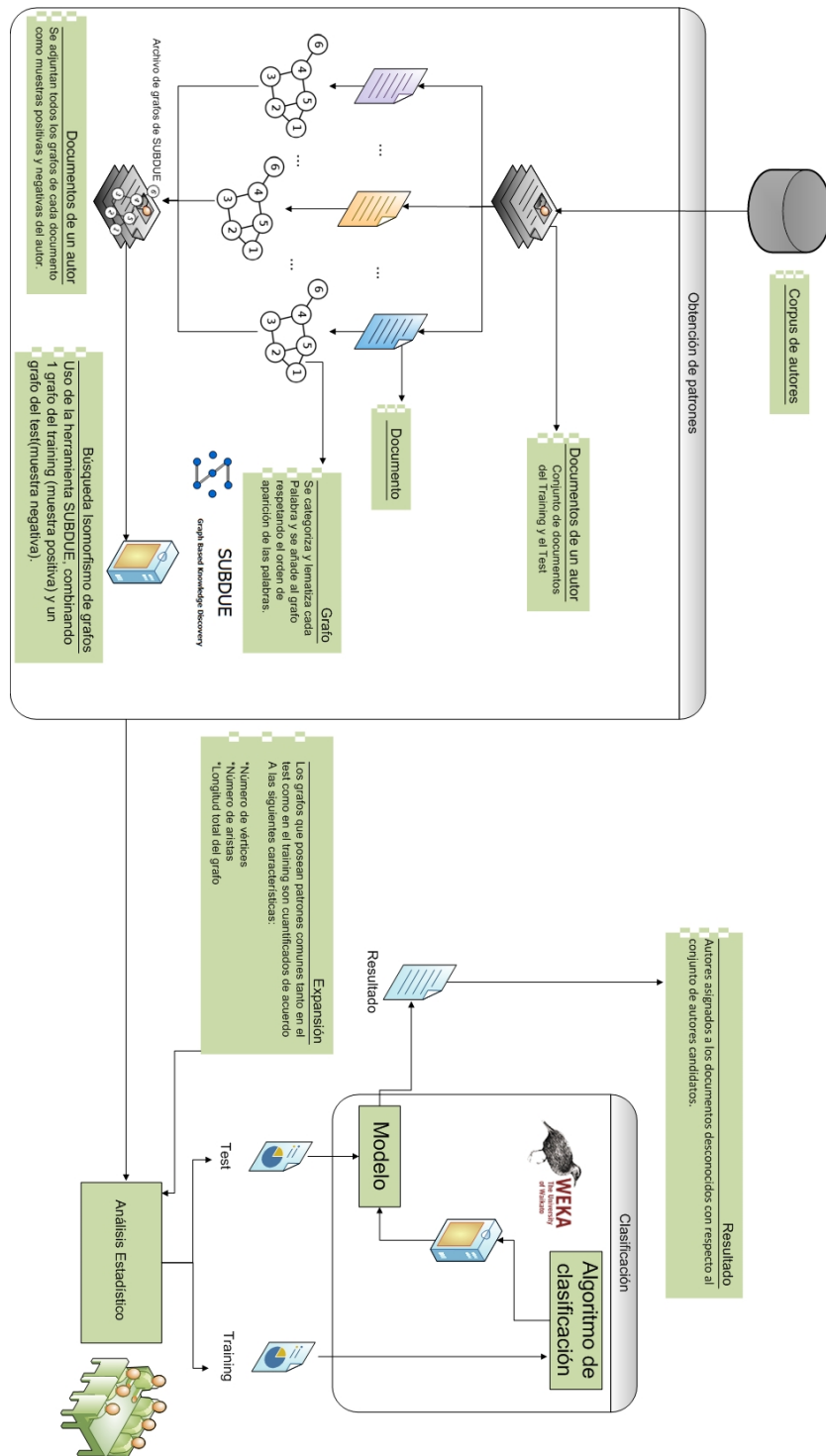


Figura 4.9: Esquema de la quinta aproximación versión 2.

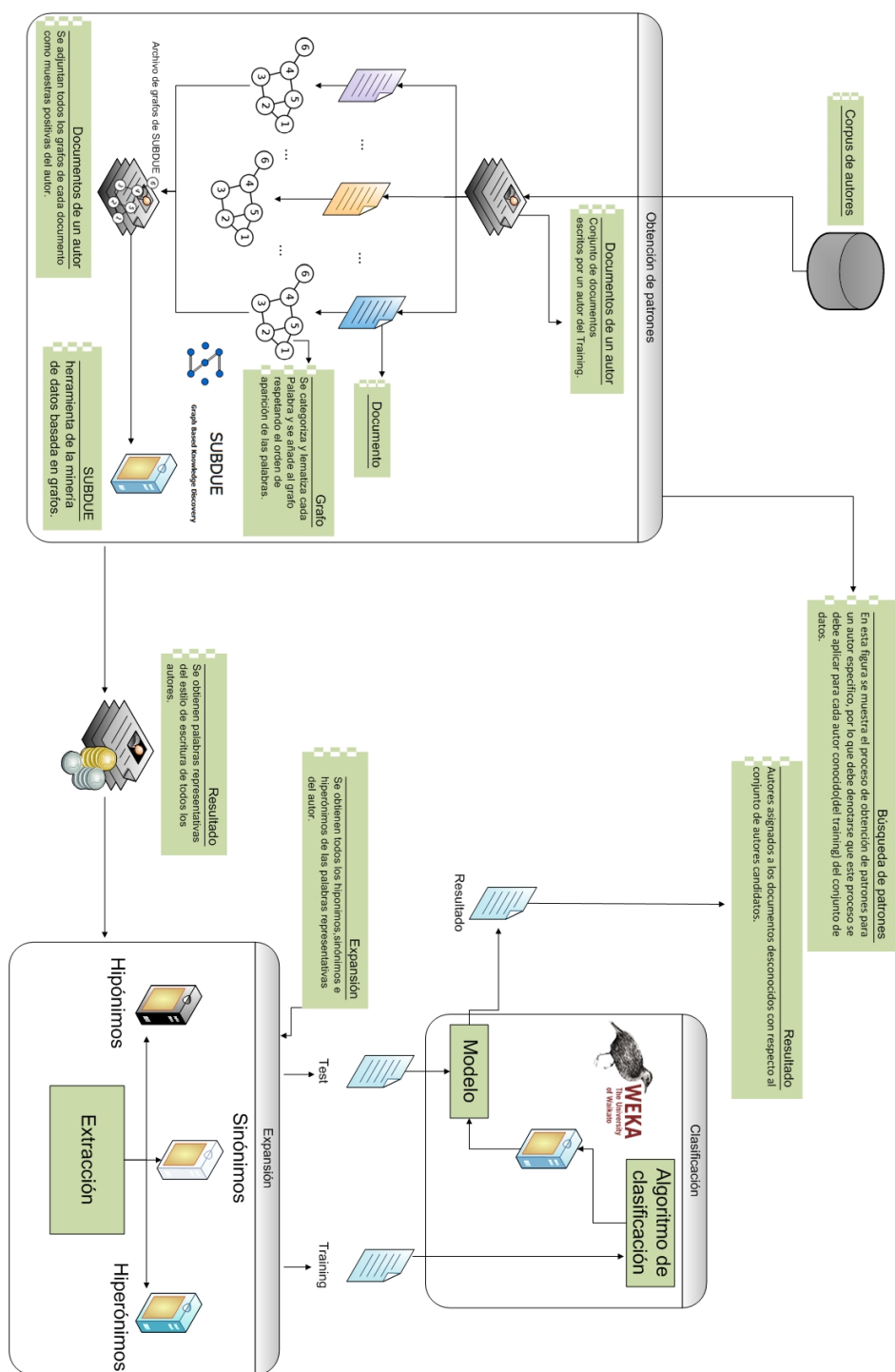


Figura 4.10: Esquema de la sexta aproximación.

página 42) de las palabras obtenidas por medio de la herramienta de minería de datos basada en grafos SUBDUE.

4.6.1. Proceso de clasificación

A continuación se muestran los pasos (ver Figura 4.10) para la correcta clasificación de un conjunto de documentos de procedencia anónima:

1. Pre-procesamiento de las palabras asociadas a los documentos tanto del *test* como del *training*.
2. Creación de un grafo representativo por documento, a partir del conjunto de datos del *training*.
3. Uso de la herramienta SUBDUE para extraer los patrones (palabras) relevantes.
4. Expansión semántica de las palabras obtenidas en el punto anterior para obtener:
 - Sinónimos
 - Hipónimos
 - Hiperónimos
5. Creación del modelo de clasificación a partir de las palabras originales y las palabras obtenidas en la expansión.
6. Prueba de los documentos del *test* a partir del modelo de clasificación.
7. Validación de los resultados a partir del *gold standard*.

4.7. Conclusión del capítulo

En esta sección se muestran y describen conceptos relacionados con el desarrollo de las distintas metodologías propuestas para lograr establecer de manera correcta a que autor le pertenece un documento dado. En especial se hace énfasis en las características extraídas así como en la forma de representarlas.

PRUEBAS Y RESULTADOS

However beautiful the strategy,
you should occasionally look at
the results.

Winston Churchill

En esta sección se presentan y discuten los resultados obtenidos en la tarea de atribución de autoría, se describen las pruebas realizadas sobre cada aproximación propuesta, y se comparan los mejores resultados. En especial, se describen los algoritmos usados para el proceso de clasificación de textos, así como los parámetros ajustados para cada aproximación.

5.1. Conjunto de datos

Para la correcta determinación de las características se trabajó con un conjunto de datos en inglés (véase Capítulo 3 en la Sección 3.5) creado por Efstathios Stamatatos(ccat C10) a partir del RCV1 [44]. En la Tabla 5.1 se muestra el número de documentos para el entrenamiento y el test del conjunto de datos usado para determinar la autoría de un documento de origen desconocido.

Tabla 5.1: Número de documentos para training y test asociado al corpus C10

Conjunto de datos	Training	Test
C10 Stamatatos	500	500

5.2. Primera aproximación

En la primera aproximación (véase el Capítulo 4) se utilizan 2 algoritmos de clasificación para determinar el correcto autor de un documento dado:

- Rocchio. El algoritmo de clasificación Rocchio crea un centroide por cada clase presente en el conjunto de datos (conjunto de autores), el centroide se calcula utilizando la Ecuación 5.1,

$$\overrightarrow{\mu(a)} = \frac{1}{|D_a|} \sum_{d \in D_a} \overrightarrow{v}(d), \quad (5.1)$$

donde D_a es el conjunto de documentos en D cuyo autor es a : $D_a = \{d : \langle d, a \rangle \in D\}$, y $\overrightarrow{v}(d)$ es el vector normalizado de d . En este algoritmo se omite el componente de consulta del algoritmo original (Rocchio relevance feedback) dado que no existe una consulta en la tarea de clasificación de textos así como se hace uso de una representación vectorial usando el esquema de peso TF-IDF [48] y el criterio de clasificación usa la medida de distancia Euclidiana. El proceso usado para la fase de training y test se muestran en el Algoritmo 1 y 2.

- Multinomial Naïve Bayes. Se usa un método probabilístico de aprendizaje supervisado llamado Multinomial Naïve Bayes para determinar la atribución de autoría. La probabilidad de que un documento d escrito por un autor a se calcula utilizando la Ecuación 5.2.,

$$P(a|d) \approx P(a) \prod_{1 \leq k \leq n_d} P(t_k|a) \quad (5.2)$$

donde $P(t_k|a)$ es la probabilidad condicional del k -th término (t_k) de un documento escrito por el autor a . Dado que estamos interesados en encontrar la mejor clase (autor) para un documento, calculamos el máximo a posteriori (MAP) como se muestra en la Ecuación 5.3,

$$a_{map} = \arg \max_{a \in A} P^*(a|d) = \arg \max_{a \in A} P^*(a) \prod_{1 \leq k \leq n_d} P^*(t_k|a) \quad (5.3)$$

donde $P^*(t_k|a)$ es estimado usando el metodo de suavizado de Laplace, como se muestra en la Ecuación 5.4,

$$P^*(t_k|a) = \frac{T_{at_k} + 1}{\sum_{t' \in V} (T_{at'} + 1)} \quad (5.4)$$

donde T_{at_k} es el número de ocurrencias de t_k y $T_{at'}$ representa el número de ocurrencias del término $T_{t'}$ del vocabulario en documentos de entrenamiento de la clase a y V es el vocabulario del corpus. El proceso usado para el *training* y *test* se muestra en los Algoritmos 3 y 4.

Algoritmo 1 Algoritmo Rocchio para la fase de *training*

Entrada: A : Conjunto de autores

Entrada: D : Conjunto de documentos

Salida: Centroides de cada conjunto de documentos de un autor

```

1: para  $a_j \in A$  hacer
2:    $D_j \leftarrow \{d : \langle d, a_j \rangle \in D\}$ 
    $\vec{\mu}_j \leftarrow \frac{1}{|D_a|} \sum_{d \in D_a} \vec{v}(d)$  (vector de frecuencias para el documento  $d$ )
3: fin para
4: devolver  $\{\vec{\mu}_1, \dots, \vec{\mu}_J\}$ 

```

Algoritmo 2 Algoritmo Rocchio para la fase de *test*

Entrada: $\{\vec{\mu}_1, \dots, \vec{\mu}_J\}$: Centroides

Entrada: d : Documento a evaluar

Salida: La mejor clase(autor) para cada documento a identificar

devolver $\arg \min_j |\vec{\mu}_j - \vec{v}(d)|$

Algoritmo 3 Algoritmo Multinomial Naïve Bayes para la fase de *training*

Entrada: A : Conjunto de autores

Entrada: D : Conjunto de documentos

Salida: Condiciones de probabilidad

$V \leftarrow \text{Extrae Vocabulario}(D)$

$N \leftarrow \text{Contar Documentos}(D)$

para $a \in A$ **hacer**

$N_a \leftarrow \text{Contar Documentos Clase}(D, a)$

$\text{prior}[a] \leftarrow \frac{N_a}{N}$

$\text{text}_a \leftarrow \text{Concatenar Texto Clase}(D, a)$

para $k \in V$ **hacer**

$T_{at_k} \leftarrow \text{Contar Tokens Termino}(\text{text}_a, k)$

fin para

para $k \in V$ **hacer**

$\text{condprob}[t][a] \leftarrow \frac{T_{at_k} + 1}{\sum_{t' \in V} (T_{at'} + 1)}$

fin para

fin para

devolver $V, \text{prior}, \text{condprob}$

Algoritmo 4 Algoritmo Multinomial Naïve Bayes para la fase de *test*

Entrada: A : Conjunto de autores
Entrada: V : Vocabulario
Entrada: $prior, condprob$: Condiciones de probabilidad
Entrada: d : Documento a evaluar
Salida: La mejor clase(autor) para cada documento a identificar
 $W \leftarrow ExtraeTokensDocumento(V, d)$
para $a \in A$ **hacer**
 $score[a] \leftarrow log(prior[a])$
 para $t \in W$ **hacer**
 $score[a] += log(condprob[t][a])$
 fin para
fin para
devolver $argMax_{a \in A} score[a]$

Los resultados obtenidos utilizando el algoritmo de clasificacion Rocchio y Multinomial Naïve Bayes sobre el conjunto de características escogidas se muestran en la Tabla 5.2 y en la Figura 5.1.

Algoritmo	Número de características (Vocabulario)	C10 (500 docs) %(correctos)
Multinomial Naïve Bayes	18489	22.8 %(114)
Rocchio	18489	29.4 %(147)

Tabla 5.2: Resultados de la primera aproximación.

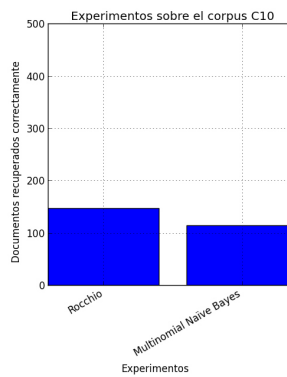


Figura 5.1: Gráfica de resultados de la primera aproximación.

5.3. Segunda aproximación

En la segunda aproximación (véase el capítulo 4) se utiliza como clasificador una máquina de soporte vectorial, la cual es un sistema de aprendizaje basado en el uso de un espacio de hipótesis de funciones lineales en un espacio de mayor dimensión inducido por un *Kernel*, en el cual las hipótesis son entrenadas por un algoritmo tomado de la teoría de optimización el cual utiliza elementos de la teoría de generalización.

Debido a las limitaciones computacionales de las máquinas de aprendizaje lineal, éstas no pueden ser utilizadas en la mayoría de las aplicaciones del mundo real. La representación por medio del *Kernel* ofrece una solución alternativa a este problema, proyectando la información a un espacio de características de mayor dimensión, el cual aumenta la capacidad computacional de las máquinas de aprendizaje lineal. Se mapea el espacio de entrada X a un nuevo espacio de características como se muestra en la Ecuación 5.5.

$$X = \{x_1, x_2, \dots, x_n\} \rightarrow \phi(x) = \{\phi(x)_1, \phi(x)_2, \dots, \phi(x)_n\} \quad (5.5)$$

Utilizando la función *Kernel* no es necesario calcular explícitamente el mapeo $\phi : X \rightarrow F$ para aprender en el espacio de características.

En el caso de esta investigación se utilizó como *Kernel* el mapeo polinomial que es un método muy popular para modelar funciones no lineales, como se muestra en la Ecuación 5.6,

$$K(x, x) = (\langle x, x \rangle + C)^d \quad (5.6)$$

donde $C \in R$.

La máquina de soporte vectorial utilizada en esta aproximación (y en aproximaciones subsecuentes) es la implementada en la plataforma de software para aprendizaje automático y minería de datos WEKA [14], ya que está disponible como software libre, es portable y contiene una extensa colección de técnicas para el preprocesamiento de datos y modelado.

Los parámetros usados en WEKA para la clasificación de un conjunto de documentos son los siguientes:

- **Número de características.** Parámetro utilizado para crear el vector de características, así como para la definición de cada uno de los atributos en el archivo creado por WEKA (extensión .arrf)
- ***BuildLogisticModels(Logistic Model)***. Parámetro utilizado para adaptar las estimaciones hechas por la máquina de soporte vectorial a un modelo logístico, el cual proporciona una mejor aproximación para características de tipo nominal.
- ***NumFolds***. Parámetro utilizado para determinar el número de particiones hechas en el training, usado para crear el modelo de clasificación, mientras mas particiones existan mayor será la precisión de las estimaciones hechas, aunque los recursos utilizados para crear el modelo se duplicaran.

Los resultados obtenidos utilizando una máquina de soporte vectorial con un kernel polinomial (*polyKernel* en WEKA) sobre el conjunto de características escogidas se muestran en la Tabla 5.3 y en la Figura 5.2. Entre las características que obtienen un mejor rendimiento, es decir que lograron recuperar mejor los rasgos distintivos de cada autor se encuentran las siguientes:

- Trigramas de las categorías gramaticales (45.4 %).
- Sufijos de palabras (32.4 %).
- Palabras cerradas (31.2 %).
- Combinación de vocales (26.2 %).
- Permutación de vocales (34.6 %).
- Letras del abecedario (27.8 %).

Sin embargo, la selección de las características contempló no solamente el porcentaje de valores correctos, sino también el cubrimiento de los documentos que corresponden a los autores. El proceso de selección se llevó a cabo de la siguiente manera: Primeramente se eligió la característica que obtuvo el mejor rendimiento global (en este caso, los trigramas de las categorías gramaticales), posteriormente, se fueron agregando aquellas características con un buen porcentaje de documentos clasificados correctamente, pero que

además recuperan documentos que las características anteriores no habían recuperado.

Con estos resultados en mente, se decide entonces realizar un nuevo experimento llamado combinación de características, el cual considera ahora la frecuencia de las seis características (figura 5.3) anteriormente comentadas. El número de documentos recuperados correctamente por el modelo de clasificación que conjuga la seis características es significativamente mejor que cualquiera de los resultados obtenidos de manera individual.

Característica	Parametros WEKA			C10 (500 docs) %(correctos)
	Número de características	Logistic Model	Num Folds	
Palabras con frecuencia igual a 1	1305	false	4	16.4 %(82)
Palabras con frecuencia igual a 2	976	false	4	18.6 %(93)
Palabras más frecuentes	10	false	5	23 %(115)
Colocaciones	543	false	5	15.4 %(77)
Prefijos de palabras	46	true	5	24.8 %(124)
Sufijos de palabras	131	true	5	32.4 %(162)
Trigramas de palabras	16359	false	4	14.2 %(71)
Palabras cerradas	195	true	3	31.2 %(156)
Número de palabras	1000	false	2	25.8 %(129)
Número de palabras por oración	500	false	2	16 %(80)
Número de oraciones	16	false	2	11 %(55)
Número de palabras capitalizadas	50	false	2	12.6 %(63)
Espacios utilizados en la escritura	50	false	2	7.6 %(38)
Trigramas de puntuación	1575	true	1	22.2 %(111)
Trigramas de las categorías gramaticales	13969	true	1	45.4 %(227)
Palabras al inicio del documento	50	false	2	18 %(90)
Palabras al final del documento	50	false	2	16.2 %(81)
Sinónimos	10544	false	3	15.2 %(78)
Hipónimos	5929	false	3	13 %(65)
Hiperónimos	7542	false	3	18.4 %(92)
Palabras longitud 10	2520	true	4	23.8 %(119)
Combinación de vocales	31	true	3	26.2 %(131)
Permutación de vocales	325	true	3	34.6 %(173)
Letras del abecedario	25	true	5	27.8 %(139)
Trigrama de caracter	35907	false	2	12.2 %(61)
Número de caracteres por oración	50	false	3	17.6 %(88)
Signos de puntuación	32	false	5	19.2 %(96)
Combinación de características	14676	true	2	56.8 %(284)

Tabla 5.3: Resultados de la segunda aproximación.

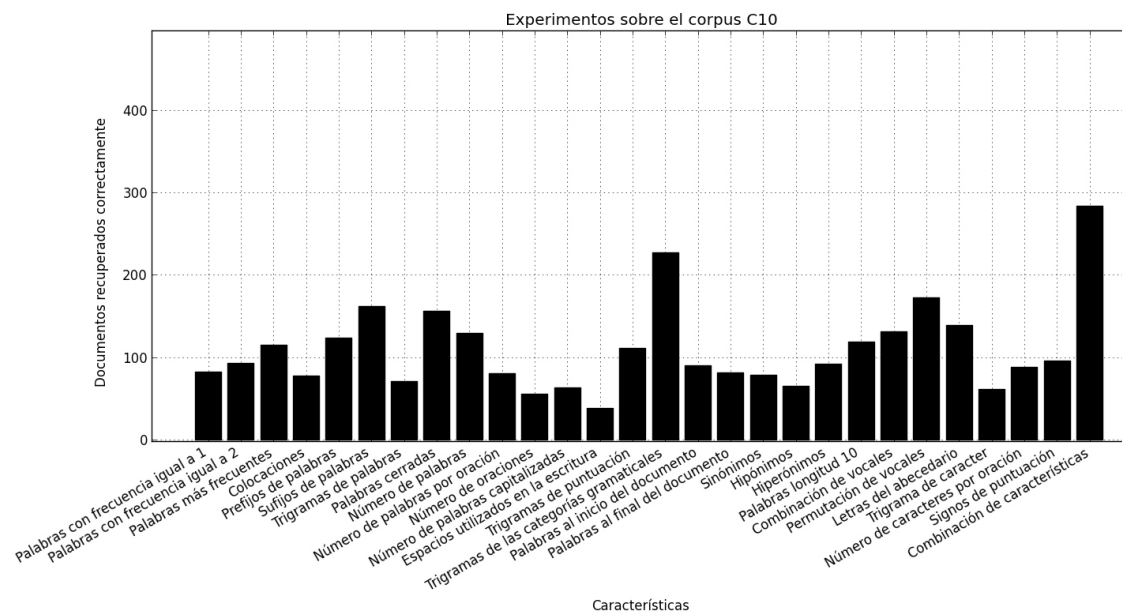


Figura 5.2: Gráfica de resultados de la segunda aproximación.

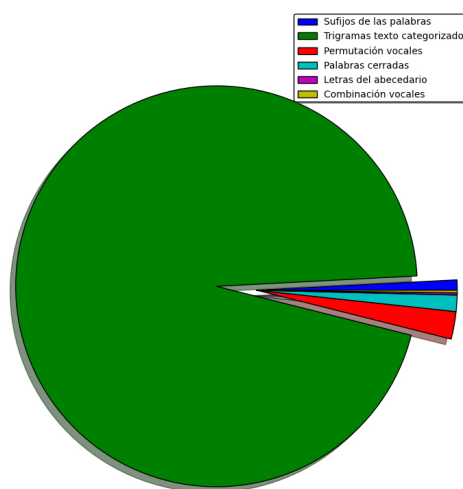


Figura 5.3: Proporción de características usadas en la combinación.

5.4. Tercera y cuarta aproximación

En la tercera y cuarta aproximación (véase el capítulo 4) se utiliza una máquina de soporte vectorial con un kernel de tipo polinomial (vease la sección 5.3).

Los resultados obtenidos para la tercera aproximación sobre el conjunto de características escogidas se muestran en la Tabla 5.4 y en la Figura 5.4, mientras que los resultados para la cuarta aproximación sobre el conjunto de características escogidas se muestran en la Tabla 5.5 y en la Figura 5.5.

5.5. Quinta aproximación

En la quinta aproximación (véase el capítulo 4) se utiliza la herramienta SUBDUE para obtener las palabras o patrones relevantes de un autor específico, los parámetros a ajustar para realizar las pruebas incluyen los siguientes:

- **Número de características.** Parámetro utilizado para crear los grafos asociados a cada autor, donde los vértices y las aristas son creados de la siguiente manera:

$$Vértice \rightarrow v \langle \text{característica} \# \rangle \langle \text{etiqueta} \rangle \quad (5.7)$$

$$Arista \rightarrow e \langle \text{vértice 1} \# \rangle \langle \text{vértice 2} \# \rangle \langle \text{etiqueta} \rangle$$

- **-nsubs.** Número de patrones (palabras en el caso de esta aproximación) a obtener por cada autor
- **-iterations.** Número de iteraciones hechas para encontrar los patrones relevantes de un autor
- **-eval** Método usado para encontrar los subgrafos o subestructuras comunes en los grafos introducidos como muestras positivas. Existen 2 valores usados para este parámetro:

- 1 (*Minimum Description Length Method, MDL*). Usado como método default para la extracción de subestructuras óptimas.
 - 2 (*Size Method*). Usado como método alternativo para encontrar la subestructura óptima, es más rápido que MDL pero menos eficiente en los cálculos de subgrafos candidatos
- **-beam**. Parámetro usado para determinar el grado de complejidad en las búsquedas de los sugrafos candidatos.

Una vez obtenidas las palabras relevantes se utiliza la herramienta WEKA con una máquina de soporte vectorial con un kernel de tipo polinomial, donde para todas las pruebas se ajustan los siguientes parámetros:

- BuildLogisticModels(Logistic Model)= true
- NumFolds = 5

Los resultados obtenidos sobre el conjunto de características escogidas se muestran en la Tabla 5.6 y en la Figura 5.6

5.6. Sexta aproximación

En la sexta aproximación (véase el Capítulo 4) se utiliza la herramienta SUBDUE de la misma forma que en la aproximación anterior, para obtener las palabras o patrones relevantes de un autor específico. De cada palabra obtenida se buscan todos los sentidos de los sinónimos, hipónimos e hiperónimos que se puedan encontrar. Teniendo dicho conjunto de palabras (tanto las palabras originales como las palabras expandidas) se utiliza la herramienta WEKA con una máquina de soporte vectorial con un kernel de tipo polinomial, donde para todas las pruebas se ajustan los mismos parámetros que la aproximación anterior.

Los resultados obtenidos sobre el conjunto de características escogidas se muestran en la Tabla 5.7 y en la Figura 5.7

Característica	Abreviatura		
Palabras con frecuencia igual a 1	A		
Palabras con frecuencia igual a 2	B		
Palabras más frecuentes	C		
Prefijos de palabras	D		
Sufijos de palabras	E		
Palabras cerradas	F		
Número de palabras	G		
Trigramas de puntuación	H		
Trigramas de las categorías gramaticales	I		
Palabras longitud 10	J		
Combinación de vocales	K		
Permutación de vocales	L		
Letras del abecedario	M		
Experimento	Parámetros WEKA		
	Número de características	Logistic Model	Num Folds
$A \cup B$	2281	false	4
$\bar{C}$	18477	false	2
$J - E$	2433	true	2
$K \Delta L$	247	true	4
$H \cup I$	15544	false	2
$E \cap J$	42	false	5
$M \cap F \cup G$	1220	false	2
$C \Delta D$	17	true	5
			C10 (500 docs)
			%(correctos)
			12.6 % (63)
			10.4 % (52)
			9.4 % (47)
			26.4 % (132)
			34.4 % (172)
			19.4 % (97)
			25.6 % (128)
			21.2 % (106)

Tabla 5.4: Resultados de la tercera aproximación.

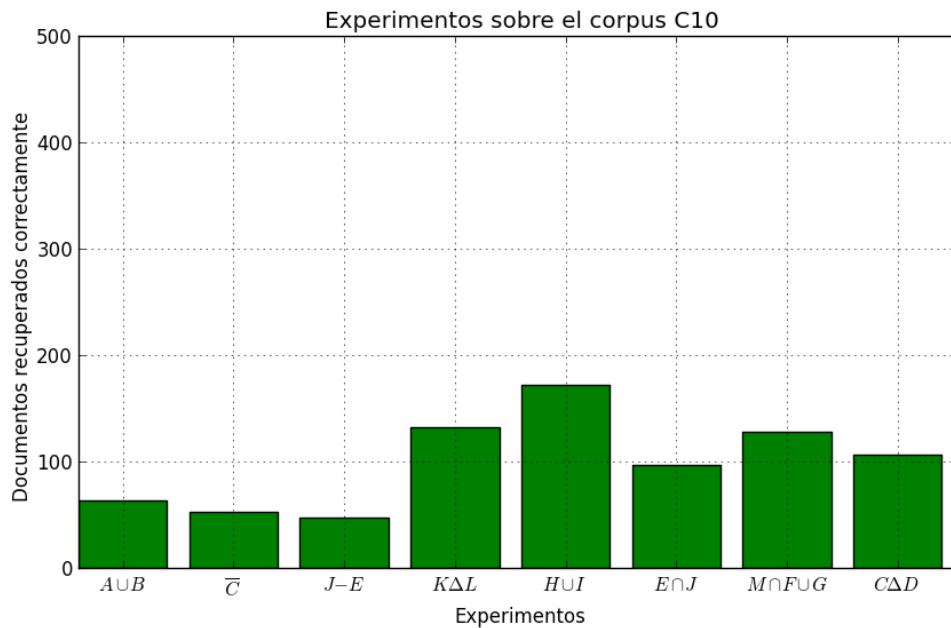


Figura 5.4: Gráfica de resultados de la tercera aproximación.

Característica	Parámetros WEKA			C10 (500 docs) %(correctos)
	Número de características	Logistic Model	Num Folds	
Palabras con frecuencia igual a 1	1305	false	4	6.2 %(31)
Palabras con frecuencia igual a 2	976	false	4	9 %(45)
Palabras más frecuentes	10	true	5	7.4 %(37)
Colocaciones	543	false	5	4.2 %(21)
Prefijos de palabras	46	true	5	10.6 %(53)
Sufijos de palabras	131	true	5	13 %(65)
Trigramas de palabras	16359	true	4	11.6 %(58)
Palabras cerradas	195	true	3	15.4 %(77)
Trigramas de puntuación	1575	false	5	6 %(30)
Trigramas de las categorías gramaticales	13969	true	5	17 %(85)
Palabras al inicio del documento	50	false	2	3.2 %(16)
Palabras longitud 10	2520	true	4	5.2 %(26)
Combinación de vocales	31	false	5	12.6 %(63)
Permutación de vocales	325	false	5	13.2 %(66)
Letras del abecedario	25	true	5	3 %(15)

Tabla 5.5: Resultados de la cuarta aproximación.

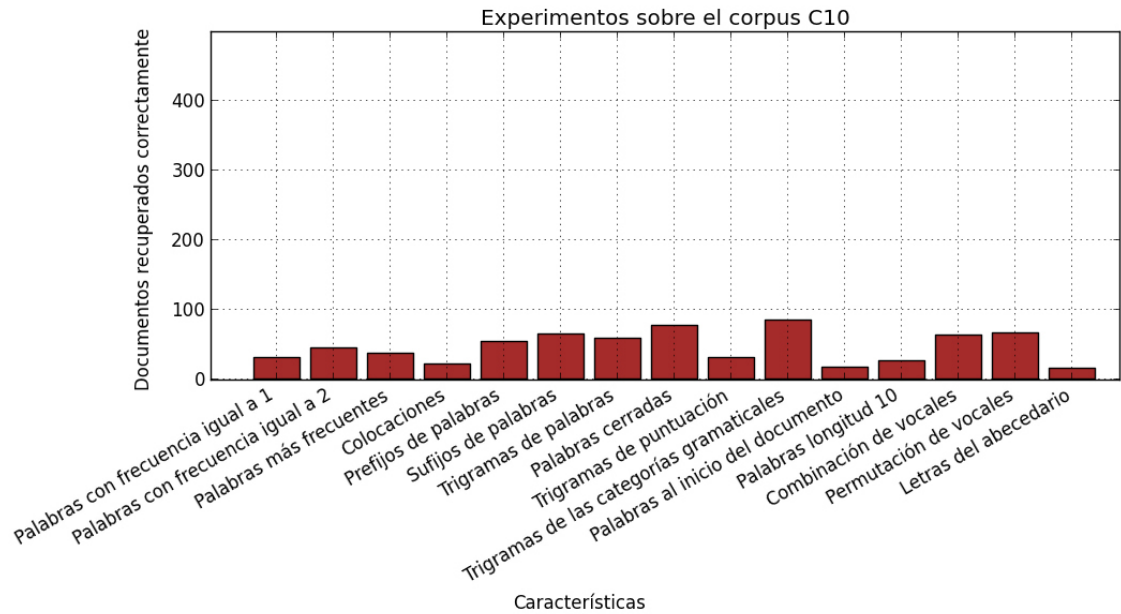


Figura 5.5: Gráfica de resultados de la cuarta aproximación.

Grafo	Parámetros SUBDUE					C10 (500 docs) %(correctos)
	Número de características	-nsubs	-iterations	-eval	-beam	
Grafo representación 1	296	30	5	1	10	70.6%(353)
Grafo representación 2	265	30	5	1	10	75.8%(379)
Combinación de las 2 representaciones	561	30	4	1	15	77.6%(388)

Tabla 5.6: Resultados de la quinta aproximación.

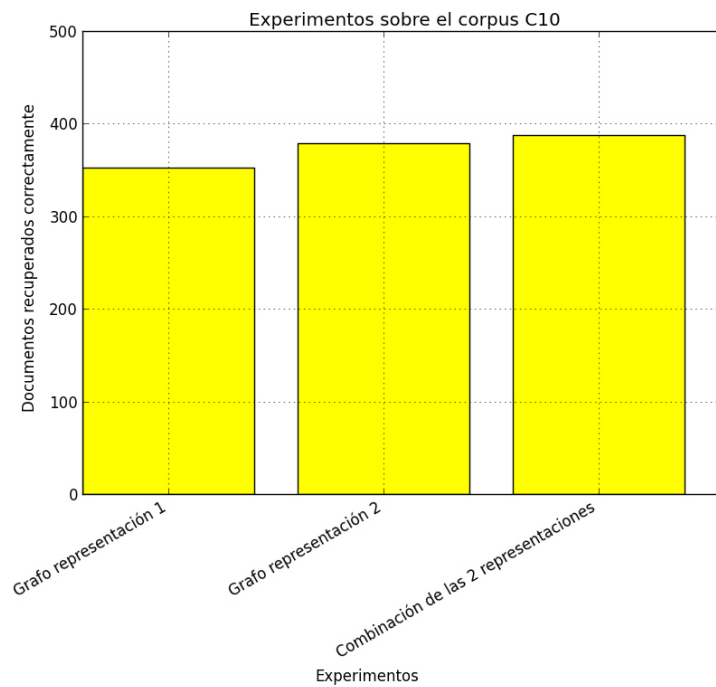


Figura 5.6: Gráfica de resultados de la quinta aproximación.

Grafo	Parámetros SUBDUE					C10 (500 docs) %(correctos)
	Número de características	-nsubs	-iterations	-eval	-beam	
Grafo representación 1 expansión sinónimos	1972	30	2	1	10	68 %(340)
Grafo representación 2 expansión sinónimos	2161	30	2	1	10	70 %(350)
Grafo representación 1 expansión hiperónimos	1963	30	2	2	10	79.2 %(396)
Grafo representación 2 expansión hiperónimos	2027	30	2	2	10	80.4 %(402)
Grafo representación 1 expansión hipónimos	6116	30	3	2	10	69.2 %(346)
Grafo representación 2 expansión hipónimos	6447	30	4	2	15	68.8 %(343)
Combinación sinónimo de las 2 representaciones	5320	30	2	1	10	71.6 %(358)
Combinación Hiperónimo de las 2 representaciones	1812	30	2	1	10	78.8 %(392)
Combinación hipónimo de las 2 representaciones	5320	30	2	1	10	73.8 %(369)

Tabla 5.7: Resultados de la sexta aproximación.

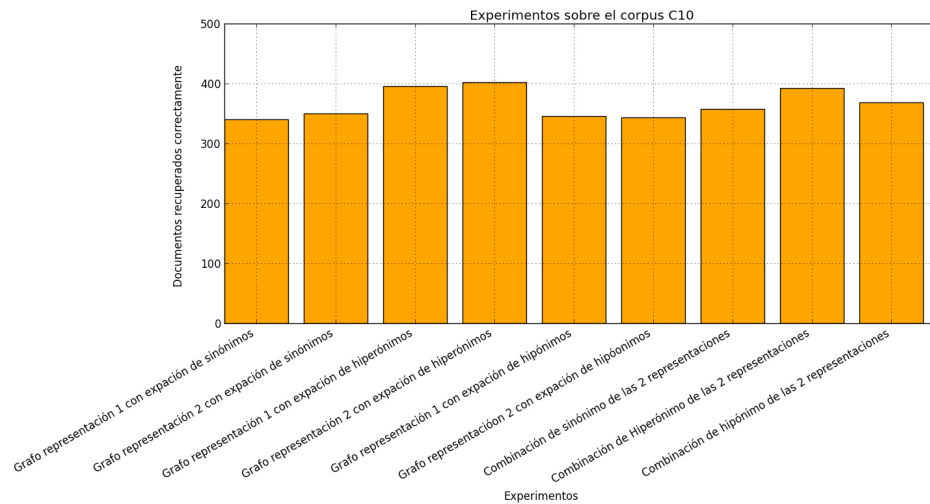


Figura 5.7: Gráfica de resultados de la sexta aproximación.

5.7. Participación PAN CLEF 2012

En el marco de la competencia **6th Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN'12)**¹, se probaron las 2 mejores aproximaciones (segunda y quinta aproximación) sobre la tarea **Traditional Authorship Attribution** de PAN'12. En dicha tarea dado un conjunto de textos de procedencia anónima, el objetivo es asignar los textos a su autor real tomado de un conjunto de autores candidatos (clase cerrada, atribución de autoría).

El conjunto de datos proporcionados en la tarea consiste en 3 colecciones de textos, formadas por distintos autores (Tabla 5.8). La primera colección (A) consta de 3 autores diferentes, la segunda (C) consta de 8 autores diferentes y la tercera (I) consta de 14 autores diferentes.

Tarea	Problema	Tipo	Autores	Documentos training	Documentos test
Traditional	A	clase cerrada	3	6	6
Traditional	C	clase cerrada	8	16	8
Traditional	I	clase cerrada	14	28	14

Tabla 5.8: Conjunto de datos usados en PAN'12.

Los resultados obtenidos en la competencia se muestran en la Tabla 5.9

Tarea	A(6 Docs Test) % correctos	C(8 Docs Test) % correctos	I(14 Doc Test) % correctos
Segunda aproximación	66.666%(4)	25%(2)	71.42%(10)
Quinta aproximación	83.33%(5)	62.5%(5)	57.142%(8)

Tabla 5.9: Resultados obtenidos en PAN'12.

5.8. Herramientas utilizadas y desarrolladas

En la Tabla 5.10 se presentan las distintas herramientas y paquetes utilizados en cada una de las aproximaciones realizadas para resolver el problema de atribución de autoría.

¹<http://pan.webis.de/>

Aproximación 1			
Herramienta	Lenguaje	Paquete	Implementación
Word tokenize	NLTK	Tokenizer	Usada para dividir un texto en tokens(palabras)
Aproximación 2			
Herramienta	Lenguaje	Paquete	Implementación
Punkt	NLTK	Tokenizers	Implementada para partir un documento en ingles en sentencias
Permutations	NLTK	Itertools	Usada para realizar la permutación de un conjunto de elementos
Combinations	NLTK	Itertools	Usada para realizar la combinación de un conjunto de elementos
Sufijos	Python	Re	Implementada para hallar la parte final de una palabra
Prefijos	Python	Re	Implementada para hallar la parte inicial de una palabra
Pos_tag	Stanford	Libre	Usado para obtener la etiqueta gramatical de las palabras
Colocaciones	NLTK	collocations	Implementado para obtener los bigramas más frecuentes
Sinónimos	Awk	Wornet-crawler	Se obtienen todos los sinónimos asociados a una palabra
Hipónimos	Awk	Wornet-crawler	Se obtienen todos los hipónimos asociados a una palabra
Hiperónimos	Awk	Wornet-crawler	Se obtienen todos los hiperónimos asociados a una palabra
Aproximación 3			
Herramienta	Lenguaje	Paquete	Implementación
Operaciones de conjuntos	Python	Libre	Herramienta utilizada para delimitar las características por medio de operaciones matemáticas de conjuntos
FeatureExtractor	Python	Libre	Herramienta creada para el manejo y creación de los vectores de características
AuthorList	Python	Libre	Herramienta creada para el manejo de los autores y sus documentos asociados
Aproximación 4			
Herramienta	Lenguaje	Paquete	Implementacion
Math	Python	Math tools	Herramienta usada para implementar diversas operaciones matemáticas complejas
Decimal	Python	Math tools	Herramienta usada para implementar operaciones de punto flotante
Aproximación 5			
Herramienta	Lenguaje	Paquete	Implementación
TreeTagger	java	Libre	Usado para obtener la etiqueta gramatical de las palabras
Porter stemmer	java	Libre	Usado para reducir una palabra a su raíz o (en inglés) a un stem o lema
Grafo texto	python	Libre	Herramienta creada para implementar un grafo basado en secuencias de palabras categorizadas gramaticalmente de un documento
Grafo oracion	python	Libre	Herramienta creada para implementar un grafo basado en secuencias de palabras categorizadas gramaticalmente de una oración
Aproximación 6			
Herramienta	Lenguaje	Paquete	Implementación
Sinónimos	NLTK	Wordnet	Se obtienen todos los sinónimos asociados a una palabra
Hipónimos	NLTK	Wordnet	Se obtienen todos los hipónimos asociados a una palabra
Hiperónimos	NLTK	Wordnet	Se obtienen todos los hiperónimos asociados a una palabra

Tabla 5.10: Herramientas usadas en la aproximación 1,2,3,4,5 y 6.

5.9. Conclusión del capítulo

En este capítulo se presentan los resultados obtenidos con las diferentes aproximaciones para resolver la autoría de un documento desconocido. La importancia de este capítulo radica en la explicación del uso de los algoritmos o herramientas utilizadas para determinar el autor de un documento, así como los parámetros usados para ajustar los modelos de clasificación, con el objetivo de obtener un mayor número de documentos recuperados.

CONCLUSIONES Y TRABAJO A FUTURO

A conclusion is the place where
you get tired of thinking.

Arthur Bloch

En este capítulo se presentan las conclusiones finales acerca de los distintos métodos para determinar la autoría de un documento, así como las contribuciones hechas por el trabajo en el área de atribución de autoría y el trabajo a futuro propuesto para una posible extensión de esta tesis.

6.1. Conclusiones finales

En este trabajo de tesis se cumplieron todos los objetivos planteados en la investigación para detectar el autor real de un documento de origen desconocido. En particular se realizó un estudio de los métodos existentes para extraer características léxicas, sintácticas y semánticas, así como el desarrollo de distintos métodos y representaciones para determinar la autoría de un documento. En estos métodos se destacan la combinación de características por medio de una representación basada en una bolsa de palabras ponderada y la creación de una representación basada en la secuencia sintáctica de las oraciones por medio de grafos.

Haciendo un análisis del proceso de atribución de autoría desarrollado en este trabajo de investigación, se puede observar que en los primeros enfoques para la extracción de elementos representativos del autor, se utilizaba todo el conjunto de datos. Esto provocaba que la dimensionalidad de las características fuera demasiado grande en comparación con otros métodos.

Tomando en cuenta la importancia de hacer una selección de características óptima, en aproximaciones subsecuentes se extrajeron todas las características léxico-sintácticas individuales a nivel de oración y de carácter. Además como se planteó la posibilidad de usar una combinación de dichas características, intentando en este proceso obtener características más representativas del autor.

A pesar de que el uso de una combinación de características proporcionó una buena representación del estilo de escritura del autor ésta no era una representación estructurada que contemplara la estructura sintáctica de los textos. Por ello en aproximaciones subsecuentes se planteó la necesidad de una representación basada en grafos la cual ayudara a encontrar patrones relevantes que no eran posibles de determinar por el uso o extracción de características individuales o combinadas. Los patrones relevantes obtenidos en la representación basada en un grafo permitieron a su vez, extraer características semánticas por medio de una expansión de sinónimos, hipónimos e hiperónimos. Dichas características permitieron encontrar palabras que no estaban presentes en el conjunto de datos original, pero que están fuertemente relacionadas con los patrones obtenidos.

Tomando en cuenta el análisis anterior, en la Figura 6.1 se presentan los mejores resultados de cada aproximación, donde se puede hacer notar que la aproximación que hace uso de una expansión de sinónimos sobre las palabras obtenidas por la herramienta Subdue obtienen la mejor precisión (80.4 % de documentos correctamente clasificados) con respecto a aproximaciones basadas en la extracción de características individuales. Este resultado indica que el uso de una representación basada en la estructura sintáctica de los textos en conjunto con la utilización de características semánticas puede mejorar considerablemente el número de documentos asignados correctamente a un autor. A su vez se puede apreciar que el uso de un método para la combinación de características puede proporcionar un conjunto de elementos léxico-sintácticos capaces de obtener resultados bastante competitivos con respecto a otros encontrados en la extensa biografía relacionada con la atribución de autoría de textos.

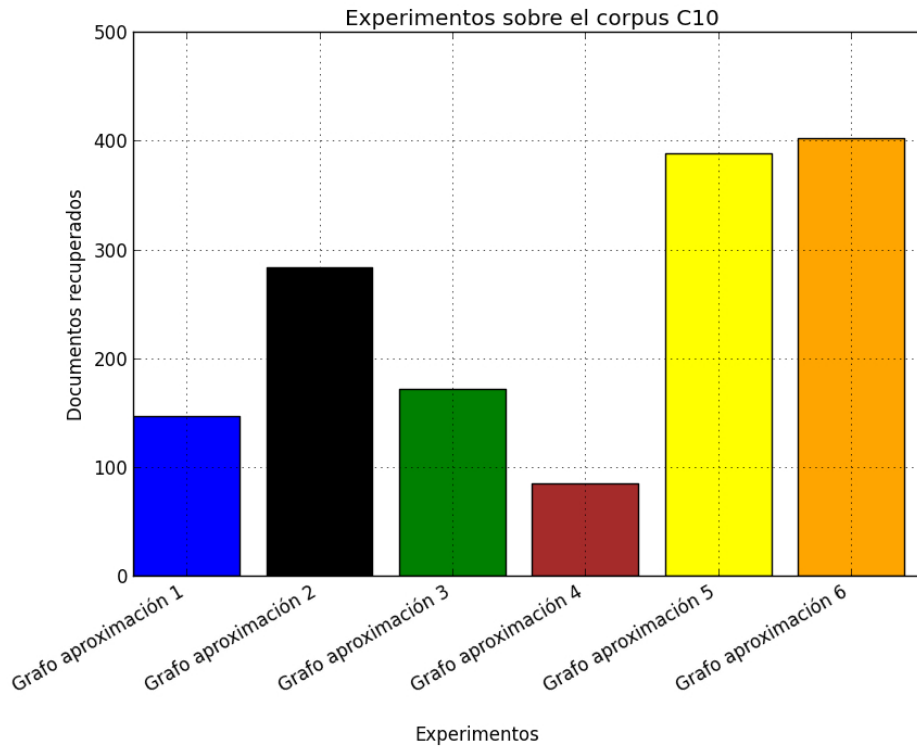


Figura 6.1: Gráfica de los mejores resultados de cada aproximación.

6.1.1. Conclusión PAN CLEF 2012

Los resultados obtenidos en la competencia CLEF PAN 2012 demuestran que el uso de la combinación de características y de la representación basada en grafos obtuvo un comportamiento estable en los tópicos usados en el corpus del PAN 2012(noticias corporativas, chats y libros de tipo literario) para la tarea de atribución de autoría. Observamos que la representación basada en grafos logra descubrir patrones más relevantes que la combinación reduciendo la dimensionalidad de los datos de manera clara. Por otra parte se puede apreciar que las características y representación antes mencionadas tienen un comportamiento distinto dependiendo de la longitud de los textos a evaluar.

6.2. Contribuciones del trabajo

A continuación se presentan las contribuciones hechas en el proceso de investigación:

1. Creación de una metodología para la tarea de atribución de autoría.
2. Creación de un *Baseline* para la tarea de atribución de autoría.
3. Extracción de características léxicas no definidas en trabajos previos del área.
4. Extracción de características sintácticas no definidas en trabajos previos del área.
5. Adaptación de algoritmos de clasificación a la tarea de atribución de autoría.
6. Creación de una representación basada en la combinación de características léxico-sintácticas.
7. Desarrollo de dos representaciones basadas en grafos para obtener patrones relevantes de un autor.
8. Obtención de características semánticas relevantes por medio de la expansión de sinónimos, hiperónimos e hipónimos.
9. Obtención de un conjunto de características estables sobre corpora con documentos acerca de noticias corporativas, chats y documentos literarios de tamaño considerable.
10. Escritura de artículos [49, 50, 51, 52] relacionados al problema de atribución de autoría sobre distintos conjuntos de datos usando las mejores técnicas implementadas en este trabajo de investigación.

6.3. Trabajo a futuro

Se considera que una posible extensión de este trabajo de investigación incluya los siguientes aspectos:

1. La extracción de nuevas características semánticas que no hayan sido consideradas en este trabajo.
2. Prueba de la metodología creada en este trabajo sobre distintos corpora, los cuales no hayan sido preparados para la tarea de atribución de autoría.
3. Uso de la representación basada en grafos (aproximación 5, proceso de clasificación versión 2) por oración para determinar la autoría de un documento dado, por medio del uso de una herramienta de isomorfismo de grafos.
4. Creación de distintas representaciones basadas en grafos, las cuales contemplen no sólo la secuencia sintáctica de las oraciones.
5. Uso de otros algoritmos de clasificación no considerados en esta investigación con anterioridad.
6. Creación de una metodología para la tarea de verificación de autoría, donde se use una extracción de elementos léxicos, sintácticos y semánticos.

Apéndice A. Clasificación con WEKA

A continuación se muestran los pasos para realizar una clasificación supervisada con la herramienta WEKA utilizando modo gráfico:

1. Se crea un archivo Training.arff como el que se muestra a continuación:

```
@relation < nombre de relacion >
@attribute < atributo1 > < tipo de dato >
:
@attribute < atributon > < tipo de dato >
@attribute autor {autor1, autor2, ..., autorn}
@data
v1, v2, v3, v4, ..., vn, autor1 (documento1)
:
v1, v2, v3, v4, ..., vn, autorm (documentom)
```

donde *< tipo de dato >* puede ser numeric, string, date o nominal. El atributo *autor* corresponde al atributo clasificador y v_n en el vector de características corresponde al valor de *< atributo_n >*.

2. Se crea un archivo Test.arff que debe tener la misma cantidad de atributos (*@attribute*) y en el mismo orden que en el archivo de Training:

```

@relation < nombre de relacion >
@attribute < atributo1 > < tipo de dato >
:
@attribute < atributon > < tipo de dato >
@attribute autor {autor1, autor2, ..., autorn}
@data
v1, v2, v3, v4, ..., vn, ? (documento1)
:
v1, v2, v3, v4, ..., vn, ? (documentom)

```

donde ? indica que el documento es de origen desconocido.

3. Se abre WEKA en modo gráfico (en *Windows*) de la siguiente manera:
 - a) Se abre una consola de comandos.
 - b) Se sitúa en la consola en *ArchivosdeProgramas/weka*.
 - c) Se escribe *java-Xmx < cantidaddememoriaenMb > -jarweka.jar*.



Figura A1: Menú inicial de WEKA.

4. Se elige la opción de *Explorer* como se muestra en la Figura A2.

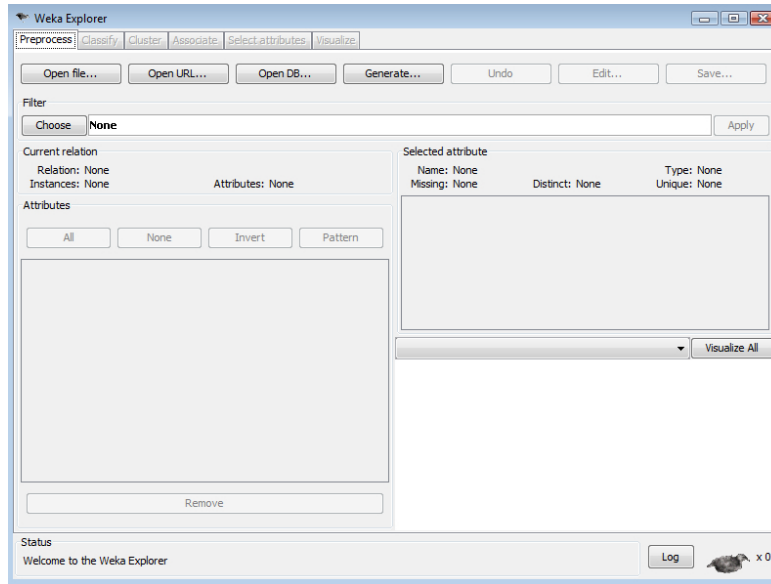


Figura A2: Menú de exploración de WEKA.

5. Se elige la opción *open file* y se busca el archivo *Training.arff* como se muestra en la Figura A3.

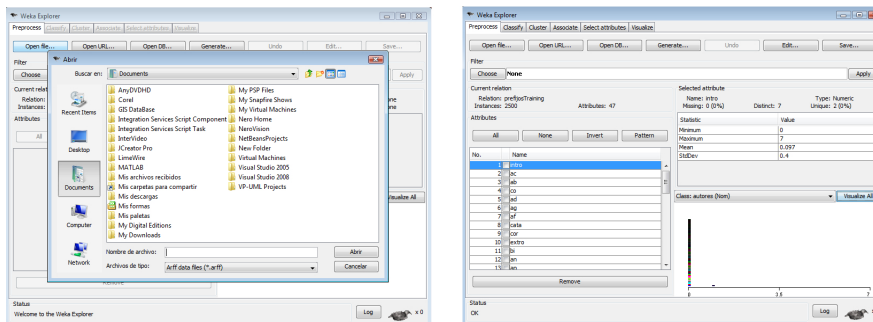


Figura A3: Menú de exploración con la opción de un archivo training.

6. Se selecciona la pestaña de *Classify* como se muestra en la Figura A4.

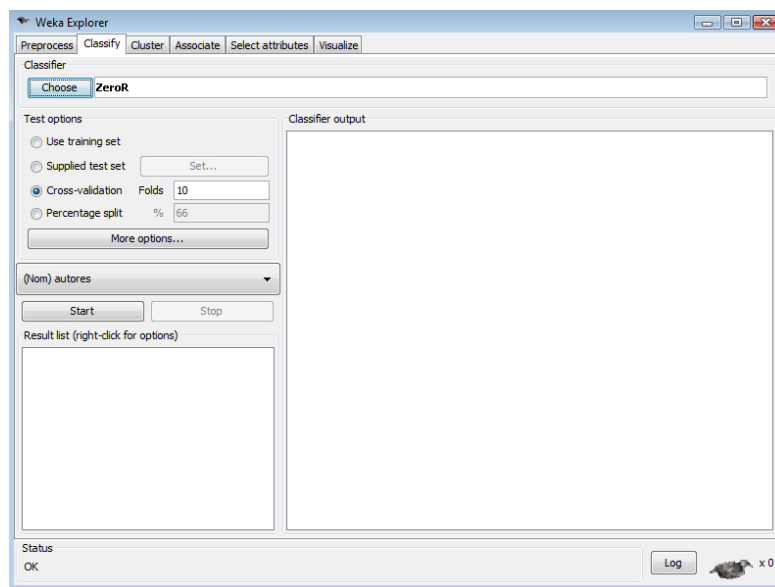


Figura A4: Menú de clasificación de WEKA.

7. Se escoge la opción de *use Training set* y se escoge el algoritmo de clasificación en la opción *Choose* como se muestra en la Figura A5.

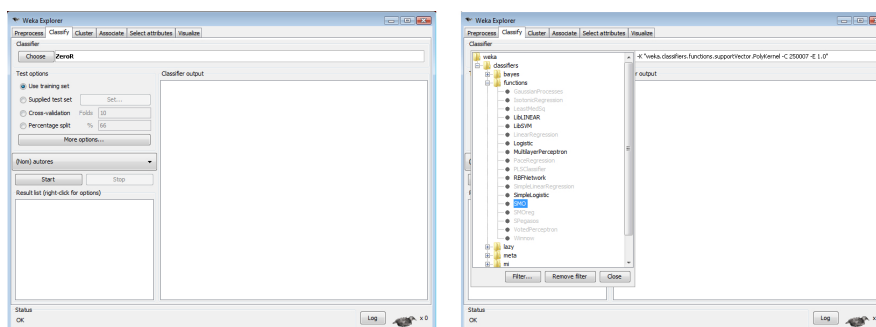


Figura A5: Menú de tipo de entrenamiento y algoritmo de clasificación.

8. Se cambian los parámetros del algoritmo de clasificación como se muestra en la Figura A6.

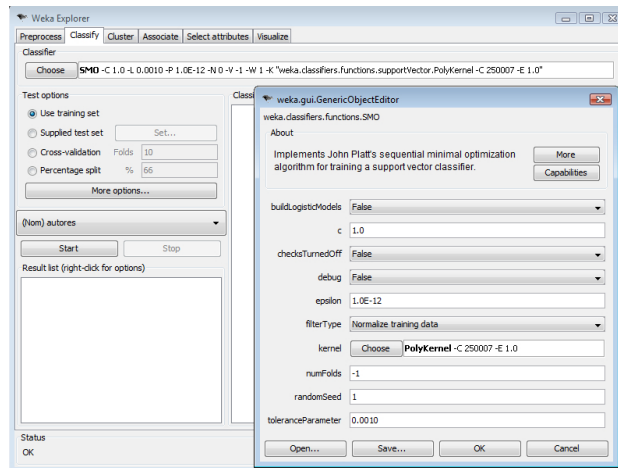


Figura A6: Menú de parámetros del algoritmo de clasificación.

9. Se escoge la opción *start* para iniciar con el *Training* y creación del modelo de clasificación como se muestra en la Figura A7.

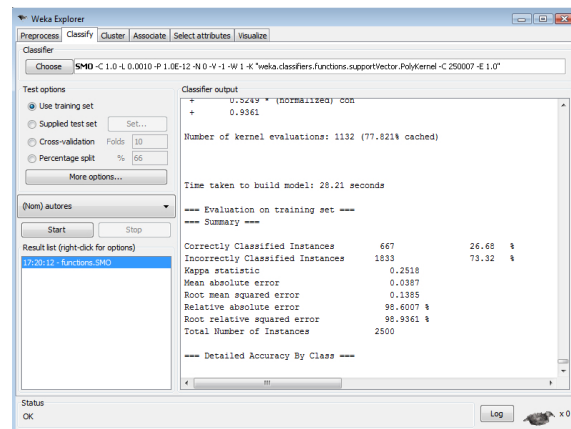


Figura A7: Menú de inicio de clasificación con training.

10. Se selecciona la opción de *supplied test set* y después la opción *set*, seguido de esto se selecciona *open file* y se busca el archivo *Test.arff* como se muestra en la Figura A8.

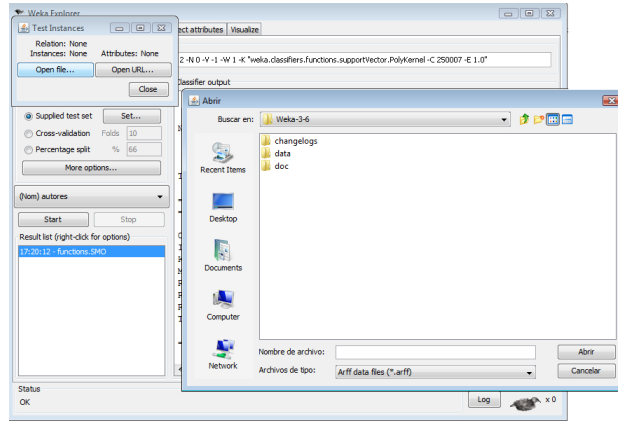


Figura A8: Menú de selección del archivo test.

11. Se selecciona del menú emergente del modelo creado la opción *Re-evaluate model on current test set* como se muestra en la Figura A9.

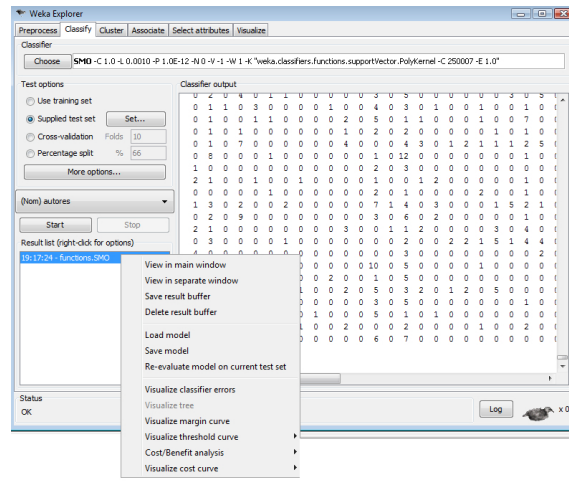


Figura A9: Menú de prueba del Test.

12. Se selecciona del menú emergente de la prueba del test creado la opción *Visualize classifier errors* como se muestra en la Figura A10.

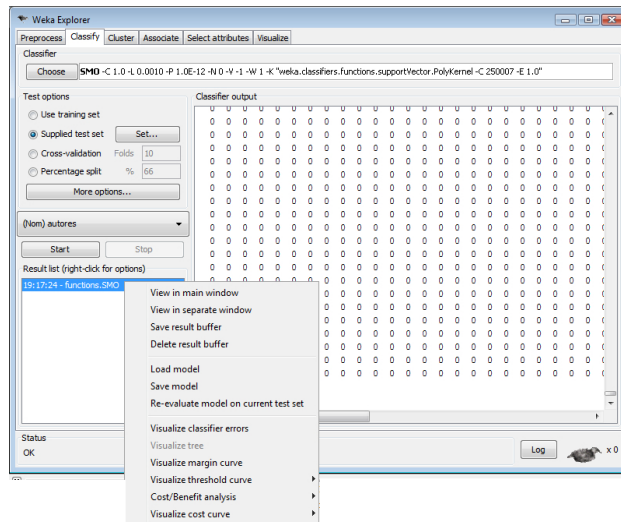


Figura A10: Menú de visualización de errores del test.

13. Se guardan los resultados con la opción *save* en un archivo *Results.arff* como se muestra en la Figura A11.

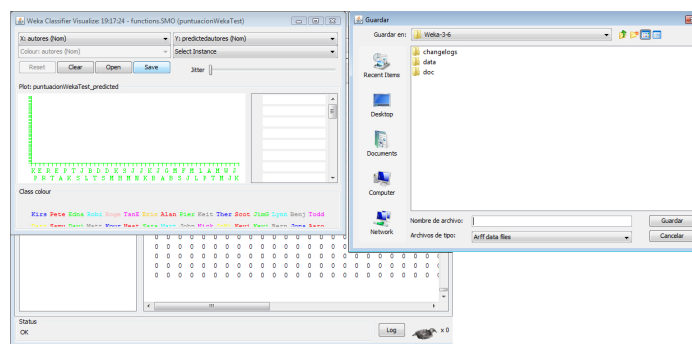


Figura A11: Menú de resultados del test.

14. El archivo *Results.arff* tendrá una estructura similar al archivo del test pero cambiará en los vectores de características como se muestra a continuación:

$$\begin{array}{c} v_1, v_2, v_3, v_4, \dots, v_n, ?, autor \ (documento_1) \\ \vdots \\ v_1, v_2, v_3, v_4, \dots, v_n, ?, autor \ (documento_m) \end{array}$$

donde el *autor* es el resultado del método de clasificación supervisado para cada uno de los documentos del test.

Apéndice B. Búsqueda de patrones con SUBDUE

A continuación se muestran los pasos para obtener patrones relevantes con la herramienta SUBDUE utilizando modo consola:

1. Se descarga SUBDUE de la página: <http://ailab.wsu.edu/subdue/>
2. Se descomprime y se abre la carpeta(en modo consola de unix-linux) de **subdue-<release>.zip**, la cual contiene las siguientes subcarpetas:
 - **/bin/**– Directorio con los archivos ejecutables de la aplicación.
 - **/docs/**– Directorio que contiene el manual de usuario proporcionado por la herramienta.
 - **/graphs/**– Directorio que contiene archivos ejemplos de la herramienta.
 - **/src/**– Directorio que contiene el código fuente de la aplicación.
3. Se escribe en la consola el comando **make**, el cual compila los programas de la aplicación.
4. Se escribe en la consola el comando **make install**, el cual copia los ejecutables a la carpeta **subdue-<release>/bin**.
5. Se escribe en la consola el comando **make clean**, el cual limpia el directorio **src** (removiendo los archivos contenidos en el).
6. Se crea un archivo prueba.txt, el cual contiene uno o más grafos con el siguiente formato:

$$\begin{aligned}
&< XP/XN > \\
&Vértice \rightarrow v \langle característica_1 \# \rangle \langle etiqueta_1 \rangle \\
&\vdots \\
&Vértice \rightarrow v \langle característica_n \# \rangle \langle etiqueta_n \rangle \\
&Arista \rightarrow \langle e/d \rangle \langle vértice 1 \# \rangle \langle vértice 2 \# \rangle \langle etiqueta \rangle \\
&\vdots \\
&Arista \rightarrow \langle e/d \rangle \langle vértice n \# \rangle \langle vértice m \# \rangle \langle etiqueta \rangle
\end{aligned}$$

donde XP significa una muestra positiva, mientras que XN significa una muestra negativa, $\langle característica_n \# \rangle$ es un identificador de vértice único y $\langle e/d \rangle$ indica que la unión entre los vértices es de forma **e** dirigida y **d** no dirigida

7. Para la ejecución de subdue nos situamos en la carpeta **subdue- $\langle release \rangle$ /bin** y ejecutamos **./subdue $\langle parametros \rangle$ prueba.txt**, donde los parámetros usados en este trabajo son los siguientes (para mayor referencia revisar el manual de SUBDUE¹):

- -beam
- -eval
- -iterations
- -nsbs

¹<http://ailab.wsu.edu/subdue/>

Bibliografía

- [1] Stamatatos, E.: A survey of modern authorship attribution methods. *J. Am. Soc. Inf. Sci. Technol.* **60**(3) (March 2009) 538–556
- [2] Juola, P.: Authorship attribution. *Found. Trends Inf. Retr.* **1**(3) (December 2006) 233–334
- [3] Koppel, M., Schler, J.: Authorship verification as a one-class classification problem. In: *Proceedings of the twenty-first international conference on Machine learning. ICML '04*, New York, NY, USA, ACM (2004)
- [4] Joachims, T., Finley, T., Yu, C.N.J.: Cutting-plane training of structural svms. *Mach. Learn.* **77**(1) (October 2009) 27–59
- [5] Argamon, S., Juola, P.: Overview of the international authorship identification competition at pan-2011. In Petras, V., Forner, P., Clough, P.D., eds.: *CLEF (Notebook Papers/Labs/Workshop)*
- [6] Juola, P.: An overview of the traditional authorship attribution subtask. In: *CLEF (Online Working Notes/Labs/Workshop)*
- [7] Mosteller, F., Wallace, D.L.: *Inference and Disputed Authorship: The Federalist*. Springer-Verlag, New York (1964)
- [8] Mendenhall, T.C.: *The characteristic curves of composition*. *Science ns-9*(214S) (1887) 237–246
- [9] Yule, G.U.: *On sentence-length as a statistical characteristic of style in prose: with application to two cases of disputed authorship*. *Biometrika* **30**(3-4) (1939) 363–390
- [10] Yule, G.U.: *The Statistical Study of Literary Vocabulary*. Cambridge University Press (1944)

-
- [11] Zipf, G.K.: *Selective Studies and the Principle of Relative Frequency in Language*. Harvard University Press (1932)
- [12] HOLMES, D.I.: *The evolution of stylometry in humanities scholarship*. *Literary and Linguistic Computing* **13**(3) (1998) 111–117
- [13] Manning, C.D., Raghavan, P., Schütze, H.: *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA (2008)
- [14] Witten, I.H., Frank, E.: *Data Mining: Practical Machine Learning Tools and Techniques*. Second edn. Morgan Kaufmann Series in Data Management Sys. Morgan Kaufmann (June 2005)
- [15] Manning, C.D., Schütze, H.: *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, USA (1999)
- [16] Madigan, D., Genkin, A., Lewis, D.D., Lewis, E.G.D.D., Argamon, S., Fradkin, D., Ye, L., Consulting, D.D.L.: *Author identification on the large scale*. In: *Proc. of the Meeting of the Classification Society of North America*
- [17] Abbasi, A., Chen, H.: *Applying authorship analysis to extremist-group web forum messages*. *IEEE Intelligent Systems* **20**(5) (September 2005) 67–75
- [18] Chaski, C.E.: *Empirical evaluations of language-based author identification techniques*. *Forensic Linguistics* **8**(1) (2001) 1–65
- [19] Grant, T.D.: *Quantifying evidence for forensic authorship analysis*. *International Journal of Speech Language and the Law* **14**(1) (2007) 1–25
- [20] Argamon, S.: *Interpreting burrows’s delta: Geometric and probabilistic foundations*. *Literary and linguistic computing* **23**(2) (2008) 131–147
- [21] Hoover, D.L.: *Testing burrows’s delta*. *LLC* **19**(4) (2004) 453–475
- [22] Meyer, S., Eissen, Z., Stein, B., Kulig, M.: *Plagiarism detection without reference collections*. In: *GfKl.* (2006) 359–366
- [23] Koppel, M., Argamon, S., Shimoni, A.: *Automatically categorizing written texts by author gender* (2003)

- [24] Collins, J., Kaufer, D., Vlachos, P., Butler, B., Ishizaki, S.: *Detecting collaborations in text - comparing the authors' rhetorical language choices in the federalist papers*. *Computers and the Humanities* **38**(1) (2004) 15–36
- [25] Zhao, Y., Zobel, J.: *Effective and scalable authorship attribution using function words*. In: *Proceedings of the Second Asia conference on Asia Information Retrieval Technology*. AIRS'05, Berlin, Heidelberg, Springer-Verlag (2005) 174–189
- [26] Coyotl-Morales, R.M., Villaseñor Pineda, L., Montes-y Gómez, M., Rosso, P.: *Authorship attribution using word sequences*. In: *Proceedings of the 11th Iberoamerican conference on Progress in Pattern Recognition, Image Analysis and Applications*. CIARP'06, Berlin, Heidelberg, Springer-Verlag (2006) 844–853
- [27] Negri, M., Kouylekov, M.: *Fbk nk: A wordnet-based system for multi-way classification of semantic relations*. In: *Proceedings of the 5th International Workshop on Semantic Evaluation*. SemEval 10, Stroudsburg, PA, USA, Association for Computational Linguistics (2010) 202–205
- [28] Rink, B., Harabagiu, S.: *Utd: Classifying semantic relations by combining lexical and semantic resources*. In: *Proceedings of the 5th International Workshop on Semantic Evaluation*. SemEval '10, Stroudsburg, PA, USA, Association for Computational Linguistics (2010) 256–259
- [29] Tratz, S., Hovy, E.: *Isi: Automatic classification of relations between nominals using a maximum entropy classifier*. In: *Proceedings of the 5th International Workshop on Semantic Evaluation*. SemEval '10, Stroudsburg, PA, USA, Association for Computational Linguistics (2010) 222–225
- [30] Nigam, K.: *Using maximum entropy for text classification*. In: *In IJCAI-99 Workshop on Machine Learning for Information Filtering*. (1999) 61–67
- [31] Solorio, T., Pillay, S., y Gómez, M.M.: *Authorship identification with modality specific meta features - notebook for pan at clef 2011*. In: *CLEF (Notebook Papers/Labs/Workshop)*

-
- [32] Kern, R., Seifert, C., Zechner, M., Granitzer, M.: *Vote/veto meta-classifier for authorship identification - notebook for pan at clef 2011*. In: *CLEF (Notebook Papers/Labs/Workshop)*
- [33] Cook, D.J., Holder, L.B.: *Graph-based data mining*. *IEEE Intelligent Systems* **15**(2) (2000) 32–41
- [34] Talbot, P.J.: *Semantic networks: A unifying framework for multistrategy reasoning*. In: *Technology Review Journal-2003*
- [35] Stamatatos, E.: *Author identification using imbalanced and limited training texts*. In: *Proceedings of the 18th International Conference on Database and Expert Systems Applications. DEXA '07, Washington, DC, USA, IEEE Computer Society (2007)* 237–241
- [36] Catherine Marneffe, M., Maccartney, B., Manning, C.D.: *Generating typed dependency parses from phrase structure parses*. In: *In LREC 2006*
- [37] Plakias, S., Stamatatos, E.: *Tensor space models for authorship identification*. In: *Proceedings of the 5th Hellenic conference on Artificial Intelligence: Theories, Models and Applications. SETN '08, Berlin, Heidelberg, Springer-Verlag (2008)* 239–249
- [38] Escalante, H.J., Solorio, T., Montes-y Gómez, M.: *Local histograms of character n-grams for authorship attribution*. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. HLT '11, Stroudsburg, PA, USA, Association for Computational Linguistics (2011)* 288–298
- [39] Bird, S., Klein, E., Loper, E.: *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly, Beijing (2009)
- [40] Joachims, T.: *Text categorization with support vector machines: Learning with many relevant features*. In: *Proceedings of the 10th European Conference on Machine Learning. ECML '98, London, UK, UK, Springer-Verlag (1998)* 137–142
- [41] Dai, W., Xue, G.R., Yang, Q., Yu, Y.: *Transferring naive bayes classifiers for text classification*. In: *AAAI*. 540–545

- [42] Mikros, G.K., Perifanos, K.: *Authorship identification in large email collections: Experiments using features that belong to different linguistic levels - notebook for pan at clef 2011*. In Petras, V., Forner, P., Clough, P.D., eds.: *CLEF (Notebook Papers/Labs/Workshop)*
- [43] Argamon, S., Juola, P.: *Overview of the international authorship identification competition at pan-2011*. In: *CLEF (Notebook Papers/Labs/Workshop)*
- [44] Lewis, D.D., Yang, Y., Rose, T.G., Li, F.: *Rcv1: A new benchmark collection for text categorization research*. *J. Mach. Learn. Res.* **5** (December 2004) 361–397
- [45] Hernández Orallo, J., Ramírez Quintana, M.J., Ferri Ramírez, C.: *Introducción a la Minería de Datos*. Pearson Educación (2004)
- [46] Olmos, I., Gonzalez, J.A., Osorio, M.: *Subgraph isomorphism detection using a code based representation*. In: *FLAIRS Conference*. 474–479
- [47] Yan, X., Han, J.: *gspan: Graph-based substructure pattern mining*. In: *ICDM*. 721–724
- [48] Salton, G., Wong, A., Yang, C.: *A vector space model for automatic indexing*. *Communications of the ACM* **18**(11) (1975) 613–620
- [49] Vilarino, D.A., Castillo, E., Pinto, D., León, S., Tovar, M.: *Baseline approaches for the authorship identification task- notebook for pan at clef 2011*. In: *CLEF (Notebook Papers/Labs/Workshop)*
- [50] Vilarino, D.A., Castillo, E., Pinto, D., Olmos, I., León, S.: *Information retrieval and classification based approaches for the sexual predator identification- notebook for pan at clef 2012*. In: *CLEF (Online Working Notes/Labs/Workshop)*
- [51] Castillo, E., Vilarino, D.A., Pinto, D., Olmos, I., Gonzalez, J.A., Carrillo, M.: *Graph-based and lexical-syntactic approaches for the authorship attribution task- notebook for pan at clef 2012*. In: *CLEF (Online Working Notes/Labs/Workshop)*
- [52] Castillo, E., Vilarino, D.A., Pinto, D., Carrillo, M., Olmos, I.: *Evaluación de características para el proceso de atribución de autoría*. *Research in computing science, Avances en inteligencia artificial* **55** (2012) 37–44

