SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Tecnológico Nacional de México

Centro Nacional de Investigación y Desarrollo Tecnológico

Tesis de Doctorado

Detección automática de cambio de estilo de escritura utilizando aprendizaje automático

presentada por

MC. Germán Ríos Toledo

como requisito para la obtención del grado de

Doctor en Ciencias de la Computación

Director de tesis

Dr. Noé Alejandro Castro Sánchez

Codirector de tesis

Dr. Grigori Sidorov

Cuernavaca, Morelos, México. Junio de 2019.

SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Centro Nacional de Investigación y Desarrollo Tecnológico

"2019, Año del Caudillo del Sur, Emiliano Zapata"

ESC\FORDOC09

Cuernavaca, Morelos, 18/junio/2019

ASUNTO: ACEPTACIÓN DEL TRABAJO DE TESIS DOCTORAL

DR. NOÉ ALEJANDRO CASTRO SÁNCHEZ
JEFE DEL DEPARTAMENTO DE CIENCIAS COMPUTACIONALES
PRESENTE

Los abajo firmantes, miembros del Comité Tutorial de la Tesis Doctoral del alumno M.C. GERMÁN RÍOS TOLEDO, manifiestan que después de haber revisado su trabajo de tesis doctoral titulado "DETECCIÓN AUTOMÁTICA DE CAMBIO DE ESTILO DE ESCRITURA UTILIZANDO APRENDIZAJE AUTOMÁTICO", realizado bajo la dirección del DR. NOÉ ALEJANDRO CASTRO SÁNCHEZ y co-dirección del DR. GRIGORI SIDOROV, el trabajo se ACEPTA para proceder a su impresión.

ATENTAMENTE
EXCELENCIA EN EDUCACIÓN TECNOLÓGICA®
"CONOCIMIENTO Y TECNOLOGÍA AL SERVICIO DE MÉXICO"

DR. NOÉ ALEJANDRO CASTRO SÁNCHEZ
CENIDET

DR. GRIGORI SIDOROV
CIC IPN

DRA. ALICIA MARTÍNEZ REBOLLAR
CENIDET

DR. DANTE MÚJICA VARGAS
CENIDET

DR. JOAQUÍN PÉREZ ORTEGA
CENIDET

DR. ALEXANDER GELBUKH
CIC IPN

C.c.p.: M.E. Guadalupe Garrido Rivera / Jefa del Depto. de Servicios Escolares
Dr. Gerardo Vicente Guerrero Ramírez / Subdirector Académico
Expediente

"2019, Año del Caudillo del Sur, Emiliano Zapata"

ESC\FORDOC010

Cuernavaca, Morelos, 18/Junio/2019

M.C. GERMÁN RÍOS TOLEDO
CANDIDATO AL GRADO DE DOCTOR
EN CIENCIAS DE LA COMPUTACIÓN
PRESENTE

Después de haber sometido a revisión su trabajo final de tesis titulado "**DETECCIÓN AUTOMÁTICA DE CAMBIO DE ESTILO DE ESCRITURA UTILIZANDO APRENDIZAJE AUTOMÁTICO**", y habiendo cumplido con todas las indicaciones que el jurado revisor de tesis le hizo, le comunico que se le concede autorización para que proceda a la impresión de la misma, como requisito para la obtención del grado.

Reciba un cordial saludo.

ATENTAMENTE
EXCELENCIA EN EDUCACIÓN TECNOLÓGICA®
"CONOCIMIENTO Y TECNOLOGÍA AL SERVICIO DE MÉXICO"

DR. NOÉ ALEJANDRO CASTRO SÁNCHEZ
JEFE DEL DEPTO. DE CIENCIAS COMPUTACIONALES

Agradecimientos

A mi director de tesis, el Dr. Noé Alejandro Castro Sánchez, mi agradecimiento por su dedicación en este trabajo doctoral, por la motivación que me transmitió y la confianza depositada en mí cuando me aceptó como su estudiante.

Al Dr. Grigori Sidorov, del cual tuve la oportunidad de aprender mucho de sus conocimientos durante mi estancia en el Centro de Investigación en Computación del IPN y a lo largo de estos cuatro años. Gracias Doctor.

A los distinguidos integrantes de mi comité tutorial: Dra. Alicia Martínez Rebollar, Dr. Joaquín Pérez Ortega, Dr. Dante Mújica Vargas y el Dr. Alexander Gelbukh, quienes con mucha paciencia me aconsejaron y me indicaron los caminos a seguir para enriquecer mi trabajo de tesis. Muchas gracias por compartirme toda su experiencia.

Al Tecnológico Nacional de México (TecNM) y al Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), por abrirme sus puertas para mejorar mis conocimientos y contribuir en mi formación académica.

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo económico que me brindó para realizar estos estudios.

Al Instituto Tecnológico de Tuxtla Gutiérrez, en especial al M.E.H. José Luis Méndez Navarro y a la Ing. Delina Culebro Farrera, por todo el apoyo que me brindaron todo este tiempo. Asimismo, agradezco a mis compañeros de la Academia de Ingeniería en Sistemas Computacionales de dicho Instituto, por darme su anuencia para iniciar este proyecto.

A ti, mi Reina, Dayani Bravo, gracias porque tuviste la entereza de apoyarme, aun sabiendo que esto significaba no vernos todos los días. A pesar de las circunstancias, seguimos juntos y felices. A Don JoseStalin y Doña Cielo, les doy las gracias por traer al mundo a una mujer tan linda.

A mis papas, Justo Ríos, Teodorita Toledo; a mis hermanos Florentino Ríos, Justita Ríos; a mis abuelitos Germán Toledo y Berta Aquino, gracias por ser como son y por la manera en que viven la vida. Mi más profunda admiración para ustedes.

A mis tíos: Juanita Aquino, Casimiro Cruz, Anabel Cruz y Andrea Cruz, les doy las gracias por todo su cariño y sus “terapias” aleccionadoras :)

Tabla de contenido

Pág.

Lista de figuras

Lista de tablas

Resumen

El objetivo principal de esta tesis es determinar el cambio de estilo de escritura a través del tiempo por medio de una característica estilométrica conocida como n-gramas, los cuales están por formados con caracteres, palabras, etiquetas POS y relaciones sintácticas. Los n-gramas se obtuvieron de un conjunto de novelas de autores de habla inglesa, con carreras literarias de alrededor de 30 años. Las novelas se organizaron de forma cronológica desde la más antigua a la más reciente. Se predefinieron tres etapas procurando que la duración de estas fuera proporcional al periodo que comprendían las novelas evaluadas.

En el contexto de esta investigación, el cambio de estilo de escritura se refiere a la variación de la frecuencia de uso de n-gramas entre las etapas. La detección de cambio de estilo se abordó como un problema de clasificación supervisada con el enfoque de aprendizaje automático. Los algoritmos de aprendizaje automático entrenan y aprenden patrones de escritura para representarlos en modelos de inferencia. En la fase de clasificación, los modelos se evaluaron con textos “no vistos” de los mismos autores en la etapa de aprendizaje. Idealmente se espera que todas las muestras sean asignadas correctamente a la etapa a la que pertenecen.

Bajo las condiciones previamente establecidas, el esquema propuesto permitió confirmar la hipótesis de que el estilo de escritura cambia a través de tiempo. Los distintos tipos de n-gramas identificaron cambios significativos en el estilo de los autores. Se encontró que los n-gramas sintácticos de relaciones de dependencia son una excelente opción para caracterizar el estilo de escritura de un autor. Otras disciplinas del Procesamiento del Lenguaje Natural, tales como: Atribución de Autoría, Identificación de Autoría, creación de perfiles de autor y detección de plagio, reportan que las características sintácticas tienen la peculiaridad de que su manipulación consciente es difícil y además, son independientes al tema que se está tratando en los documentos.

Capítulo 1. Introducción

La *estilometría* es una disciplina que se basa en la presunción de que cada persona tiene un estilo de escritura. En otras palabras, es una forma de reconocimiento del estilo de escritura que se basa en la información lingüística que se encuentra en un documento (Brennan, Afroz y Greenstadt, 2012). Esto resulta de gran utilidad en áreas como el derecho penal y civil debido a que ayuda a la detección de plagio, la creación de perfiles de autor y a la protección del anonimato. En términos computacionales, el estilo de escritura se refiere a *la frecuencia* de uso de elementos del texto conocidos como *características estilométricas*.

Generalmente, el estilo utilizado en la producción lingüística de un hablante o escritor mostrará algún tipo de variación a través del tiempo (Turell y Gavaldà, 2013). Dicha variación puede ocurrir por factores sociales, individuales e incluso geográficos. Aspectos como el género, la edad y el nivel educativo, también influyen en el uso del lenguaje. El lenguaje individual es el resultado del contacto del individuo con el resto de los miembros de su entorno lingüístico a lo largo de su vida, esta interacción influye en la ideología y los hábitos lingüísticos de una persona.

En el análisis del cambio de estilo de escritura, comúnmente se utilizan las palabras y lo que a ellas concierne: palabras con contenido semántico (sustantivos, verbos, adjetivos, adverbios); palabras funcionales (preposiciones, adverbios, artículos, pronombres, adjetivos); longitud de palabras; categorías gramaticales; errores de escritura; lemas; entre otras.

Actualmente, el desafío principal en el análisis de estilo es identificar características estilométricas apropiadas para cada tarea en particular. Dado que la estilometría se enfoca principalmente en la forma del texto y no en su contenido, es importante explorar la funcionalidad de características estilométricas independientes de la temática. En este sentido, el uso de la información sintáctica es una alternativa potencial. Los analizadores sintácticos modernos obtienen la información sintáctica de cada oración y la representan

de forma estructurada en un *árbol de constituyentes* o en un *árbol de dependencia*. Dichos árboles muestran el orden no lineal entre las palabras, las categorías gramaticales y los nombres de las relaciones sintácticas existentes. Las características estilométricas basadas en información sintáctica posibilitan el desarrollo de análisis de estilo más completos.

En esta investigación para la detección de cambio de estilo de escritura a través del tiempo, se propone el uso de una característica estilométrica denominada *n-grama sintáctico*, la cual se obtiene de una representación conocida como *árbol de dependencias*. Esta característica es robusta a la influencia temática y ha sido utilizada en las tareas de atribución de autoría y detección de plagio con resultados favorables. Además, en el análisis incluye otros tipos de características estilométricas, particularmente *n*-gramas de caracteres, palabras y categorías gramaticales de palabras. A las categorías gramaticales también se les conoce como etiquetas POS (*Part Of Speech*). La tarea de detección de cambio de estilo se planteó como un problema de clasificación bajo un enfoque de aprendizaje automático supervisado. Para la evaluación de la propuesta se utilizó un corpus de novelas de escritores con largas carreras literarias. Los resultados mostraron un cambio de estilo significativo en todos los autores y en los distintos n-gramas. Los *n*-gramas de relaciones sintácticas mostraron resultados similares a las otras características propuestas.

1.1 Planteamiento del problema

El análisis de estilo de escritura muestra la presencia de patrones generados de forma consciente o inconsciente por parte del autor. Este conjunto de patrones conforma su estilo de escritura. No obstante, dicho estilo puede sufrir cambios con el paso del tiempo debido a factores como la edad del autor, nivel educativo, el periodo o época en que escribió los textos, un mayor dominio del lenguaje, cambios de género literario, entre otras causas. Generalmente, estos cambios ocurren de manera gradual.

El problema radica en que los enfoques actuales para la detección de cambio de estilo de escritura se basan principalmente en el estudio del vocabulario empleado. Sin embargo, estas características no son inmunes al control consciente del autor y además pueden variar en función del tópico del texto. No obstante, una alternativa prometedora son las características basadas en información sintáctica, que ya han sido propuestas en trabajos relacionados al cambio de estilo de escritura, a saber: profundidad del árbol, complejidad sintáctica, longitud de sentencias (en palabras), voz activa, voz pasiva, oraciones simples y compuestas, clausulas por sentencias, verbos principales, subordinados y embebidos. El análisis sintáctico arroja otras alternativas aún no exploradas para resolver esta tarea, como lo es el uso de una estructura conocida árbol de dependencias. El árbol de dependencias tiene la cualidad de mostrar la relación o *dependencia* entre pares de palabras, que no necesariamente son vecinas en una oración. Este hecho permite identificar patrones que no se detectan a simple vista.

En esta investigación se propone el uso de analizadores sintácticos para procesar la información del árbol de dependencias y obtener una característica estilométrica llamada *n*-gramas sintácticos. Además, para un análisis más confiable, se proponen otros tipos de *n*-gramas: de caracteres, palabras y etiquetas POS. Todos estos n-gramas se procesan desde un enfoque de aprendizaje automático supervisado, lo que permite construir modelos confiables para identificar cambios de estilo de escritura.

1.2 Justificación

Los sociolingüistas han demostrado durante décadas que el lenguaje está sometido a un constante cambio a través del tiempo. Además, estos cambios ocurren incluso en la forma particular en que cada persona los utiliza (Labov 1972, Chambers 2008). Turell y Gavaldà (2013) indicaron que el estilo en la producción lingüística de un hablante o escritor generalmente mostrará algún cambio a través del tiempo en función de su experiencia o de nuevos conocimientos.

La detección de los cambios de estilo por las causas previamente mencionadas se ha convertido en un asunto de interés dentro del campo de Procesamiento de Lenguaje Natural, a saber: la detección de plagios de documentos, la creación de perfiles de autor, la identificación de autores y la atribución de autoría. Otras tareas que abordan los cambios de estilo son: la predicción de la personalidad (Luyckx y Daelemans, 2008); la detección de tendencias a la depresión (Rude, Gortner y Pennebaker, 2004); y el aprendizaje de una segunda lengua (Yoon y Bhat, 2012). Otra aplicación importante se relaciona con enfermedades neurológicas que repercuten en el uso del lenguaje (Williams et al. 2003, Garrard et al. 2005, Lancashire y Hirst 2009, Le 2010, Hirst y Wei Feng 2012).

Por otro lado, el cambio de estilo de escritura a través del tiempo es un aspecto importante a considerar en la tarea de atribución de autoría: existe la posibilidad de que entre los momentos en los que fueron escritos los documentos haya transcurrido un considerable período de tiempo. Si un autor tiende a variar de forma significativa el uso de ciertas características entre una obra, las pruebas de atribución podrían no ser concluyentes.

1.3 Hipótesis

Es posible identificar cambios de estilo de escritura a través del tiempo utilizando características estilométricas léxicas, morfológicas y sintácticas haciendo uso del enfoque de aprendizaje automático supervisado.

1.4 Objetivo general

Crear un modelo basado en un enfoque automático supervisado para la detección automática de cambio de estilo de escritura analizando un corpus de novelas de varios autores.

1.5 Objetivos específicos

- Construir modelos de aprendizaje a partir de diferentes características estilométricas.
- Obtener un modelo sobre el estilo en términos de los n-gramas sintácticos.
- Evaluar la viabilidad de la aplicación de algoritmos de reducción de dimensiones por selección basada en frecuencia.
- Evaluar la viabilidad de la aplicación de algoritmos de reducción de dimensiones por extracción de características.
- Evaluar la eficiencia de los distintos tipos de *n*-gramas, ya sea de forma individual o combinada.

1.6 Organización del documento

El documento está organizado de la siguiente manera: el Capítulo 2 describe el marco teórico, el Capítulo 3 proporciona al lector una serie de trabajos relacionados con la detección automática de cambio de estilo de escritura, el Capítulo 4 expone la propuesta metodológica para la detección de cambio de estilo de escritura, el Capítulo 5 contiene la experimentación y evaluación de resultados y el Capítulo 6 las conclusiones y trabajos futuros

Capítulo 2. Marco teórico

En este capítulo se dan a conocer los campos relevantes de investigación en el análisis estilométrico de textos.

2.1 Análisis de estilo

Desde el punto de vista computacional, el término *estilo* se refiere al análisis de la frecuencia de uso de elementos del texto conocidos como *características estilométricas*. Es importante destacar que el término “estilo” tiene otras acepciones dentro del campo del análisis automático de textos. A continuación, se mencionan dos de ellas. El estilo se conceptualiza como las diferencias sutiles pero regulares entre textos que idealmente comparten lenguaje, género y tema, pero difieren con respecto a la autoría (Golcher, 2007). El estilo hace referencia a los elementos lingüísticos que independientemente del contenido del documento, persisten a lo largo de todos los trabajos de un autor (Uzuner y Katz, 2005).

Los estilos de habla y comunicación escrita no solo incluyen aspectos de identidad, origen étnico, edad, género y origen social, sino también indican los contextos en los que se usa el idioma. Al mismo tiempo, la forma en que las personas escriben cartas, correos electrónicos, mensajes de texto y entradas de *blog*s, indica la conciencia de las diferentes audiencias de estos diferentes géneros. Adaptamos el lenguaje a nuestra audiencia (J. Holmes, 2013).

Los factores antes mencionados influyen en las elecciones (que consciente o inconscientemente) una persona hace al momento de componer un texto, es probable que esas elecciones tendrán una presencia repetitiva en el resto del texto o en los diferentes textos del autor (Alzahrani, Salim y Abraham, 2012). Desde este enfoque, un texto puede considerarse como una secuencia de elecciones realizadas sobre los elementos del lenguaje para expresar una idea. A este patrón de elecciones sobre el uso del lenguaje se le conoce como *estilo de escritura* del autor. Así, el análisis del estilo de

escritura consiste en proponer características estilométricas que al ser cuantificadas permitan identificar al autor del texto.

2.2 Características estilométricas

Una *característica estilométrica* hace referencia a un rasgo de la forma que un autor compone sus textos. La Tabla 2.1 muestra la clasificación de algunas características estilométricas de acuerdo con el tipo de información lingüística que representan.

Tabla 2.1. Clasificación de características estilométricas.

Tipo de característica	Descripción
Caracteres	Alfabéticos y numéricos, letras mayúsculas y minúsculas, marcas de puntuación, *n*-gramas a nivel de carácter.
Léxicas	Longitud de palabras, longitud de sentencias, riqueza de vocabulario, hapaxes (palabras que aparecen una y dos veces), palabras frecuentes, *n*-gramas de palabras, errores de escritura.
Sintácticas	Etiquetas POS, estructura de sentencia y frase, racimos, reglas de reescritura.
Semánticas	Sinónimos, hiperónimos, dependencias semánticas.
Específicas de aplicación	Estructurales (características HTML como tipo, tamaño y color de la letra.), específicas de contenido (tipo y número de emoticones utilizados.), especificas del lenguaje (forma de saludo y despedida, tipo de firmas, indentación, modismos)

Una forma simple y natural de ver un texto es como una secuencia de elementos (palabras, dígitos, signos de puntuación) agrupados en oraciones. Un análisis de estilo común consiste en utilizar una lista de palabras. La lista de palabras no considera la información contextual que acompaña a cada palabra ni el orden en que aparecen. Sin embargo, estas características capturan efectivamente las correlaciones entre autores y temas (Genkin y Lewis 2005; Kaster, Siersdorfer y Weikum 2005).

Las *palabras funcionales* expresan una relación estructural o gramatical con otras palabras y no aportan información sobre la temática del texto: artículos, pronombres personales, preposiciones, conjunciones y disyunciones. Los trabajos de (Burrows 1987; DavidI. Holmes, Robertson y Paez 2001; Binongo 2003) utilizaron este tipo de palabras en la tarea de atribución de autoría. Las *palabras de contenido* proveen el contenido temático de un texto: sustantivos, verbos, adjetivos y adverbios. Los trabajos de Burrows

(1987) y Joula (2007), utilizaron estas características para clasificación de textos basada en tópico.

Las etiquetas POS también son características útiles para el análisis de estilo. Las etiquetas POS para una palabra pueden ser sustantivo, verbo, adjetivo, pronombre, entre otros. Por ejemplo, en el proyecto Penn Treebank[1] la nomenclatura indica que CC es conjunción coordinada, NNS sustantivo plural, DT determinante, PRP pronombre personal. La descripción de cada etiqueta POS se encuentra en el Anexo A. Las etiquetas pueden incluir características morfológicas más detalladas. Por ejemplo, la etiqueta VIP1S, podría significar "verbo, indicativo, presente, primera persona, singular". El etiquetado se aplica de igual forma a signos de puntuación, números, cantidades, entre otros. Las etiquetas se obtienen por medio de programas conocidos como *etiquetadores de partes de la oración*.

2.3 Árbol de dependencias

Para representar la estructura sintáctica de una oración existe un formalismo denominado *gramática de dependencias*. La gramática de dependencias muestra *la relación* entre pares de palabras, donde una de ellas es la palabra rectora y la otra palabra es dependiente de la primera. La información sobre las relaciones de dependencia entre las palabras también se puede representar de manera gráfica por medio de un *árbol de dependencias*. Considerando la oración "*John smoked with a little more dignity and surveyed them in silence*", el analizador sintáctico *Stanford Parser*[2] genera las relaciones de dependencia que se muestran en la Tabla 2.2.

[1] http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

[2] https://nlp.stanford.edu:8080/parser/

Tabla 2.2 Ejemplo de relaciones de dependencia.

Relaciones de dependencia
nsubj(smoked-2,john-1)
root(ROOT-0,smoked-2)
prep(smoked-2,with-3)
det(dignity-7,a-3)
advmod(more-6,little-5)
amod(dignity-7,more-6)
pobj(with-3,dignity-7)
cc(smoked-2,and-8)
conj(smoked-2,surveyed-9)
dobj(surveyed-9,them-10)
prep(surveyed-9,in-11)
pobj(in-11,silence-12)

Esta representación muestra al inicio el nombre de la relación, el primer argumento dentro de los paréntesis representa al elemento rector y el segundo al dependiente. Así, *amod(dignity-7, more-6)* significa que hay una relación o dependencia desde *dignity* a *more* con una relación de modificador adjetival (*amod*). Todas las posibles relaciones sintácticas se definen en el Anexo B. Los números en las palabras indican la posición que estas tienen dentro de la oración. La información de la Tabla 2.2 se representa de forma gráfica en un árbol de dependencias como se muestra en la Figura 2.1.

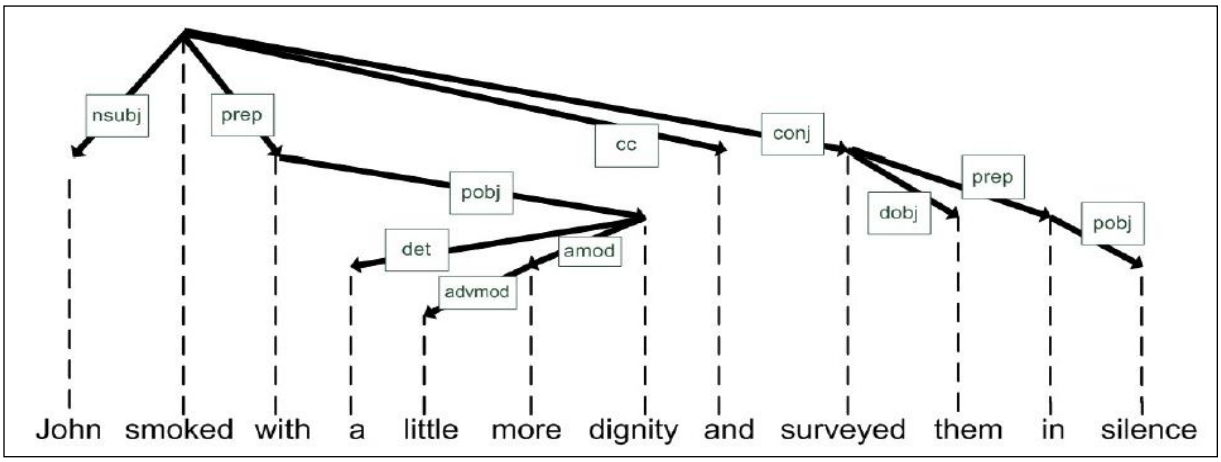


Figura 2.1 Árbol de dependencias.

2.4 Modelado del lenguaje por medio de n-gramas

Los *n-gramas* son una característica estilométrica ampliamente utilizada en el análisis de estilo de escritura. Un *n*-grama es una secuencia de elementos que se obtienen siguiendo el orden lineal que mantienen en el texto. Los elementos pueden ser caracteres, palabras o etiquetas POS. Los *n*-gramas suponen una ventana imaginaria de tamaño *n* que se desplaza *n* elementos en cada iteración, hasta alcanzar el final del texto. Los elementos que se encuentran dentro de la ventana corresponden a un *n*-grama. Sidorov (2013) denominó a estos *n*-gramas como "tradicionales" porque se crean siguiendo el orden lineal en que los elementos aparecen en el texto.

Los *n*-gramas de caracteres obtienen las frecuencias de caracteres alfabéticos, dígitos, mayúsculas, minúsculas, signos de puntuación, pueden capturar prefijos, sufijos y subcadenas de palabras. Estas características han demostrado ser bastante útiles para cuantificar el estilo de escritura (Grieve, 2007), sin embargo no toman en cuenta la información contextual. Conforme se incrementa el valor de n, los n-gramas de caracteres pueden capturar información léxica, como por ejemplo palabras en inglés formadas con tres caracteres: *you*, *the*, *for*, *was*, entre otras.

Si $n \geq 2$, las frecuencias de n-gramas de palabras son más bajas, ya que la probabilidad de que dos o más palabras aparezcan juntas es mucho menor. Si el análisis de estilo requiere más detalles acerca de las palabras, es posible lematizarlas para conocer a raíz y sus derivaciones. Las Tabla 2.3, Tabla 2.4 y Tabla 2.5 muestran ejemplos de *n*-gramas de caracteres, palabras y etiquetas POS obtenidos de la oración "*Juan lee un libro interesante*".

Tabla 2.3. n-gramas de caracteres de la oración "Juan lee un libro interesante".

n	**n-gramas de caracteres**
1	J, u, a, n, l, e, e, u, n, l, i, b, r, o, i, n, t, e, r, e, s, a,n, t, e
2	ju, ua, an, nl, le, ee, eu, un, nl, li, ib, br, ro, oi, in, nt, te, er, rs, sa, an, nt, te
3	jua, anl, lee, eun, nli, ibr, roi, int, ter, rsa, ant
4	Juan, nlee, eunl, libr, roin, nter, resa,ante

Tabla 2.4. n-gramas de palabras de la oración "Juan lee un libro interesante".

n	n-gramas de palabras
1	Juan, lee, un, libro, interesante
2	Juan lee, lee un, un libro, libro interesante
3	Juan lee un, un libro interesante
4	Juan lee un libro

Tabla 2.5. n-gramas de etiquetas POS de la oración "Juan lee un libro interesante".

n	n-gramas de etiquetas POS
1	NP, VM, DI, NC, AQ
2	NP VM, VM DI, DI NC, NC AQ
3	NP VM DI, DI NC AQ
4	NP VM D NC

No es posible generalizar cuál es el valor apropiado para la longitud de un n-grama. Hourvardas y Stamatatos (2006) afirmaron que la selección de un valor óptimo de n depende del lenguaje. Los valores comúnmente utilizados son $n = \{2, 3, 4, 5\}$, véase Barrón-Cedeño (2009), Kešelj *et al*. (2003) y Sidorov et al. (2012). Conforme el valor de n aumenta, la frecuencia de uso disminuye y la probabilidad de encontrar n-gramas comunes en distintos documentos decrece.

2.5 N-gramas sintácticos

Los elementos que conforman este tipo de n-gramas no se obtienen conforme al orden de aparición en el texto, sino al orden en que aparecen en el árbol de dependencias. Los n-gramas sintácticos representan las relaciones entre palabras desde un punto de vista sintáctico. Los elementos pueden ser palabras, etiquetas POS y relaciones sintácticas o incluso combinaciones de ellos (Durán, 2017). Algunas de las ventajas de los n-gramas sintácticos son:

- Se aprovechan las relaciones sintácticas entre palabras.
- Cada palabra se usa con sus vecinos basados en el árbol sintáctico.
- Permiten ignorar los fenómenos superficiales del lenguaje.
- Las palabras auxiliares pueden ignorarse durante del proceso de generación.

- Permiten introducir información lingüística dentro de los métodos basados en estadística y los métodos de aprendizaje automático.
- Se aplican de la misma forma que los otros tipos de en tareas del Procesamiento del Lenguaje Natural.

En la Tabla 2.6 se muestran los 2-gramas de palabras que se obtienen de la oración "*John smoked with a little more dignity and surveyed them in silence*" al recorrer el árbol de dependencias previamente mostrado en la Figura 2.1. Como sucede con el resto de las características estilométricas, los *n*-gramas sintácticos frecuentes son las más importantes para propósitos el análisis de estilo.

Tabla 2.6. *3*-gramas tradicionales de palabras y *3*-gramas sintácticos de palabras.

***3*-gramas tradicionales de palabras**	**3-gramas sintácticos de palabras**
John smoked with, with a little, little more dignity, dignity and surveyed, surveyed them in	smoked-John-with, smoked-John-and, smoked-John-surveyed, smoked-with-and, smoked-with-surveyed, smoked-and-surveyed, smoked-surveyed-in, smoked-with-dignity, smoked-surveyed-them, surveyed-them-in, surveyed-in-silence, with-dignity-a, with-dignity-more, dignity-a-more, dignity-more-little

Los *n*-gramas sintácticos se obtienen a partir de la salida generada por un analizador sintáctico (también conocido como *Parser*). Uno de los analizadores más conocidos es *Stanford*, este analizador genera dos tipos de salidas: el árbol de estructuras, donde el nivel del componente se representa por sus espacios identificadores, y las relaciones sintácticas que corresponden al árbol de dependencias.

2.6 El Modelo Espacio Vectorial

El Modelo Espacio Vectorial se utiliza para representar objetos por medio de sus características. Conceptualmente, el modelo es un espacio de N dimensiones donde a cada característica le corresponde una dimensión. En la práctica, dicho modelo es una matriz de dos dimensiones, donde las filas representan objetos, las columnas representan las características y las celdas la frecuencia de éstas. Este modelo permite representar documentos por medio de cualquier característica estilométrica. Además de

frecuencias, pueden utilizarse valores booleanos para indicar presencia o ausencia de una característica o bien los valores *tf-idf*.

2.7 Los valores *tf-idf*

Se denomina frecuencia de término (*term frequency*, *tf*) al número de veces que un término aparece en el texto. En primera instancia, cuanto más frecuente es un término, más importante en ese documento (Sidorov, 2013). Sin embargo, si una palabra se encuentra en toda una colección de documentos entonces es incapaz de distinguir entre los documentos, y por lo tanto tiene poca utilidad. En el otro sentido, si una palabra se encuentra exactamente en un documento, es una palabra muy útil para cálculos de similitud. Considerando estos casos extremos, *tf* se combina con una medida llamada frecuencia inversa de documento (*inverse document frequency*, *idf*). El *idf* se calcula para cada palabra en una colección mediante la Fórmula 2.1.

$$idf_t = log \frac{N}{df_t} \tag{2.1}$$

Donde N es el número de documentos en la colección, df_t es el número de documentos en el que aparece el término y idf_t es la frecuencia inversa del documento. Se recomienda combinar *tf* e *idf* de una palabra en cada documento (*tf-idf*). De acuerdo con (Manning and Raghavan 2009), el valor *tf-idf* para un término puede ser: alto cuando el termino aparece muchas veces en un número pequeño de documentos, bajo cuando aparece pocas veces en un documento y muy bajo cuando el término aparece en todos los documentos.

El método de *tf-idf* permite la discriminación de palabras frecuentes, como lo son las palabras funcionales, asignándoles pesos nulos o bajos. Otro de los aspectos a considerar al momento de utilizar las frecuencias es la longitud del documento. Una de

las modificaciones que han sido agregadas al método de *tf-idf* es el empleo de un factor de normalización que permita equilibrar casos en los que se empleen documentos de distintos tamaños. Existen dos razones principales para emplear la normalización: la primera es que los documentos grandes usualmente emplean los mismos términos repetidamente, como resultado, los factores de frecuencia de los términos pueden ser grandes para documentos largos. La segunda es que los documentos largos también tienen una gran cantidad de términos diferentes. La normalización, a grandes rasgos, permite tratar de la misma manera a todos los documentos sin importar su longitud.

2.8 Aprendizaje automático

A la construcción de programas de computadora que aprendan y mejoren automáticamente con la experiencia se le conoce como Aprendizaje Automático. El proceso de aprendizaje automático consiste en encontrar la relación entre los patrones y los resultados utilizando únicamente los ejemplos de entrenamiento. El objetivo central del aprendizaje automático es *el aprendizaje* y *la inferencia*. La Figura 2.2 muestra el proceso de aprendizaje: a partir de los datos de entrenamiento se obtienen los *vectores de características*, el programa aprende al descubrir patrones y el conocimiento adquirido se resume en un modelo. A la lista de atributos utilizados para resolver un problema se denomina vector de características.

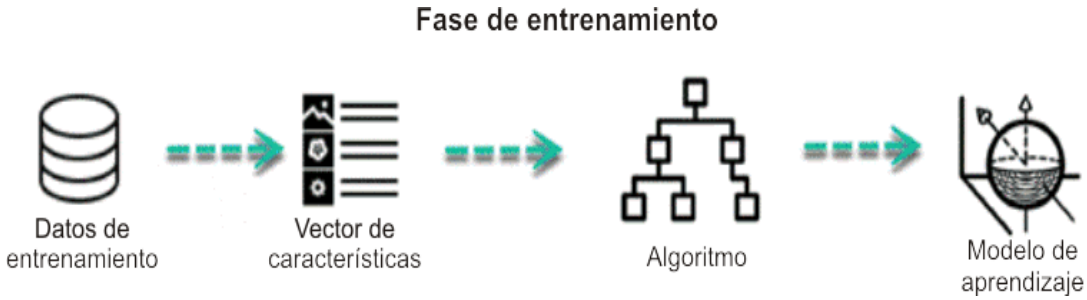


Figura 2.2 Etapa de aprendizaje.

Ahora, el modelo debe evaluarse con datos que nunca ha visto. La Figura 2.3 muestra el proceso de inferencia. Los nuevos datos se transforman en un vector de características, pasan por el modelo y ofrecen una predicción.

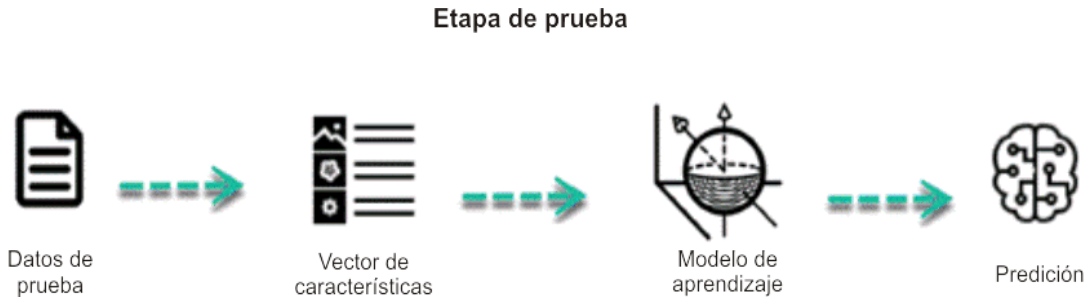


Figura 2.3 Etapa de inferencia.

En esta investigación se utilizan algoritmos de aprendizaje automático supervisados. El aprendizaje supervisado tiene variables de entrada y variables de salida (previamente etiquetadas) y utiliza un algoritmo para derivar la función de mapeo de la entrada a la salida. En el aprendizaje supervisado, existe necesariamente la suposición de que variables las variables predictivas estan de alguna manera relacionados con la variable objetivo o dependiente.

2.8.1 Regresión Logística

La regresión logística es un método de clasificación multiclase. Se usa normalmente cuando la variable dependiente es dicotómica y las variables independientes son continuas o categóricas. Cuando la variable dependiente se compone de más de dos categorías, se puede emplear una regresión logística multinomial. La Regresión logística mide la relación entre la variable dependiente y una o más variables independientes, al estimar las probabilidades utilizando la función logística subyacente.

Estas probabilidades se deben transformar en valores binarios para realizar una predicción. Esta es la tarea de la función logística, también llamada *función sigmoidea*. La función sigmoidea es una curva en forma de S que puede tomar cualquier número de valor real y asignarlo a un valor entre el rango de 0 y 1, pero nunca exactamente en esos

límites. Estos valores entre 0 y 1 se transformarán utilizando un clasificador de umbrales. La Figura 2.4 muestra los pasos descritos previamente.

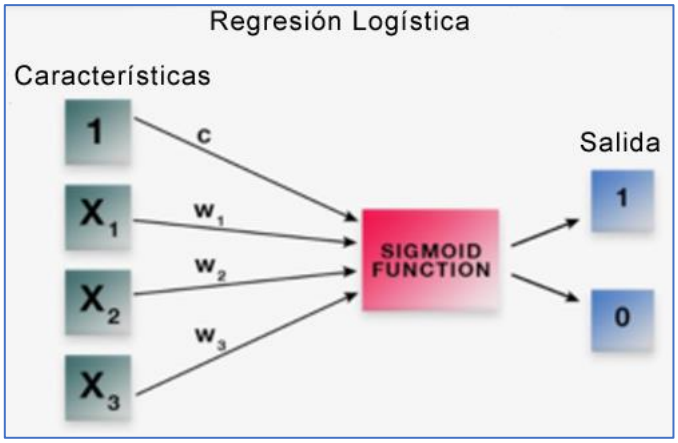


Figura 2.4 El proceso de Regresión Logística.

Para su uso, es necesario que los datos sean linealmente separables. Sea

$$p = \frac{1}{1 + b^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}}$$

Donde p es la probabilidad esperada, x_1 a x_p son las distintas variables independientes, β_0 a β_p representan los coeficientes de regresión. En la regresión logística, los coeficientes derivados del modelo (por ejemplo, β_1) indican el cambio en las probabilidades de registro esperadas en relación con un cambio de una unidad en x_1, manteniendo constantes los demás predictores. Algunos supuestos principales en regresión logística son: la variable dependiente debe ser de naturaleza dicotómica; no deben existir valores atípicos en los datos; no debe haber correlaciones altas entre los predictores.

2.8.2 Máquinas de Soporte Vectorial

En problemas de clasificación, las Máquinas de Soporte Vectorial (SVM, *Support Vector Machines*) se basan en encontrar el hiperplano que proporciona la mayor distancia mínima a los ejemplos de entrenamiento de dos clases. El hiperplano óptimo es en cierto

sentido, equidistante de las dos clases. La notación utilizada para definir formalmente un hiperplano es la siguiente: en un espacio 2-dimensiones, el hiperplano es una línea de la forma $A_0 + A_1X_1 + A_2X_2 = 0$. En un espacio de m-dimensiones, el hiperplano es de la forma $A_0 + A_1X_1 + A_2X_2 + \ldots + A_m m = 0$. En las Máquinas de Soporte Vectorial, un punto de datos se ve como un vector de p-dimensiones. Hay muchos hiperplanos que podrían clasificar los datos. Una opción razonable como el mejor hiperplano es la que representa la mayor separación, o margen, entre las dos clases. Si existe tal hiperplano, se conoce como *hiperplano de margen máximo*. La idea descrita anteriormente se observa en la Figura 2.5.

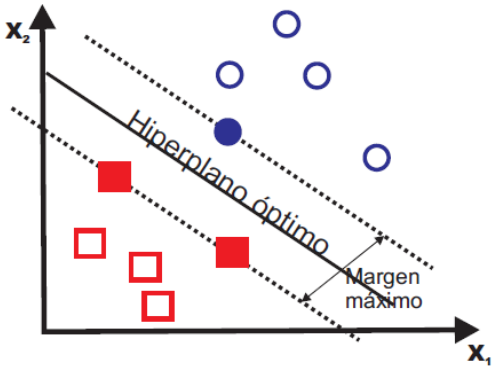


Figura 2.5 Hiperplano óptimo en Máquinas de Soporte Vectorial.

En un conjunto de datos de entrenamiento de n puntos de forma $(\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n)$, donde y_i puede ser 1 o -1 e indica a qué clase pertenece el punto $\vec{x}_i$. Queremos encontrar el "hiperplano de margen máximo" que divide el grupo de puntos $\vec{x}_i$ con $y_i = 1$ del grupo de puntos $\vec{x}_i$ para el cual $y_i = -1$, de manera que se maximice la distancia entre el hiperplano y el punto más cercano $\vec{x}_i$. Cualquier hiperplano puede escribirse como el conjunto de puntos $\vec{x}$ que satisfacen $\vec{w} \cdot \vec{x} - b = 0$. Donde $\vec{w}$ es el vector normal (no necesariamente normalizado) al hiperplano y b es una constante arbitraria. El parámetro $\frac{b}{\|\vec{w}\|}$ determina el desplazamiento del hiperplano desde el origen a lo largo del vector normal $\vec{w}$.

Si los datos de entrenamiento son linealmente separables, podemos seleccionar dos hiperplanos paralelos que separan las dos clases de datos, de modo que la distancia entre ellos sea lo más grande posible. Con un conjunto de datos normalizado o estandarizado, estos hiperplanos se pueden describir mediante las ecuaciones $\vec{w} \cdot \vec{x} - b = 1$ (valores en o por encima de este límite es de una clase) y $\vec{w} \cdot \vec{x} - b = -1$ (valores en o por debajo de este límite es de otra clase). Geométricamente, la distancia entre estos dos hiperplanos es $\frac{2}{\|\vec{w}\|}$, así que para maximizar la distancia entre los planos queremos minimizar $\| \vec{w} \|$. También se debe evitar que los puntos de datos caigan en el margen, por lo cual se agregan las siguientes restricciones:

$$\vec{w} \cdot \vec{x}_i - b \geq 1 \text{ si } y_i = 1$$

$$\vec{w} \cdot \vec{x}_i - b \leq -1 \text{ si } y_i = -1$$

Esto puede reescribirse como $y_i(\vec{w} \cdot \vec{x}_i - b) \geq 1$ para toda $1 \leq i \leq n$. Podemos poner esto juntos para obtener el problema de optimización:

"Minimizar $\| \vec{w} \|$ sujeto a $y_i(\vec{w} \cdot \vec{x}_i - b) \geq 1$ para $i = 1, \ldots, n$."

El vector $\vec{w}$ y b que resuelven este problema determinan nuestro clasificador, $\vec{x} \mapsto sgn(\vec{w} \cdot \vec{x} - b)$. Una consecuencia importante de esta descripción geométrica es que el hiperplano de margen máximo está completamente determinado por aquellos $\vec{x}_i$ que se encuentran más cerca de él. Estos $\vec{x}_i$ son llamados *vectores de soporte.* Los conceptos antes mencionados se muestran en la Figura 2.6.

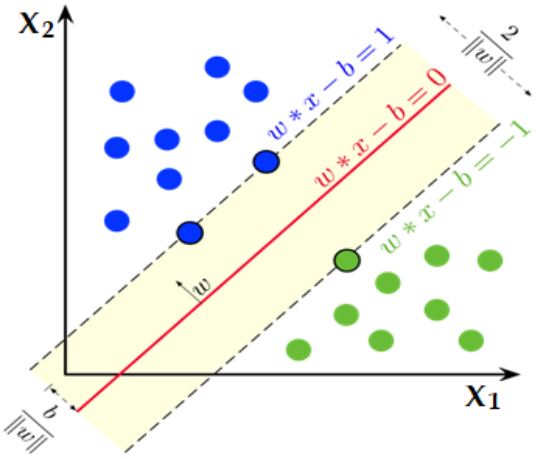


Figura 2.6 Aspectos formales de Máquinas de Soporte Vectorial.

2.8.3 Clasificador Naive Bayes

Todos los clasificadores Naive Bayes asumen que el valor de una característica particular es independiente del valor de cualquier otra característica sin considerar cualquier posible correlación entre dichas variables. Naive Bayes es un modelo de probabilidad condicional: dada una instancia de problema a clasificar, representada por un vector $x = (x_1, \dots, x_n)$ de n características, asigna a esta instancia las probabilidades $p(C_k|x_1, \dots, x_n)$ para cada posible salida K o clases C_k.

El problema con la formulación anterior es que, si el número de características n es grande o si una característica puede tomar un gran número de valores, entonces no es factible basar dicho modelo en tablas de probabilidad. Por lo tanto, reformulamos el modelo para hacerlo más manejable. Usando el teorema de Bayes, la probabilidad condicional se puede descomponer como:

$$p(C_k|\boldsymbol{x}) = \frac{\boldsymbol{p}(C_k)\, p(\boldsymbol{x}|C_k)}{\boldsymbol{p}(\boldsymbol{x})}$$

En la práctica, solo hay interés en el numerador de esa fracción, porque el denominador no depende de C y se dan los valores de las características x_i, por lo que el denominador es efectivamente constante. El numerador es equivalente al modelo de

probabilidad conjunta $p(C_k, x_1, \dots, x_n)$, que se puede reescribir usando la regla de la cadena para aplicaciones repetidas de la definición de probabilidad condicional:

$$p(C_k, x_1, \dots, x_n) = p(x_1, \dots, x_n, C_k)$$

$$= p(x_1 | x_2, \dots, x_n, C_k)\, p(x_2, \dots, x_n, C_k)$$

$$= p(x_1 | x_2, \dots, x_n, C_k)\, p(x_2 | x_3, \dots, x_n, C_k)\, p(x_3, \dots, x_n, C_k)$$

$$= p(x_1 | x_2, \dots, x_n, C_k)\, p(x_2 | x_3, \dots, x_n, C_k) \dots p(x_{n-1} | x_n, C_k)\, p(x_n | C_k)\, p(C_k)$$

Ahora entran en juego los supuestos de independencia condicional: asume que todas las características en $\boldsymbol{X}$ son mutualmente independientes, condicional en la categoría C_k.

$$p(x_i | x_{i+1}, \dots, x_n, C_k) = p(x_i | C_k)$$

Así, el modelo conjunto se puede expresar como:

$$p(C_k | x_1, \dots, x_n) \propto p(C_k, x_1, \dots, x_n)$$

$$= p(C_k)\, p(x_1 | C_k)\, p(x_2 | C_k)\, p(x_3 | C_k) \dots$$

$$= p(C_k) \prod_{i=1}^{n} p(x_i | C_k)$$

Donde $\propto$ denota *proporcionalidad*. Esto significa que, bajo los supuestos de independencia anteriores, la distribución condicional sobre la variable de clase C es

$$p(C_k | x_1, \dots, x_n) = \frac{1}{Z} p(C_k) \prod_{i=1}^{n} p(x_i | C_k)$$

Donde la evidencia $Z = p(\boldsymbol{x}) = \sum_k p(C_k)\, p(\boldsymbol{x} | C_k)$ es un factor de escala que depende solo de $x_1, \dots, x_n$. Es decir, una constante si se conocen los valores de las variables de características.

Para estimar los parámetros para la distribución de características, se debe asumir una distribución o generar modelos no paramétricos para las características del conjunto de entrenamiento (John and Langley 2013). Los supuestos sobre distribuciones de características se denominan modelo de evento del clasificador Naive Bayes. Para las funciones discretas como las que se encuentran en la clasificación de documentos (incluido el filtrado de correo no deseado), la distribución multinomial es muy popular.

Si una clase dada y un valor de característica nunca ocurren juntos en los datos de entrenamiento, entonces la estimación de probabilidad basada en la frecuencia será cero. Esto es problemático porque borrará toda la información en las otras probabilidades cuando se multipliquen. Por lo tanto, a menudo es deseable incorporar una corrección de muestra pequeña, llamada *pseudocuenta*, en todas las estimaciones de probabilidad, de modo que nunca se establezca una probabilidad exactamente igual a cero. Esta forma de regularizar Naive Bayes se llama *suavizado de Laplace* cuando la pseudocuenta es uno y *suavizado de Lidstone* en el caso general.

2.9 Métricas para evaluar algoritmos de aprendizaje automático

La matriz de confusión y sus métricas permiten conocer el desempeño de un modelo de aprendizaje. La matriz de confusión es una matriz cuadrada donde las filas se nombran según las clases reales y las columnas según las clases previstas por el modelo. La matriz muestra de forma explícita cuándo una clase es confundida con otra, lo que permite trabajar de forma separada con distintos tipos de error. La Tabla 2.7 muestra una matriz de confusión.

Tabla 2.7. Matriz de confusión.

	P (modelo)	**N (modelo)**
P (real)	VP	FN
N (real)	FP	VN

VP representa a verdaderos positivos, VN verdadero negativo, FN falso negativo y FP falso positivo. La diagonal principal contiene la suma de todas las predicciones correctas. La otra diagonal refleja los errores del clasificador (Errores Tipo I y Tipo II).

Para evaluar el modelo, podríamos calcular su exactitud (*accuracy*), la cual representa la proporción de predicciones correctas que ha hecho el modelo del total de predicciones. La exactitud se calcula de acuerdo con la Fórmula 2.2.

$$accuracy = \frac{VP+VN}{VP+VN+FP+FN} \tag{2.2}$$

La métrica exactitud resulta conveniente cuando el conjunto de datos posee Verdaderos Positivos y Verdaderos Negativos en cantidades similares. De lo contrario pueden utilizarse las métricas de precisión (*precision)* y la especificidad (*recall)*. La descripción matemática de *precision* y *recall* se muestran en las Fórmulas 2.3 y 2.4.

$$precision = \frac{VP}{VP+FP} \tag{2.3}$$

$$recall = \frac{VP}{VP+FN} \tag{2.4}$$

2.10 Reducción de dimensiones

Es deseable agregar tantas características a un modelo como sea posible esperando mejorar los resultados de una métrica. Pero el rendimiento del modelo disminuirá debido al número elevado de características. Esto ocurre porque la densidad de la muestra disminuye exponencialmente con el aumento de la dimensionalidad. Sin aumentar el número de muestras de entrenamiento, la dimensionalidad aumenta y se vuelve más y más escasa. Debido a esta escasez, es mucho más fácil encontrar una solución “perfecta” para el modelo de aprendizaje automático, lo que probablemente conduce a un sobreajuste (*overfitting*). El sobreajuste ocurre cuando el modelo se corresponde demasiado con un conjunto particular de datos y no generaliza bien. La reducción de la dimensionalidad puede hacerse mediante el uso de *técnicas de selección* y *extracción de características*.

La selección basada en frecuencia es sencilla y consiste en seleccionar un número de características utilizadas con mayor frecuencia. No existe una regla o método para definir el número apropiado de características. Por otro lado, la reducción por extracción es un proceso que toma el conjunto original de N características y las transforma en un subconjunto M donde $M \leq N$. Solo las características transformadas son utilizadas para el proceso de entrenamiento e inferencia del algoritmo de Aprendizaje Automático. Dos de las técnicas más populares en esta categoría son Análisis de Componentes Principales (PCA, *Principal Component Analysis*) y Análisis Semántico Latente (LSA, *Latent Semantic Analysis*).

2.10.1 Análisis de Componentes Principales

La idea central del Análisis de Componentes Principales es reducir la dimensionalidad de un conjunto de datos que consiste en un gran número de variables interrelacionadas, al tiempo que se conserva la mayor cantidad posible de la variación presente en el conjunto de datos. Esto se logra transformando a un conjunto nuevo de variables denominados *componentes principales*, que no están correlacionados y que están ordenados de modo que los primeros conserven la mayor parte de la variación presente en todas las variables originales.

De manera formal, la reducción de dimensiones con Análisis de Componentes Principales puede plantearse de la siguiente manera: supongamos que se tienen n observaciones de p diferentes variables. Defina a X como una matriz de $(n\ x\ p)$, donde la i-esima columna de X contiene las observaciones de la i-esima variable, $i = 1, \dots, p$. Cada renglón x_i de X puede representarse como un punto en un espacio p-dimensional. En consecuencia, X contiene n puntos en un espacio p-dimensional. Esta técnica proyecta datos p-dimensionales en un subespacio q- dimensional $(q \leq p)$ de manera que minimiza la suma de las distancias cuadradas desde los puntos hasta sus proyecciones.

2.10.2 Análisis Semántico Latente

El Análisis Semántico Latente analiza las relaciones entre un término y los conceptos contenidos en una colección de texto no estructurada. La técnica produce un conjunto de *conceptos* más pequeño que el conjunto original. Los objetos son documentos y las características son términos que aparecen en estos. La matriz $X_{p \times n}$ es una matriz de p términos y n documentos, se le conoce como matriz término-documento.

El Análisis Semántico Latente deriva factores de índice no correlacionados que podrían considerarse conceptos artificiales, es decir, la semántica latente. Esta técnica se basa en el Valor Singular de Descomposición (SVD, *Singular Value Decomposition*) de la matriz término-documento de la siguiente manera:

$$X_{p \times n} = T_{p \times m} S_{m \times m} V^{T}_{(n \times m)}$$

donde:

- $T_{p \times m}$ es la matriz de vectores propios de XX^T; m es el rango de XX^T
- $S_{m \times m}$ es una matriz diagonal que contiene la raíz cuadrada de los valores propios de XX^T
- $V_{n \times m}$ es la matriz de vectores propios de $X^T X$
- T es la traspuesta de la matriz

En esta representación, las entradas diagonales en S son los valores singulares y normalmente se ordenan primero con el valor singular más grande (valor propio más grande). La reducción se logra eliminando todos menos los k valores singulares, lo que nos da una nueva descomposición:

$$\bar{X}_{p \times n} = T'_{p \times k} S'_{k \times k} V'^{T}_{k \times n}$$

Donde S' es ahora $k \times k$ y corresponde a las columnas que han sido eliminadas. En esta situación, $V'S'$ es una matriz $(n \times k)$ que nos da las coordenadas de n

documentos en el nuevo espacio k-dimensional. La transformación que se lleva a cabo es lineal.

Capítulo 3. Estado del arte

Mosteller y Wallace (1966) fueron de los primeros en usar y popularizar las palabras funcionales para la atribución de autoría. Según su razonamiento, estas palabras aparecían en todos los textos (independientemente del tema y el género) y no estaban sujetos a un control consciente. Otra forma de caracterizar el estilo es por medio de palabras frecuentes (unigramas de palabras), este enfoque no retiene información contextual. Esta representación puede ser eficiente, pero podría mejorarse utilizando información sintáctica. La información sintáctica puede incorporarse con etiquetas POS, o utilizando la derivación de palabras. Ambas ofrecen abstracciones lejos de la palabra completa, las etiquetas de POS se centran en los atributos sintácticos de los textos, y los derivados de palabras pueden capturar un poco más de la semántica de las palabras.

Otro aspecto que destacar es si las características poseen o no información lingüística. Los 1-gramas de caracteres no poseen información lingüística, pero han demostrado ser muy útiles especialmente en tareas independientes del lenguaje. A diferencia de las características léxicas, los n-gramas de caracteres también son más tolerantes al ruido, como los errores gramaticales o la puntuación inusual (Stamatatos 2009). No obstante, al aumentar la longitud del n-grama, se aproxima a la longitud promedio de la palabra. De esta manera, este tipo de característica posiblemente se vuelvan más significativas y motivadas lingüísticamente.

Las características sintácticas son incluso más subconscientes que las palabras funcionales, ya que no están lexicalizadas y, por lo tanto, representan atributos altamente automáticos e inconscientes de producción lingüística (Chaski, 2006). Los trabajos del estado del arte utilizaron características basadas en información sintáctica tales como la complejidad sintáctica, longitud de sentencias, uso de voz activa y pasiva, longitud de sentencias en palabras, oraciones simples y compuestas, verbos subordinados, entre otras.

A continuación, se presentan los trabajos relacionados a la detección de cambio de estilo de escritura.

3.1 Cambios de estilo debido a cuestiones patológicas

William *et al*. (2003) analizaron 57 cartas escritas por el rey *James VI* (monarca del siglo XVII) que comprendieron un periodo de veinte años. El objetivo fue evaluar si las señales lingüísticas de estas cartas reflejaban el envejecimiento normal, Alzheimer o Demencia Vascular. Las características estilométricas utilizadas fueron la relación riqueza de vocabulario, la longitud de sentencias en palabras, el número de cláusulas por sentencia y la complejidad sintáctica *D-Level*. Los resultados revelaron un patrón de disminución de la complejidad sintáctica y una mayor diversidad de vocabulario a partir de los primeros años en los 50´s del rey. Los resultados les sugirieron que el rey confiaba en las funciones semánticas para compensar la disminución de la sintaxis. Con el respaldo de los registros médicos y los resultados de las autopsias, los investigadores sugirieron que el rey podría haber sufrido de hipertensión crónica, una condición que puede ser un antecedente de demencia vascular (Posner et al., 2002). El estudio no produjo un diagnóstico concluyente, debido a la aplicabilidad desconocida del análisis lingüístico moderno al estilo de escritura isabelino, así como a la diferencia en la salud y la duración de la vida en el siglo XVII, advirtieron los investigadores.

Garrard *et al*. (2005) examinaron muestras de texto de novelas escritas por *Iris Murdoch*, escritora inglesa a quien se le diagnosticó la enfermedad de Alzheimer. Para el análisis utilizaron los textos completos de las novelas *The Net* (1954), *The Sea* (1978) y *Jackson's Dilemma* (1995). Esta última, escrita cuando los primeros síntomas de Alzheimer comenzaron a emerger. El objetivo de esta investigación fue determinar si las propiedades del vocabulario utilizado en la novela *Jackson´s Dilemma* difiere significativamente de las dos novelas anteriores. Como características estilométricas se utilizaron listas de palabras, etiquetas POS y el contexto donde estas ocurren, la complejidad sintáctica (longitud de las oraciones) y la densidad de las cláusulas subordinadas. *Garrard et al* indicaron que los hallazgos más convincentes fueron a nivel

de vocabulario. La evidencia de la disponibilidad de un vocabulario más restringido durante la redacción del trabajo final fue proporcionada por el menor número de tipos de palabras únicas en relación con el recuento general de palabras, lo que implicó una mayor tasa de repetición de las palabras de la última novela y una mayor tasa de introducción de nuevas palabras en los dos trabajos anteriores.

Lancashire y Hirst (2009) analizaron la evolución en el uso de características léxicas en 16 novelas de *Agatha Christie*, escritora que se presume padecía de Alzheimer cuando compuso sus últimas obras. Las novelas fueron representadas por medio de riqueza de vocabulario, colocaciones y palabras indefinidas (*thing, anything, something*). De cada novela utilizaron 5 bloques de 10,000 palabras cada uno. Los investigadores descubrieron una disminución gradual en el tamaño del vocabulario a través del tiempo, así como un aumento en las frases repetidas y las palabras indefinidas, que fueron particularmente evidentes en las últimas novelas. La riqueza del vocabulario también disminuyó con la edad de la autora, las tres novelas que escribió en sus 80, *Nemesis*, *Elephants* y *Postern*, tuvieron un vocabulario más pequeño que cualquiera de las obras analizadas escritas entre las edades de 28 a 63 años. El número de diferentes tipos de frases que se repiten aumentó con la edad. Según los investigadores, esto implicó una disminución en la riqueza léxica de su escritura. El uso de palabras vagas e indefinidas también aumentó significativamente con la edad.

Le (2010) realizó experimentos para verificar la hipótesis de que la mayoría de los patrones de cambios lingüísticos detectados en el envejecimiento normal también están presentes en personas con demencia. Le analizó las obras de los escritores británicos *Iris Murdoch*, *Agatha Christie* y *P.D. James*. Las características estilométricas utilizadas fueron riqueza de vocabulario, etiquetas POS, palabras de contenido, palabras frecuentes, especificidad de palabra, palabras de relleno (*well*, *yeah*, *ah*, *um*), complejidad sintáctica escala D-level (Rosenberg y Abbeduto, 1987) y uso de voz pasiva.

Le consideró una variedad de medidas léxicas y sintácticas basadas en investigaciones anteriores, que sugieren que el vocabulario y la complejidad sintáctica disminuyen más rápidamente en la demencia, especialmente el uso de palabras de baja

frecuencia y más específicas, así como las repeticiones y desfluencias léxicas. Además, se supuso que la voz pasiva es un indicador de un declive lingüístico más rápido, ya que el grupo no sano utilizó menos construcciones pasivas, así como pasivos más simples sin agentes. Las hipótesis con respecto a un declive léxico más rápido en *Murdoch* están ampliamente confirmadas. Más de 20 años antes de que se manifestaran los síntomas de la enfermedad de Alzheimer, su vocabulario comenzó a disminuir, lo que resultó en un aumento significativo de las repeticiones léxicas de las palabras de contenido. Sin embargo, su especificidad léxica, medida a través de la proporción de sustantivos y verbos indefinidos específicos, permaneció intacta en todo momento.

Todos los tipos léxicos de *Christie* muestran una disminución general con solo dos excepciones. Las puntuaciones de vocabulario, repetición y especificidad varían solo en una escala muy pequeña en las novelas de *James*. Por lo tanto, los autores señalaron que, aunque *Murdoch* no comparte el aumento de *Christie* en los nombres indefinidos, ambos muestran un declive léxico común que no se encuentra en *James* que valida su hipótesis con respecto a los marcadores léxicos. De acuerdo con Le, los resultados del análisis sintáctico de complejidad son un tanto desconcertantes: si bien no se pueden encontrar tendencias lineales significativas para *Murdoch* durante todo el período, se produce una caída en sus 40 y 50 años, seguida de un período de recuperación menos intuitivo y luego un ligero descenso para sus dos últimas novelas. Los resultados de *Christie* varían en gran medida en general, lo que indica una tendencia creciente en lugar de disminuir. Tanto Murdoch como *Christie* muestran una disminución general en las construcciones pasivas, pero un aumento proporcional en las construcciones más simples y una caída de las más difíciles, aunque no todos los resultados son significativos. Sin embargo, los agentes pasivos de *Murdoch* aumentan significativamente. Los resultados sintácticos de *James* varían ligeramente.

Hirst y Feng (2012) analizaron los cambios en el estilo de escritura de tres novelistas británicos: *Iris Murdoch*, *Agatha Christie* y *P.D. James*. El estudio pretendía verificar la hipótesis de que estilo de escritura de autores con la enfermedad de Alzheimer sufre cambios con el tiempo. Las características estilométricas utilizadas fueron léxicas

(palabras frecuentes, hapaxes), de caracteres (dígitos, signos de puntuación, mayúsculas y minúsculas, *2*-gramas y *3*-gramas) y sintácticas (etiquetas POS y entropía de etiquetas POS). Hirst y Feng utilizaron textos de aproximadamente 4,000 palabras, además dividieron las novelas en etapas denominadas *prime*, *transition* y *late*. Para *Murdoch* y *Christie*, el período *late* contenía novelas en las cuales los efectos de la enfermedad de Alzheimer eran evidentes y el período *prime* contiene novelas en las que se presume que no hay efectos del Alzheimer. Construyeron un clasificador binario para discriminar novelas del período *prime* y *late*. Hirst y Feng indicaron que ignoraron las novelas del periodo *transition* para separar el trabajo de los autores en dos casos claros. La clasificación binaria pudo discriminar entre los períodos *prime* y *late* de *Christie* y *James*, pero no en el caso de *Murdoch*. Hirst y Feng concluyeron que los estilos de estos autores cambiaron en la vejez, pero no debió suceder así para *P.D. James*, quien no padecía la enfermedad. Los autores no pudieron concluir si los cambios se debieron directamente a la enfermedad de Alzheimer. Sin embargo, indicaron que los resultados apoyan la idea de que estos conjuntos de características podrían permanecer prácticamente sin cambios incluso en la enfermedad de Alzheimer.

3.2 Cambios de estilo debido al envejecimiento

Pennebaker y Stone (2003) realizaron un análisis estilométrico para verificar que conforme la edad aumenta, también se incrementa la complejidad cognitiva. En particular, plantearon hipótesis sobre el efecto de envejecimiento del lenguaje. Los dos conjuntos de datos considerados para esto fueron un corpus que contenía autoinformes de estudios de divulgación psicológica y un corpus con trabajos recopilados de 10 autores a lo largo del tiempo (entre los años 1591–1939). El estudio de Divulgación incluyó a 3,280 participantes de 45 estudios separados, de los cuales 32 fueron experimentos tradicionales de divulgación emocional fueron asignados al azar para escribir sobre un tema emocional o traumático o un tema superficial. El análisis estadístico incluyó análisis de correlación y regresión tanto lineal como cuadrática simple.

El segundo conjunto contenía muestras de texto de 10 autores diferentes, tanto británicos como estadounidenses, hombres y mujeres, de diferentes géneros. La datación se basó en el momento en que se escribió una obra, con un promedio de años tres cuando era necesario y se volvía a la fecha de publicación en los casos en que la fecha de composición no pudo ser determinada. Parte del análisis fue correlacional, analizando la relación simple entre el uso del lenguaje y la edad.

Las características LIWC (*Linguistic Inquiry and Word Count*) de cada autor fueron correlacionadas con la edad del autor en el año en que se escribió el trabajo. A continuación, calcularon las medias de las correlaciones para cada una de las variables y se sometieron a pruebas *t-tes*t de una sola muestra para evaluar si cada media era significativamente diferente de cero. Cinco de las 14 medias de correlación originales fueron significativamente diferente de cero $(p \leq .05, df = 9)$.

La otra parte del análisis consistió en crear un coeficiente de envejecimiento basado en los hallazgos de las 14 variables en el estudio de divulgación. Esto reveló que 6 de los 10 autores examinados exhibieron el mismo patrón de uso del lenguaje con respecto al envejecimiento que se encontró en el proyecto de Divulgación. El hipotético aumento de palabras de emoción positiva y la disminución de palabras de emoción negativa fueron significativas para el proyecto de Divulgación, pero no para el proyecto Autor. Por otro lado, se encontró una caída significativa en el tiempo en la primera persona en ambos Corpus (y en los datos de Divulgación hubo una disminución no significativa en los pronombres en plural en primera persona). En lugar del hipotético cambio de tiempo futuro a pasado, se encontró una disminución en el tiempo pasado y un aumento en los verbos en tiempo futuro en el proyecto Divulgación, un aumento en el tiempo futuro también fue significativo para el proyecto Autor. Aunque no se predijo ningún cambio en las secuencias de cartas largas, estas aumentaron significativamente con el tiempo para el proyecto Divulgación, con este efecto presente pero no significativo para el proyecto Autor.

3.3 Cambios de estilo debido otros factores

Tabata (1994) consideró la variación cronológica en las obras de *Charles Dickens* considerando las separaciones por estilo narrativo, Tabata examinó las narraciones en tercera persona del período inicial de *Dickens* (todas escritas en la década de 1830) y el estilo narrativo en primera persona y en tercera persona escrito después de 1849. Las palabras específicas de estilo narrativo fueron excluidas para enfocarse en diferencias más sutiles de diferencias estilísticas. Se utilizó el algoritmo de Análisis de Componentes Principales para agrupar patrones de variación entre las 74 palabras más frecuentes, que luego se proyectaron en las muestras de texto. Esto revela una clara separación entre los estilos narrativos en una dimensión y su cronología en la otra.

Por ejemplo, los pronombres relativos *wich* y *who*, como los que discriminan más fuertemente a favor de los textos escritos en la década de 1830, mientras que *that* predominan después de 1849. Tabata reportó que el estilo tardío de *Dickens* también presenta las partículas adverbiales out y *down*, el pronombre i*t*, y la preposición *like*, donde algunos de estos cambios también han sido notados por otros estudios con respecto a la narrativa de primera persona en inglés general. Para identificar los marcadores más discriminativos, Tabata usó una prueba *t-test* en cada marcador por separado para las muestras narrativas en tercera persona. Los marcadores *wich*, *it*, *out* y *like* fueron muy significativos ($p < 0.01$). El agrupamiento (*clustering*) usando solo las 21 palabras más discriminativas resulta en una distinción aún más nítida entre conjuntos.

Forsyth (1999) analizó cambios en el estilo de escritura del poeta *William Butler Yeats*. El análisis se basa en subcadenas de marcadores distintivos que se extraen de 142 poemas utilizando una versión modificada de "Búsqueda de características de Monte-Carlo". Estas subcadenas se clasifican de acuerdo con las categorías '*Young Yeats*' y '*Old Yeats*'. Los poemas se dividieron en estas categorías basándose en que se escribieron antes y después del año 1915.

Forsyth informó sobre la identificación de 20 marcadores claros de '*Young Yeats*' y '*Old Yeats*'. Para nueve de cada diez poemas de prueba su conteo es mayor en la

categoría de edad apropiada. Con la finalidad de fechar los textos, el investigador definió un índice '*Young Yeats*' como $YYIX = (YY - YO)/(YY + YO)$, donde YY se refiere al número de marcadores en la etapa joven y YO a los marcadores de la etapa vieja. Una correlación de $YYIX$ y año de composición produce un r de -0.84. Al examinar dos poemas que habían sido revisados por Yeats unos 30 años más tarde, es notable que el número de marcadores YY disminuyó en la versión revisada, mientras que el número de marcadores OY aumentó.

Can y Patton (2004) también dividieron en las categorías joven y viejo los autores turcos *Cetin Altan* y *Yasar Kemar*. Las obras de *Altan* fueron de los años 1960–1969 (joven) y 2000 (antiguo) y para *Kemal* se seleccionó una novela de 1971 (joven) y 1998 (antiguo). Para cada autor, los datos se dividieron en dieciséis bloques de tamaño fijo de 2,500 palabras para cada período. Informaron que el tipo promedio y la longitud del token aumentaron significativamente para los dos autores entre sus trabajos antiguos y nuevos utilizando una prueba t-test. Además, empleando diferentes métodos, como la Regresión Lineal, Análisis de Componentes Principales, y Análisis de Varianza, encontraron que los tipos de palabras son discriminadores ligeramente mejores que el tipo y la longitud del token.

Los autores también reportaron una fuerte relación entre la longitud promedio del token y la antigüedad del texto en las obras de *Altan*, aunque un valor R^2 de 0.24 indica que es probable que haya otros factores involucrados. El análisis de las tasas de uso de diferentes tipos y longitudes de token utilizando Regresión Logística mostró que la longitud de palabra de tres a ocho es predominante en las obras antiguas de *Altan*, mientras que una longitud de palabra de nueve o más fue más representativo de sus nuevas obras. Para *Kemal*, surge una distribución similar, lo que llevó a la conclusión de que estos resultados pueden deberse al mayor dominio del idioma de ambos autores. Can y Patton también comparan trabajos antiguos y nuevos con respecto a los rasgos característicos de las palabras, dando cinco marcadores significativos para *Altan* y dos para *Kemal*. Finalmente, utilizando un Análisis Discriminante, encontraron el mejor marcador, logrando un índice de clasificación promedio de 98.96% para *Altan* y 84.38%

para *Kemal*, la diferencia la atribuyeron a la mayor distancia de tiempo entre el trabajo de *Altan* y, en consecuencia, el desarrollo más pronunciado en estilo.

Los mismos investigadores analizaron 4 novelas del autor turco *Yasar Kemal*. Las características estilométricas que evaluaron fueron palabras frecuentes, longitud de sentencias en palabras, conteo de silabas, etiquetas POS, longitud de palabras y riqueza de vocabulario. Crearon bloques de 5000 palabras para el análisis. Mediante pruebas de significancia encontraron que estas características cambian significativamente a través de los cuatro volúmenes.

Pol (2005) realizó un estudio para medir la fidelidad estilística del autor, para ello creó un corpus de 20 artículos de opinión de 6 autores de habla hispana. Los artículos comprendieron el periodo de un año. Las características estilométricas que utilizó fueron palabras funcionales y de contenido, lemas, riqueza de vocabulario, etiquetas POS, longitud de sentencia, longitud de párrafo, longitud de palabras, hapax legomena, signos de puntuación y adverbios. Pol concluyó que los autores tienden a seguir patrones similares para las diferentes variables, pudo confirmar que los autores tienden a tomar las mismas decisiones a partir de la variedad de opciones que el sistema lingüístico les brinda. Esta conclusión se apoya en el hecho de que los artículos evaluados comprendieron únicamente un año de producción de los autores.

Spassova (2009) realizó experimentos para demostrar la hipótesis que, considerando el tiempo de producción, el comportamiento lingüístico de los individuos no cambia de forma significativa a través del tiempo. La investigadora utilizó como características estilométricas *n*-gramas de etiquetas POS. Organizó a los autores en dos grupos: el grupo 1 con tiempos de producción entre novelas de 20 a 30 años y el grupo 2 con tiempos de producción entre 5 y 8 años. Del grupo 1 concluyó que, del total de variables analizadas, menos de la mitad se muestran susceptibles a la variación. Spassova afirmó que en *2*-gramas, la variación estadísticamente significativa en el cambio de la frecuencia se manifiesta tras períodos de escritura de entre 20 y 30 años. En cuanto a los 3-gramas, afirmó que la variación estadísticamente significativa se dio en novelas que han sido escritas con un mínimo de 10 años de diferencia. Sobre el grupo 2,

concluyó que el estilo de los autores no cambia a corto plazo (cinco años) pero el cambio es patente cuando ha transcurrido un período más largo. Spassova confirmó su hipótesis de que los individuos no tienden a variar en gran medida cuando se trata de un intervalo de tiempo corto.

Klaussner (2017) propuso el desarrollo de métodos para el análisis del estilo de autor a lo largo del tiempo, en particular con respecto a cómo el envejecimiento afecta al lenguaje a lo largo de la vida, qué tipo de características lingüísticas son particularmente cambiantes y cómo interactúan entre sí las diferentes características. Los datos analizados para esta investigación fueron divididos en dos conjuntos principales: veintidós autores literarios que van desde 1847 como año de la primera publicación hasta 1923 como año de la última publicación y el corpus de referencia que proporciona un lenguaje que abarca de 1830 a 1929. Las características estilométricas (marcadores de estilo) utilizadas fueron etiquetas POS, n-gramas (de caracteres y de etiquetas POS), pronombres personales y los tiempos verbales de las palabras. La herramienta estadística utilizada fue Modelos de Regresión. Las conclusiones a las que llegó Klaussner fueron: Las características estilométricas examinadas no proporcionaron evidencia del envejecimiento lingüístico en los autores literarios. Sin embargo, hubo evidencia de interferencia en el lenguaje de fondo. Klaussner indicó que cuando los autores literarios se analizan como un grupo, las características más generales, como *3*-gramas y *4*-gramas y probablemente los ngramas sintácticos, sean las características más reveladoras del cambio literario. Las n-gramas sintácticos parecen variar más fuertemente dependiendo del autor y pueden ser más útiles para análisis individuales. Finalmente, sobre la cuestión de si el cambio de los escritores *James* y *Twain* es marcadamente diferente con respecto a otros autores que componen obras al mismo tiempo, indicó que los análisis no pudieron detectar cualquier cosa que los distinga claramente de sus contemporáneos.

3.4 Análisis del estado del arte

En esta tesis se da por hecho que los autores aquí evaluados estaban clínicamente sanos y que, si ocurren cambios de estilo, estos se deben a la experiencia adquirida conforme avanzaron en sus carreras literarias. Los textos se caracterizaron por medio de n-gramas de distinta naturaleza: caracteres, palabras, etiquetas POS y relaciones sintácticas. En cuanto a estos últimos, la literatura relacionada al análisis de estilo menciona que este tipo de características son difíciles de manipular de forma consciente, por lo que resultan candidatas idóneas por su persistencia a los cambios de tema y probablemente al tiempo. La tarea de detección de cambio de estilo se planteó como un problema de clasificación multiclase y las características se utilizaron bajo en un enfoque de aprendizaje automático supervisado.

Capítulo 4. Método propuesto

4.1 Descripción general del método propuesto

El método propuesto para la detección de cambio de estilo consta de dos etapas principales: la primera consiste en obtener las características estilométricas y la segunda en aplicar un enfoque de aprendizaje automático supervisado para clasificar documentos nuevos mediante las características seleccionadas en la etapa anterior. A la vez, cada etapa involucra un conjunto de fases. La Figura 4.1 muestra el diagrama de bloques de la propuesta.

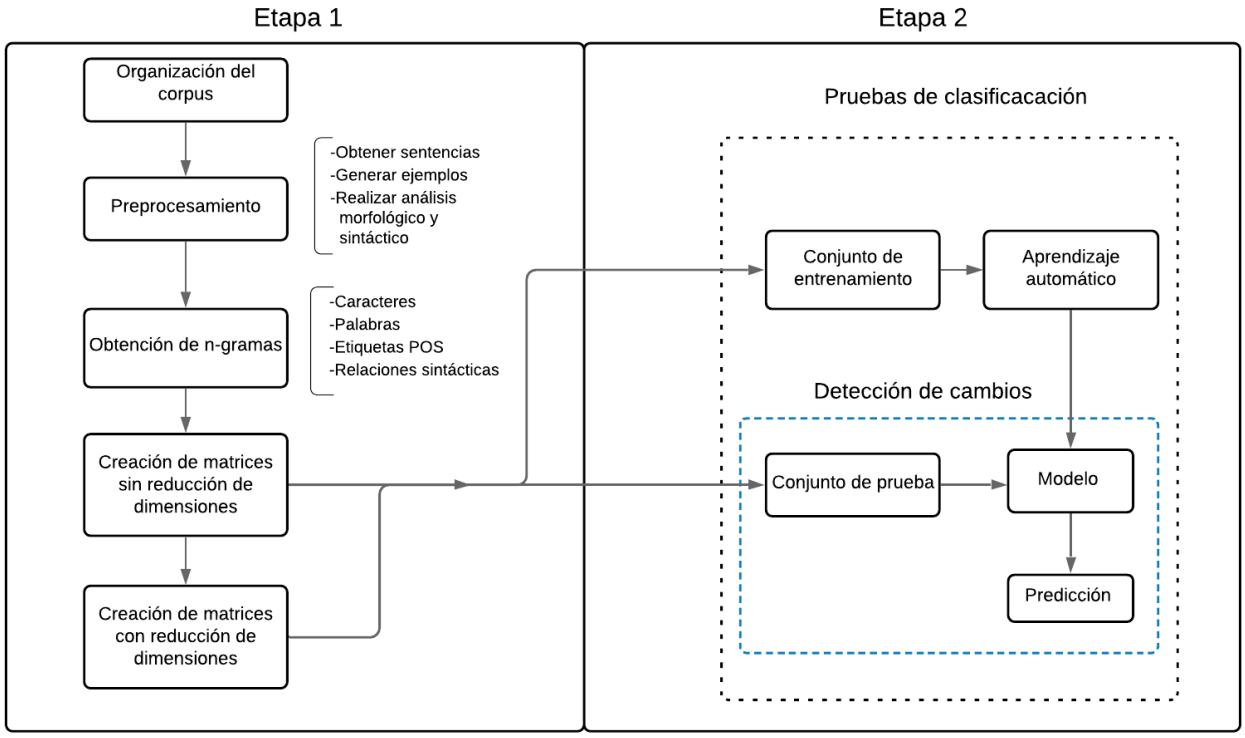


Figura 4.1 Metodología para la detección de cambio de estilo de escritura.

4.2 Descripción por fases

A continuación, se proporciona una descripción detallada de cada actividad planteada en el método propuesto.

4.2.1 Organización del corpus

Se creó un corpus de novelas de autores de habla inglesa con largas carreras literarias, las cuales comprenden al menos 20 años. La organización del corpus se muestra en la Tabla 4.1. Todos los autores cuentan con nueve novelas, todas descargadas del Proyecto Gutenberg[3]. Las novelas se organizaron de forma cronológica de acuerdo con la fecha de publicación.

Tabla 4.1. Corpus de novelas para evaluar el cambio de estilo.

Autor	Etapas					
	Inicial		Media		Final	
	Año	Nombre	Año	Nombre	Año	Nombre
Booth Tarkington (BT)	1899	*Gentleman*	1914	*Penrod*	1919	*Ramsey*
	1902	*Vanrevels*	1915	*Turmoil*	1921	*Alice Adams*
	1905	*Canaan*	1916	*Seventeen*	1922	*Gentle Julia*
Charles Dickens (CD)	1838	*Nicholas Nickleby*	1848	*Dombey and Son*	1859	*Two Cities*
	1838	*Oliver Twist*	1850	*Copperdfield*	1861	*Expectations*
	1841	*Barnaby*	1853	*Bleak house*	1865	*Our mutual friend*
Frederick Marryat (FM)	1830	*The King's Own*	1839	*The panthom ship*	1845	*The Mission*
	1831	*Jacob Faithful*	1839	*A diary in America*	1847	*New Forrest*
	1831	*Newton Forster*	1840	*Olla Podrida*	1848	*The Little Savage*
George Macdonald (GM)	1863	*David Elginbrod*	1873	*Gutta Percha*	1888	*Electrical Lady*
	1864	*Adela*	1875	*A double story*	1891	*Flight of Shadow*
	1865	*Alec Forbes*	1876	*Thomas Wingfold*	1892	*hope of góspel*
George Vaizey (GV)	1901	*School Story*	1908	*Flaming June*	1914	*Cassandra*
	1902	*Pixie*	1908	*Big Game*	1914	*College Girl*
	1902	*Houseful of Girls*	1910	*Marriage*	1915	*Claire*
Louis Tracy (LT)	1903	*wings of morning*	1907	*The captain*	1912	*Romance of NY*
	1904	*the revelers*	1909	*inmortals*	1916	*The day of wrath*
	1905	*disapperance*	1909	*the stoneway girl*	1919	*Mortimer fenley*
Mark Twain (MT)	1869	*Innocents Abroad*	1883	*Mississippi*	1897	*The Equator*
	1872	*Roughing It*	1884	*Huckleberry Finn*	1905	*What is man?*
	1876	*Tom Sawyer*	1889	*King Arthur*	1906	*Dollar*

[3] https://www.gutenberg.org/

Las obras de cada autor se dividieron en etapas considerando que, si existen variaciones de su estilo de escritura, estas puedan detectarse comparando los patrones detectados en cada una de las etapas. Se establecieron 3 etapas denominadas *inicial*, *media* y *final*. De forma general, se trató de dividir de forma proporcional los años transcurridos entre la primera y la última novela de la colección de cada autor.

4.2.2 Preprocesamiento

Cada novela fue dividida en oraciones por medio de la herramienta NLTK[4] (*Natural Language ToolKit*). Específicamente con el método sent_tokenize(), el cual usa un algoritmo no supervisado denominado *Unsupervised Multilingual Sentence Boundary*, propuesto por Strunk y Kiss (2006). Este algoritmo construye un modelo para palabras abreviadas, colocaciones y palabras que comienzan oraciones; luego usa ese modelo para encontrar límites de oraciones. Las oraciones de 1 y 2 palabras fueron eliminadas, principalmente porque los n-gramas sintácticos donde $n < 3$ no capturan las relaciones no lineales del árbol de dependencia. En su lugar capturan las relaciones lineales como lo hacen los *n*-gramas tradicionales. Otra razón para eliminar estas sentencias se debe a que los *3*-gramas de palabras y etiquetas POS requieren de oraciones de al menos tres palabras. Las sentencias resultantes se convirtieron a caracteres en minúsculas para hacer que cadenas de caracteres superficialmente diferentes tengan la misma forma (Turney y Pantel, 2010).

Posteriormente, las novelas fueron divididas en textos más pequeños de tamaño proporcional. De esta forma, una novela puede dividirse en tantos fragmentos como se requieran y con la misma cantidad de sentencias. Las pruebas de clasificación preliminares mostraron que los textos que contienen de 50 a 500 sentencias producían tasas de acierto bajas. Se optó por dividir las novelas en cuatro diferentes tamaños:

[4] http://www.nltk.org/

novelas completas, mitades de novelas, tercios de novelas y cuartos de novelas. A manera de ejemplo, si una novela consta de 1,000 oraciones, entonces cada texto contendrá 1000, 500, 333 y 250 sentencias respectivamente. La Tabla 4.2 muestra los textos generados con las novelas del autor *Booth Tarkington*. Las novelas completas, medias novelas, tercios de novelas y cuartos de novelas se identifican con las leyendas 1, 2, 3 y 4, respectivamente.

Tabla 4.2. Número de sentencias en las novelas de *Booth Tarkington*.

Novelas	Tamaño del texto			
	1	**2**	**3**	**4**
AliceAdams	5,602	2,801	1,867	1,400
Canaan	4,598	2,299	1,532	1,149
Gentleman	5,350	2,675	1,783	1,337
Julia	4,312	2,156	1,437	1,078
Penrod	3,841	1,740	1,160	870
Ramsey	2,179	1,089	726	544
Seventeen	3,917	1,958	1,305	979
Turmoil	5,892	2,946	1,964	1,473
Vanrevels	2,802	1,401	934	700

Al terminar la división de los textos, se realizó el proceso de etiquetado y el análisis sintáctico. Para obtener las etiquetas POS se utilizó el etiquetador de la herramienta NLTK. El análisis sintáctico se llevó a cabo con el analizador sintáctico *Stanford Parser*. Comelles *et al* (2010) realizaron un estudio para evaluar cinco analizadores, reportaron que Stanford Parser fue el que cometió menos errores en oraciones que presentaban ambigüedad sintáctica.

Considere como ejemplo la oración "*Victor sat at the counter on a plush red stool*", cuyo árbol de dependencias se muestra en la Figura 4.2.

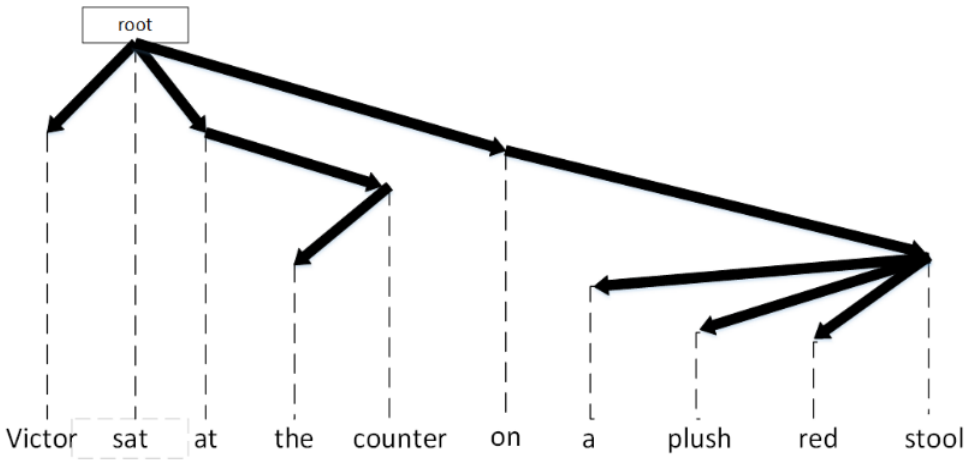


Figura 4.2 Árbol de dependencias de la oración "*Victor sat at the counter on a plush red stool*".

En la Tabla 4.3 se muestran las frecuencias de *3*-gramas tradicionales de palabras y *3*-gramas sintácticos de palabras de la oración "*Victor sat at the counter on a plush red stool*".

Tabla 4.3. *3*-gramas tradicionales de palabras y *3*-gramas sintácticos de palabras.

***3*-gramas tradicionales**	**Frecuencia**	***3*-gramas sintácticos**	**Frecuencia**
Victor-sat-at	1	sat[victor,at]	1
sat-at-the	1	sat[victor,on]	1
at-the-counter	1	sat[at,on]	1
the-counter-on	1	sat[on[stool]]	1
counter-on-a	1	sat[at[counter]]	1
on-a-plush	1	on[stool[plush]]	1
a-plush-red	1	on[stool[a]]	1
plush-red-stool	1	on[stool[red]]	1
		stool[a,plush]	1
		stool[a,red]	1
		stool[plush,red]	1
		sat[at[counter]]	1
		at[counter[the]]	1

Los *n*-gramas sintácticos capturan las relaciones sintácticas entre las palabras, que corresponden a las reglas gramaticales del lenguaje. Por ejemplo, al comparar el *n*-grama *Victor-sat-at* y el *n*-grama sintáctico *sat[Victor,at]* se observa que ambos *n*-gramas se conforman por las mismas palabras, pero en el segundo se establece una relación entre las palabras *Victor* y *at*, así como una relación entre éstas y la palabra *sat*.

4.2.3 Generación de *n*-gramas

En cada estudio con *n*-gramas es común modificar el valor de *n* para permitir que los *n*-gramas contengan más elementos. Se obtuvieron n-gramas de longitud $n = \{1, 2, 3, 4, 5\}$. Los que mejores resultados se obtuvieron con $n = 3$, valor similar a lo reportado en detección de plagio (Barrón-Cedeño y Rosso, 2009), atribución de autoría (Escalante y otros, 2011; Sidorov y otros, 2014; Sapkota y otros, 2014), categorización de textos (Addellatif y Zakaria, 2007) e identificación de autores (Houvardas y Stamatatos, 2006).

Al trabajar con n-gramas también es importante establecer la frecuencia mínima de uso dentro del texto. Inicialmente, se consideraron todos los *3*-gramas (a partir de la frecuencia 1). Con esta acción se obtuvieron una enorme cantidad de caracteristicas de frecuencias 1 y 2, las cuales representaban más de la mitad del total de *3*-gramas generados. La literatura relacionada a estilometría sugiere que los n-gramas de baja frecuencia contribuyen poco o nada al estilo de escritura del autor. Por ello, se seleccionaron los *3*-gramas con frecuencia ≥ 3. Este umbral permitió acotar el número de *3*-gramas. A pesar de establecer este corte, se observó que sólo una pequeña parte de los *3*-gramas son utilizados con mucha frecuencia (Ver Anexo C). Existe una relación entre la longitud del *n*-grama y la frecuencia de uso: al aumentar la longitud, el número de *n*-gramas se incrementa de forma significativa y las frecuencias tienden a disminuir.

Los *n*-gramas de caracteres, palabras y etiquetas POS se generaron con el programa *text2ngram*[5], un software libre bajo licencia GPL. Dicho programa requiere como parámetros el tipo de n-grama a generar y su longitud, además de la frecuencia de uso mínima que debe tener en el archivo de entrada. Los *3*-gramas de relaciones

[5] https://homepages.inf.ed.ac.uk/lzhang10/ngram.html

sintácticas se obtuvieron con un algoritmo desarrollado en Python[6] (Sidorov y otros, 2014; Posadas-Durán y otros, 2015; Markov y otros, 2017), adicionalmente el algoritmo también obtiene n-gramas sintácticos de palabras y etiquetas POS. El algoritmo construye una representación vectorial de cada texto. Cada entrada se compone de dos partes: el n-grama sintáctico y su frecuencia de uso. De acuerdo con Durán (2017), el *n*-grama sintáctico que mejor resultado presentó fue el de relaciones sintácticas.

La Tabla 4.4 muestra el número de *3*-gramas obtenidos en la colección de novelas de cada autor. Los *n*-gramas de caracteres, palabras, etiquetas POS y relaciones sintácticas se identifican con las leyendas *car*, *pal*, *pos* y *sr* respectivamente. Las siglas de la columna autor representan los nombres de estos (Véase Tabla 4.1).

Tabla 4.4 Total de características obtenidas por tamaño de texto.

Autor	car				pal				pos				sr			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
BT	7,584	6,820	6,388	6,114	3,467	2,157	1,635	1,411	5,003	4,186	3,747	3,457	4,574	3,538	3,049	2,800
CD	8,129	7,645	7,319	7,132	32,960	21,870	17,272	14,601	8,647	7,615	7,028	6,594	8,652	6,839	5,911	5,391
FM	8,645	7,879	7,528	7,233	9,523	6,220	4,715	3,959	6,247	5,368	4,854	4,536	5,291	4,196	3,646	3,291
GM	8,254	7,527	7,079	6,798	7,791	4,799	3,654	2,935	6,192	5,266	4,792	4,456	6,118	4,809	4,215	3,766
GV	7,397	6,634	6,212	5,909	3,381	1,954	1,334	1,092	4,878	4,060	3,634	3,317	4,455	3,504	3,067	2,737
LT	7,530	6,739	6,297	5,986	3,281	1,850	1,310	977	4,680	3,881	3,486	3,173	5,032	3,962	3,385	3,059
MT	9,187	8,421	7,961	7,704	7,933	5,294	4,122	3,630	6,252	5,408	4,891	4,581	5,345	4,264	3,745	3,391

4.2.4 Creación de modelos sin reducción de dimensiones

Como se puede apreciar en la tabla previamente mostrada, el número de dimensiones es alto, esto podría afectar la exactitud de la clasificación. La cantidad de textos (instancias) en la matriz esta determinada por el tamaño de los textos disponibles. Cada matriz se divide en conjuntos de entrenamiento y prueba. La Tabla 4.5 muestra el

[6] http://www.cic.ipn.mx/~sidorov/MultiSNgrams_3_3.py

número de instancias disponibles en los distintos tamaños de texto, así como la proporción de ellas utilizadas para entrenamiento y prueba.

Tabla 4.5 Distribución de instancias de acuerdo al tamaño de texto.

Tamaño Del texto	Instancias	Entrenamiento	Prueba
Novelas completas	9	6	3
Medias novelas	18	12	6
Tercios de novelas	24	18	9
Cuartos de novelas	36	24	12

En novelas completas el conjunto de prueba se formó con 3 instancias (1 por etapa) y las 6 instancias restantes sirvieron para el conjunto de entrenamiento. En medias novelas, el conjunto de prueba contó con 6 instancias (2 por etapa) y el conjunto de prueba con 12 instancias, así sucesivamente. La proporción de instancias fue de 33% para pruebas y 66% para entrenamiento. La frecuencia de *3*-gramas de las novelas de mayor extensión puede predominar sobre las más pequeñas, por ello se requiere una normalización para evitar que los *n*-gramas con el mayor rango de variación dominen a los *n*-gramas de menor frecuencia. Existen diversos métodos de estandarización. Aqui se aplicó el método que consiste en extraer la media del grupo del valor de cada variable y dividir el valor resultante por la desviación estándar.

Para tener una idea más clara de lo obtenido en esta fase, se proporcionan los siguientes ejemplos: el autor *Booth Tarkington* (BT) tiene una matriz de 9 x 7,584 en novelas completas (1) y *3*-gramas de caracteres (car), *Charles Dickens* (CD) tiene una matriz de 9 *x* 32,960 en novelas completas (1) y *3*-gramas de palabras (pal) y *Mark Twain* (MT) una matriz de 18 x 5,408 en medias novelas (2) y *3*-gramas de etiquetas POS (pos) (Véase Tabla 4.4).

4.2.5 Creación de modelos con reducción de dimensiones

Las matrices de la etapa anterior poseen una gran cantidad de características (dimensiones), del orden de miles (Véase Tabla 4.4). En el aprendizaje automático, cuando los objetos tienen alta dimensión, a menudo es conveniente reducirlas a un número mínimo (Cunningham, 2007). Como se mencionó en la sección 2.10, la reducción de dimensiones tiene múltiples beneficios. En este caso se espera que la reducción de dimensiones mejore la exactitud de clasificación de los modelos. Para aplicar la reducción de dimensiones se evaluaron dos estrategias:

- Seleccionar los *3*-gramas más frecuentes.
- Utilizar algoritmos de extracción de características.

4.2.5.1 Selección de n-gramas frecuentes

Un método simple para definir un conjunto de características representativas del estilo de un autor es obtener aquellas que se utilizan con más frecuencia. Independientemente del tipo de característica, se debe decidir qué cantidad de ellas se utilizarán. No existe un valor ideal para todas las tareas del Procesamiento del Lenguaje Natural, ya que depende de la cantidad de información y el tipo de característica estilométrica utilizada. Por ejemplo, (Baayen, Halteren y Tweedie, 1996) utilizaron 50 palabras, (Burrows, 1987) 100 palabras, (Stamatatos, 2006a) 1,000 palabras, (Velázquez, 2014) con un máximo de 11,000 utilizando caracteres, palabras y etiquetas POS como características.

Se realizaron experimentos de clasificación seleccionando *3*-gramas frecuentes en bloques de 500, 1000, 1500, 2000, ..., n. Donde n es el número máximo de *3*-gramas disponibles en el modelo. En esta etapa se observó que los *3*-gramas con mayor frecuencia de uso representaron aproximadamente un tercio del total existente en los modelos sin reducción de dimensiones (Ver Anexo C). La selección de los n-gramas frecuentemente utilizados abre la posibilidad de eliminar características de baja frecuencia que podría resultar útil en la detección de cambio de estilo de escritura.

4.2.5.2 Algoritmos de extracción de características

La reducción de dimensiones vía extracción de características crea un conjunto nuevo de características "sintéticas" a partir de la combinación de las características originales (Stamatatos, 2009). Se evaluaron 2 algoritmos para extracción de caracteristicas, Análisis de componentes principales y Análisis Semántico Latente, ambos algoritmos son implementaciones de la herramienta *scikit-learn*[7]:

En *scikit-learn*, el Análisis de Componentes Principales centra, pero no escala los datos de entrada para cada característica antes de aplicar el método. El parámetro opcional *whiten=True* permite proyectar los datos en el espacio singular mientras se escala cada componente a la variación de la unidad. Esto suele ser útil en el caso de las Máquinas de Soporte Vectorial con el *kernel* RBF y el algoritmo de agrupamiento K-Means. El parámetro *n_components*, indica el número de componentes a retener. Si este parámetro no se especifica, se retienen todos los componentes iguales al total de instancias disponibles. Si $0 < n\_components < 1$, se seleccionan el número de componentes de modo que la cantidad de variación sea menor o igual al porcentaje especificado en *n_components*.

El Análisis Semántico Latente realiza una reducción de la dimensionalidad lineal mediante la técnica Valor Singular de Descomposición. Contrario al Análisis de Componentes Principales, este estimador no centra los datos antes de calcular la descomposición del valor singular. Esto significa que puede trabajar con matrices escasamente pobladas de manera eficiente. Esta técnica trabaja con matrices de términos *tf-idf*. El parámetro *n_components* indica la dimensionalidad de los datos de salida, debe ser estrictamente menor que el número de características. El valor predeterminado 2 es útil para la visualización.

[7] http://scikit-learn.org/stable/

En la reducción de dimensiones por selección, se debe determinar el número de dimensiones del nuevo conjunto (Rea y Rea, 2016). Este parámetro importante determina cuan aceptable es la perdida de información debido a la transformación. Se evaluaron dos estrategias:

1. Utilizando un umbral que representa el porcentaje de variación retenida de los datos originales. Con esta heurística, el número de dimensiones es variable. Puede ocurrir que el umbral se alcance incluso con un solo componente. Se evaluaron los siguientes umbrales: 70, 80, 90 y 100%. Este último valor equivale a obtener un numero de dimensiones igual al total de instancias del conjunto de entrenamiento (Véase Tabla 4.5). Por defecto, las implementaciones de *scikit-learn* funcionan de esta manera. La implementación de Análisis de Componentes Principales de *scikit-learn*, permite indicar el porcentaje de varianza acumulada que se requiere de los componentes. Así, el número de componentes retenidos cambia con el modelo. Por ejemplo, suponga que se requiere el 95% de varianza. Si ocurre que el primer componente retiene 87% el segundo 10%, juntos hacen un total de 97%. Sin embargo, se comprobó que el método retiene un porcentaje menor al indicado, es decir 87% en este ejemplo.

2. Utilizando únicamente 2 dimensiones. De acuerdo a Binongo (2003), cuando se analizan los textos de diferentes autores, la diferencia de autor a menudo se puede reflejar principalmente en el gráfico de las dos primeras componentes principales. Cabe aclarar que, según la naturaleza de los datos, seleccionar un par de dimensiones puede conducir a una perdida sustancial de la información original.

El algoritmo PCA mostró mejores resultados con 2 dimensiones, mientras que el algoritmo LSA lo hizo con 6, 12, 18 y 24 dimensiones. Estos últimos valores representan el número de instancias del conjunto de entrenamiento según el tamaño del texto.

Al comparar la reducción de dimensiones basada en frecuencia contra la extracción de características, esta última logró mejores resultados en las pruebas de clasificación. En lugar de eliminar características, la extracción de características crea nuevas dimensiones a partir de las originales sin descartar *n*-gramas de baja frecuencia. Aparentemente, los *n*-gramas utilizados con poca frecuencia contienen información importante sobre el estilo de escritura del autor.

4.2.6 Pruebas de clasificación

Las matrices obtenidas en las fases previas se dividieron en conjuntos de entrenamiento y prueba. Para detectar cambios de estilo de escritura de forma confiable, se diseñó un esquema en el que cada novela en sus respectivas etapas fuera utilizada para el entrenamiento y prueba de los algoritmos de aprendizaje. Para ejemplificar lo anterior, la Tabla 4.6 muestra las novelas del autor *Booth Tarkington*.

Tabla 4.6 Novelas del autor *Booth Tarkington*.

Etapas					
Inicial		**Media**		**Final**	
Novela	Año	Novela	Año	Novela	Año
Gentleman	1899	*Penrod*	1914	*Ramsey*	1919
Vanrevels	1902	*Turmoil*	1915	*AliceAdams*	1921
Canaan	1905	*Seventeen*	1916	*Julia*	1922

El primer conjunto de prueba se formó con las instancias *Gentleman*, *Penrod* y *Ramsey* y el resto forman el conjunto de entrenamiento. El segundo conjunto de prueba se forma con *Gentleman*, *Penrod* y *Alice Adams*, el tercer conjunto de prueba lo forman *Gentleman*, *Penrod* y *Julia* y así sucesivamente. De esta forma se obtienen 27 diferentes conjuntos de entrenamiento y prueba. La Tabla 4.7 muestra los conjuntos de prueba resultantes del proceso anterior. Los conjuntos de entrenamiento no se muestran por razones de espacio.

Tabla 4.7 Conjuntos de prueba del autor *Booth Tarkington*.

Conjunto de prueba	Novelas	Conjunto de prueba	Novelas	Conjunto de prueba	Novelas
1	*Gentleman-Penrod-Ramsey*	10	*Vanrevels, Penrod, Ramsey*	19	*Canaan, Penrod, Ramsey*
2	*Gentleman-Penrod-AliceAdams*	11	*Vanrevels, Penrod, AliceAdams*	20	*Canaan, Penrod, AliceAdams*
3	*Gentleman-Penrod-Julia*	12	*Vanrevels, Penrod, Julia*	21	*Canaan, Penrod, Julia*
4	*Gentleman-Turmoil- Ramsey*	13	*Vanrevels, Turmoil, Ramsey*	22	*Canaan, Turmoil, Ramsey*
5	*Gentleman-Turmoil-AliceAdams*	14	*Vanrevels, Turmoil, AliceAdams*	23	*Canaan, Turmoil, AliceAdams*
6	*Gentleman-Turmoil- Julia*	15	*Vanrevels, Turmoil, Julia*	24	*Canaan Turmoil Julia*
7	*Gentleman-Seventeen-Ramsey*	16	*Vanrevels, Seventeen, Ramsey*	25	*Canaan, Seventeen, Ramsey*
8	*Gentleman-Seventeen-AliceAdams*	17	*Vanrevels, Seventeen, AliceAdams*	26	*Canaan, Seventeen, AliceAdams*
9	*Gentleman-Seventeen- Julia*	18	*Vanrevels, Seventeen, Julia*	27	*Canaan, Seventeen, Julia*

Como el número de instancias a clasificar de cada etapa es el mismo, la métrica exactitud (*accuracy*) resulta apropiada para medir la eficiencia de cada modelo (Garc, 2009). Con esta métrica se conoce la proporción de predicciones correctas que ha hecho el modelo del total de instancias de prueba. La exactitud de cada prueba de clasificación se acumula y al final se promedia entre los 27 experimentos. Se utilizaron cuatro algoritmos de clasificación supervisados: Máquinas de soporte vectorial SVM (SVC y LinearSVC), clasificador bayesiano con distribución multinomial (NBM) y regresión logística (LG).

La implementación del algoritmo SVM (SVC, *Support Vector Clasification*) está basada en *Libsvm*. Cuando se entrena un SVM con el núcleo de la función de base radial (RBF), se deben considerar dos parámetros: C y gamma. El parámetro C, común a todos los kernels SVM, intercambia errores de clasificación de los ejemplos de entrenamiento contra la simplicidad de la superficie de decisión. El parámetro *gamma* define cuánta influencia tiene un solo ejemplo de entrenamiento. La elección correcta de C y gamma es

crítica para el rendimiento del SVM. Otro de los clasificadores evaluados fue una variante de SVM (LinearSVC), cuyo uso se recomienda para grandes conjuntos de datos. Similar a SVC con el parámetro kernel = 'linear'.

En el algoritmo La regresión logística también se conoce como clasificación de Máxima Entropía (MaxEnt). En el caso multiclase, el algoritmo de entrenamiento usa el esquema one-vs-rest (OvR) si el parámetro multi_class='ovr'. El solucionador 'liblinear' admite la regularización con norma L1 y L2. Por defecto, el parámetro penalty='l2'. El parámetro C indica el grado o fuerza de la regularización, su valor por defecto es 1.

El algoritmo MNB implementa el algoritmo Naive Bayes para datos distribuidos multinomialmente, y es una de las dos variantes clásicas de Naive Bayes utilizadas en la clasificación de texto, donde los datos se representan típicamente como conteos de vectores de palabras. La distribución está parametrizada por $\theta_y = (\theta_{y1}, \dots, \theta_{yn})$ para cada clase y, donde n es el número de características y θ_{yi} es la probabilidad $P(x_i|y)$ de que la característica i aparezca en un ejemplo perteneciente a la clase y. El parámetro θ_y se estima mediante una versión suavizada de máxima verosimilitud, es decir, conteo de frecuencia relativa:

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n}$$

Donde $N_{yi} = \sum x \in T^{x_i}$ es el número de veces la característica i en una muestra de clase en el conjunto de entrenamiento T. $N_y = \sum_{i=1}^{n} N_{yi}$ es el recuento total de todas las características para la clase y. Si $\alpha = 1$ se denomina suavizado Laplace, Si $\alpha < 1$ se llama suavizado Lidstone. Dentro de método, por defecto el parámetro alpha=1. Los antecedentes de suavizado $\theta \geq 0$ da cuenta de las características que no están presentes en las muestras de aprendizaje y evita cero probabilidades en otros cálculos.

Los clasificadores descritos manejan espacios vectoriales con una alta dimensionalidad y son ampliamente utilizados en problemas de clasificación supervisada.

Capítulo 5. Experimentación y evaluación de resultados

Con la finalidad de simplificar la presentación de resultados, se realizó una prueba piloto con novelas completas y los diferentes *3*-gramas. La Figura 5.1 muestra la exactitud promedio de los clasificadores de Maquinas de Soporte Vectorial (Liblinear y SVM), Naive Bayes Multinomial (MNB) y Regresión Logística (LG). Liblinear y LG presentaron resultados muy similares. Se optó por este último para la presentación de resultados.

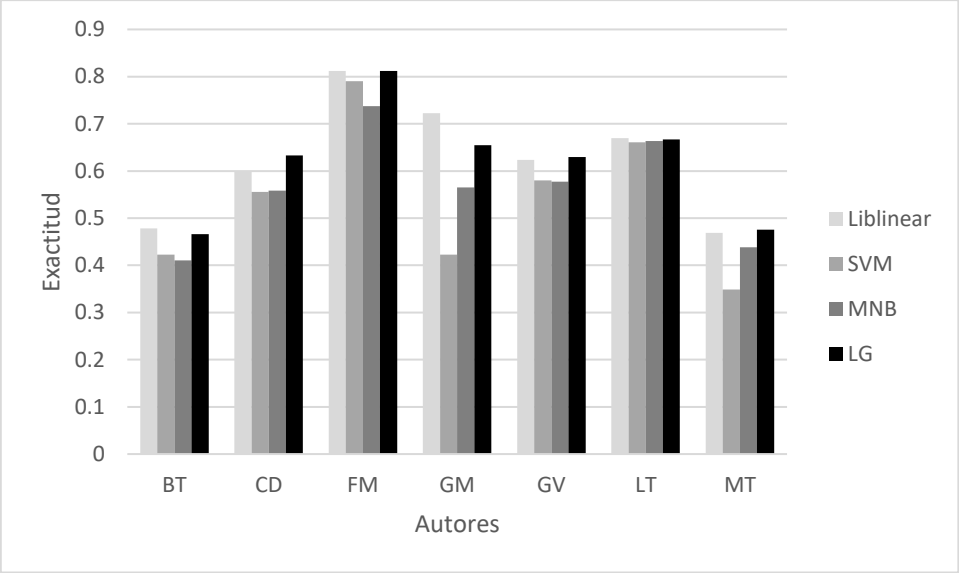


Figura 5.1 Exactitud de los algoritmos de aprendizaje automático.

Las pruebas de clasificación sirvieron para determinar en qué grado de cambio de estilo detectado por los diferentes n-gramas propuestos. Los experimentos comprendieron dos grupos: el Grupo 1 compuesto por siete autores cuyas producciones se dividieron en tres etapas y el Grupo 2, compuesto de 11 autores y dos etapas. En el Grupo 2 se suprimió la etapa media para que el intervalo de tiempo entre las clases inicial y final fuera mayor. Se espera que la exactitud sea mayor en el segundo grupo, dado que se tienen dos clases y una brecha de tiempo mayor entre ellas. Para los experimentos del Grupo 1 la línea base es de 33% de exactitud y para el Grupo 2 es de 50%. Estos porcentajes representan la probabilidad de realizar una asignación aleatoria correcta entre las clases disponibles.

5.1 Pruebas de clasificación con 3 etapas y 7 autores

Los autores y novelas evaluadas del Grupo 1 se muestran en la Tabla 5.1. La diferencia en años entre una etapa y otra no supera los 10 años en la mayoría de los casos. Si ocurren cambios de estilo, se espera una exactitud máxima de 66%: dos de cada tres muestras clasificadas correctamente.

Tabla 5.1 Experimentos de 3 etapas y 7 autores.

Autor	Etapas					
	Inicial		Media		Final	
	Año	Novela	Año	Novela	Año	Novela
Booth Tarkington (BT)	1899	*Gentleman*	1914	*Penrod*	1919	Ramsey
	1902	*Vanrevels*	1915	*Turmoil*	1921	Alice Adams
	1905	*Canaan*	1916	*Seventeen*	1922	Gentle Julia
Charles Dickens (CD)	1838	*Nicholas*	1848	*Dombey and Son*	1859	Two Cities
	1838	*Oliver Twist*	1850	*Copperdfield*	1861	Expectations
	1841	*Barnaby*	1853	*Bleak house*	1865	Our mutual friend
Frederick Marryat (FM)	1830	*The King's Own*	1839	*The panthom ship*	1845	The Mission
	1831	*Jacob Faithful*	1839	*A diary in America*	1847	New Forrest
	1831	*Newton Forster*	1840	*Olla Podrida*	1848	The Little Savage
George Macdonald (GM)	1863	*David Elginbrod*	1873	*Gutta Percha*	1888	Electrical Lady
	1864	*Adela*	1875	*A double story*	1891	Flight of Shadow
	1865	*Alec Forbes*	1876	*Thomas Wingfold*	1892	hope of góspel
George Vaizey (GV)	1901	*School Story*	1908	*Flaming June*	1914	Cassandra
	1902	*Pixie*	1908	*Big Game*	1914	College Girl
	1902	*Houseful of Girls*	1910	*Marriage*	1915	Claire
Louis Tracy (LT)	1903	*wings of morning*	1907	*The Captain*	1912	Romance of NY
	1904	*the revelers*	1909	*Inmortals*	1916	The day of wrath
	1905	*disapperance*	1909	*the stoneway girl*	1919	Mortimer fenley
Mark Twain (MT)	1869	*Innocents*	1883	*Mississippi*	1897	The Equator
	1872	*Roughing It*	1884	*Huckleberry Finn*	1905	What is man?
	1876	*Tom Sawyer*	1889	*King Arthur*	1906	Dollar

La Tabla 5.2 muestra la exactitud promedio de los *3*-gramas a través de los tamaños de texto. Todos los autores superan la línea base de 33%. Como referencia, se marcan en negritas los porcentajes iguales o superiores a 50%. Los autores *Booth Tarkington* (BT), *Charles Dickens* (CD), *Frederick Marryat* (FM), *George Macdonald* (GM) y *George*

Vaizey (GV) obtienen los porcentajes de exactitud más altos con respecto a la línea base, mientras que los autores *Louis Tracy* (LT) y *Mark Twain* (MT) los porcentajes más bajos. El porcentaje más alto de 70%, se obtiene en *3*-gramas de relaciones sintácticas (sr) para los autores *Frederick Marryat* (FM) y *George Macdonald* (GM). Los modelos con reducción (PCA) y (LSA) no superaron la exactitud de los modelos sin reducción de dimensiones (todo).

Tabla 5.2 Exactitud promedio en 3 etapas.

AUTOR	todo				PCA				LSA			
	car	**pal**	**pos**	**sr**	**car**	**pal**	**pos**	**sr**	**car**	**pal**	**pos**	**Sr**
BT	**0.62**	**0.61**	**0.68**	0.47	**0.52**	**0.50**	**0.52**	0.19	**0.62**	**0.61**	**0.68**	0.47
CD	0.44	0.48	**0.57**	**0.59**	**0.58**	**0.60**	**0.60**	**0.56**	0.45	0.48	**0.57**	**0.59**
FM	**0.58**	**0.59**	**0.62**	**0.70**	**0.52**	0.44	**0.66**	**0.56**	**0.58**	**0.60**	**0.62**	**0.70**
GM	**0.59**	**0.63**	**0.54**	**0.70**	**0.56**	**0.63**	**0.56**	**0.66**	**0.59**	**0.63**	**0.53**	**0.70**
GV	0.40	0.40	**0.61**	**0.67**	**0.56**	**0.50**	**0.64**	**0.51**	0.40	0.40	**0.61**	**0.67**
LT	0.35	0.56	0.46	0.35	0.44	0.48	0.41	0.64	0.35	**0.56**	0.46	0.35
MT	0.39	0.55	0.49	0.47	0.39	0.44	0.39	0.19	0.40	**0.55**	**0.50**	0.47

5.1.1 Modelos sin reducción de dimensiones en 3 etapas

En esta sección es posible observar cómo operan los modelos conforme se incrementa la cantidad de información en las instancias de entrenamiento y prueba. La Tabla 5.3 muestra que, en la mayor parte de los casos, la exactitud más alta se obtiene con textos de novelas completas (1). Sin embargo, algunos autores lo hacen con textos de medias novelas (2). También se observa que ningún *3*-grama predomina claramente sobre los demás. Aunque la exactitud de los *3*-gramas de relaciones sintácticas (sr) es ligeramente superior. En esta categoría, el autor *Frederick Marryat* (FM) obtuvo 76% en medias novelas (2), *George Vaizey* (GV) un 74% en novelas completas (1) y *George Macdonald* (GM) 72% en medias novelas (2). Los autores *Louis Tracy* (LT) y *Mark Twain* (MT) tienen los resultados más bajos a través del conjunto de experimentos. Los 3-gramas de etiquetas POS (pos) presentaron resultados superiores a los de caracteres (car) y palabras (pal).

Tabla 5.3 Modelos sin reducción de dimensiones y 3 etapas.

AUTOR	car				pal				pos				sr			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
BT	**0.61**	**0.61**	**0.63**	**0.64**	**0.67**	**0.62**	**0.62**	**0.55**	**0.67**	**0.70**	**0.67**	**0.68**	0.46	0.47	0.44	**0.51**
CD	0.39	0.45	0.46	0.48	0.47	0.47	0.47	0.49	**0.52**	**0.55**	**0.59**	**0.60**	**0.56**	**0.59**	**0.61**	**0.59**
FM	**0.51**	**0.58**	**0.62**	**0.60**	**0.66**	**0.59**	**0.57**	**0.57**	**0.62**	**0.59**	**0.65**	**0.63**	**0.68**	**0.76**	**0.67**	**0.69**
GM	0.36	**0.65**	**0.70**	**0.64**	0.42	**0.72**	**0.66**	**0.72**	0.23	**0.65**	**0.65**	**0.60**	**0.69**	**0.72**	**0.70**	**0.70**
GV	0.36	0.43	0.43	0.39	0.49	0.36	0.34	0.40	**0.58**	**0.62**	**0.59**	**0.66**	**0.74**	**0.64**	**0.67**	**0.64**
LT	0.32	0.32	0.38	0.36	**0.52**	**0.64**	**0.57**	**0.52**	0.44	0.49	0.45	0.45	0.38	0.32	0.34	0.35
MT	0.42	0.34	0.39	0.43	**0.54**	**0.57**	**0.56**	**0.51**	**0.51**	**0.50**	0.48	0.49	0.44	0.49	0.47	0.48

5.1.2 Modelos con reducción de dimensiones en 3 etapas

Esta sección contiene la misma información que la sección previa, pero estos modelos incluyen la reducción de dimensiones con los algoritmos PCA y LSA. La Tabla 5.4 muestra los resultados obtenidos en modelos con reducción de dimensiones PCA. El autor *Louis Tracy* (LT) mostró una mejora sustancial en la exactitud con respecto al modelo sin reducción (Véase Tabla 5.3), pasando de un máximo de 38% a un 65% de exactitud. Algo similar sucedió con los resultados de *3*-gramas de caracteres (car) y de palabras (pal) para *Charles Dickens* (CD), con un incremento en la exactitud de al menos 15%. Sin embargo, en otros autores la reducción produjo una disminución en la exactitud. Tal es caso del autor *George Vaizey* (GV). En modelos sin reducción y *3*-gramas de relaciones sintácticas (sr) pasó de 74%, a valores que apenas superan el 50%. En cuanto al tamaño de texto, los mejores resultados se obtienen con novelas completas (1) y medias novelas (2).

Tabla 5.4 Modelos con reducción de dimensiones PCA y 3 etapas.

AUTOR	car				pal				pos				sr			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
BT	**0.57**	**0.50**	**0.53**	0.48	**0.54**	**0.56**	0.47	0.41	**0.62**	**0.52**	**0.50**	0.45	0.20	0.17	0.19	0.21
CD	**0.57**	**0.58**	**0.58**	**0.60**	**0.62**	**0.59**	**0.60**	**0.59**	**0.57**	**0.61**	**0.62**	**0.61**	**0.62**	**0.55**	**0.54**	**0.53**
FM	**0.53**	**0.51**	**0.51**	**0.52**	0.46	0.43	0.43	0.43	**0.63**	**0.69**	**0.65**	**0.65**	**0.54**	**0.57**	**0.57**	**0.55**
GM	0.48	**0.58**	**0.59**	**0.59**	**0.64**	**0.62**	**0.64**	**0.63**	**0.51**	**0.57**	**0.56**	**0.58**	**0.74**	**0.68**	**0.63**	**0.60**
GV	**0.60**	**0.54**	**0.56**	**0.55**	**0.58**	0.49	0.46	0.46	**0.70**	**0.62**	**0.67**	**0.55**	0.48	**0.51**	**0.51**	**0.52**
LT	0.49	0.43	0.41	0.41	**0.56**	0.44	0.46	0.44	0.47	0.40	0.39	0.37	**0.63**	**0.62**	**0.65**	**0.65**
MT	0.41	0.42	0.37	0.34	0.41	0.45	0.42	0.46	0.42	0.40	0.37	0.37	0.22	0.15	0.18	0.20

La Tabla 5.5 muestra los resultados obtenidos en modelos con el algoritmo de reducción de dimensiones LSA. La tendencia es similar a los dos modelos expuestos previamente: a excepción de los autores *Louis Tracy* (LT) y *Mark Twain* (LT), el resto logran al menos 60% de exactitud en los distintos *3*-gramas. Sin embargo, cabe destacar la influencia del tamaño de texto en la exactitud. Al dividir una novela en bloques más pequeños ocurre que la cantidad de información disponible en los ejemplos disminuye, lo que representa un desafío mayor en la etapa de aprendizaje. Asimismo, el número de instancias a clasificar es mayor. La reducción con LSA mejoró la exactitud de autores como George Macdonald (GM) en todas las características, mientras que para los autores *Charles Dickens* (CD), *Frederick Marryat* (FM), *Louis Tracy* (LT) y *Mark Twain* (MT) en *3*-gramas de caracteres (car).

Tabla 5.5 Modelos con reducción de dimensiones LSA y 3 etapas.

AUTOR	car				pal				pos				sr			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
BT	**0.60**	**0.61**	**0.63**	**0.64**	**0.67**	**0.62**	**0.61**	**0.55**	**0.67**	**0.70**	**0.67**	**0.68**	0.46	0.47	0.44	0.51
CD	0.40	0.45	0.46	0.48	0.47	0.48	0.47	0.49	**0.52**	**0.55**	**0.59**	**0.60**	**0.56**	**0.59**	**0.61**	**0.59**
FM	**0.51**	**0.58**	**0.62**	**0.60**	**0.65**	**0.59**	**0.57**	**0.57**	**0.62**	**0.59**	**0.65**	**0.63**	**0.68**	**0.76**	**0.67**	**0.69**
GM	0.36	**0.65**	**0.70**	**0.64**	0.42	**0.72**	**0.66**	**0.72**	0.23	**0.65**	**0.65**	**0.60**	**0.69**	**0.72**	**0.70**	**0.70**
GV	0.36	0.43	0.43	0.39	0.49	0.36	0.34	0.40	**0.58**	**0.62**	**0.59**	**0.66**	**0.74**	**0.64**	**0.67**	**0.64**
LT	0.32	0.32	0.38	0.36	**0.52**	**0.64**	**0.57**	**0.52**	0.44	0.49	0.45	0.45	0.38	0.32	0.34	0.35
MT	0.42	0.34	0.40	0.43	**0.54**	**0.57**	**0.56**	**0.51**	**0.51**	**0.50**	0.48	0.49	0.44	0.49	0.47	0.48

En el Anexo D se pueden consultar los porcentajes de exactitud obtenidos en cada uno de los 27 experimentos de algunos autores.

5.2 Pruebas de clasificación con 2 etapas y 11 autores

En estos experimentos se suprimió la etapa media con la idea de que el tiempo transcurrido (en años) entre las etapas fuera mucho mayor. Como el tiempo entre etapas es mayor, se espera que la exactitud alcance máximos de 100% en algunos casos. Los 11 autores y sus novelas del Grupo 2 se muestran en la Tabla 5.6.

Tabla 5.6. Experimentos de 2 etapas y 11 autores.

Autor	Etapas				años entre etapas
	Inicial		Final		
	Año	Nombre	Año	Nombre	
Booth Tarkington (BT)	1899	*Gentleman*	1919	*Ramsey*	14
	1902	*Vanrevels*	1921	*Alice Adams*	
	1905	*Canaan*	1922	*Gentle Julia*	
Charles Dickens (CD)	1838	*Nicholas Nickleby*	1859	*Two Cities*	18
	1838	*Oliver Twist*	1861	*Expectations*	
	1841	*Barnaby*	1865	*Our mutual friend*	
Frederick Marryat (FM)	1830	*The King's Own*	1845	*The Mission*	11
	1831	*Jacob Faithful*	1847	*New Forrest*	
	1831	*Newton Forster*	1848	*The Little Savage*	
George Macdonald (GM)	1863	*David Elginbrod*	1888	*Electrical Lady*	23
	1864	*Adela*	1891	*Flight of Shadow*	
	1865	*Alec Forbes*	1892	*hope of góspel*	
George Vaizey (GV)	1901	*School Story*	1914	*Cassandra*	11
	1902	*Pixie*	1914	*College Girl*	
	1902	*Houseful of Girls*	1915	*Claire*	
Louis Tracy (LT)	1903	*Wings of morning*	1912	*Romance of NY*	7
	1904	*The revelers*	1916	*The day of wrath*	
	1905	*Disapperance*	1919	*Mortimer fenley*	
Mark Twain (MT)	1869	*Innocents Abroad*	1897	*The Equator*	21
	1872	*Roughing It*	1905	*What is man?*	
	1876	*Tom Sawyer*	1906	*Dollar*	
Arthur Conan (AC)	1887	*Study In Scarlet*	1917	*His Last Bow*	26
	1890	*Sign of the Four*	1926	*The land of mist*	
	1891	*White Company*	1927	*Sherlock Holmes*	
Edgar Rice (ER)	1912	*Princess of Mars*	1941	*Llana of Gathol*	23
	1914	*Gods of mars*	1942	*Men of Jupiter*	
	1918	*Warlord of Mars*	1944	*Land of Terror*	

Autor	Etapas				años entre etapas
	Inicial		Final		
	Año	Nombre	Año	Nombre	
John Buchan (JB)	1910	*Prester john*	1932	*Gap in the curtain*	16
	1915	*Thirtynine steps*	1936	*Island of sheeps*	
	1916	*Greenmantle*	1941	*Sick heart river*	
Irish Murdoch (IM)	1954	*Under the net*	1985	*Good Aprentice*	27
	1956	*Enchanter*	1987	*Brotherhood*	
	1958	*The bell*	1995	*Jackson's Dilema*	

La Tabla 5.7 muestra la exactitud promedio de los tipos de n-gramas a través de los tamaños de texto. Todos los autores superan la línea base de 50% de exactitud. Los resultados son considerablemente superiores al grupo 1, en algunos casos se registraron porcentajes con 100% de exactitud. Los autores *Booth Tarkington* (BT), *Edgar Rice* (ER), *John Buchan* (JB) e *Iris Murdoch* (IM) tuvieron los mejores porcentajes de clasificación.

Tabla 5.7 Exactitud promedio en 2 etapas.

AUTOR	Todo				PCA				LSA			
	car	pal	pos	sr	car	pal	pos	sr	car	pal	pos	sr
BT	**0.94**	**0.90**	**0.98**	**0.81**	**0.90**	**0.83**	**0.88**	0.54	**0.89**	**0.73**	**0.92**	**0.70**
CD	0.66	**0.87**	**0.88**	**0.89**	0.54	**0.76**	0.66	0.55	0.64	**0.70**	**0.86**	**0.87**
FM	**0.73**	**0.71**	**0.81**	**0.89**	**0.90**	**0.69**	**0.98**	**0.86**	**0.71**	0.68	**0.79**	**0.88**
GM	**0.76**	**0.69**	0.55	**0.86**	**0.69**	**0.83**	**0.69**	**0.73**	**0.72**	0.60	0.54	**0.79**
GV	0.48	0.64	0.67	**0.80**	**0.93**	**0.78**	**0.89**	**0.75**	0.48	0.50	0.61	**0.80**
LT	0.67	0.59	0.59	0.52	**0.76**	0.65	0.65	**0.99**	0.62	0.54	0.66	0.50
MT	0.65	**0.82**	**0.75**	**0.80**	0.47	0.44	0.55	0.42	0.65	**0.74**	**0.73**	**0.75**
AC	**0.76**	0.64	**0.70**	**0.81**	0.46	0.42	0.54	0.55	0.67	0.49	**0.72**	0.85
ER	**1.00**	**0.87**	**0.98**	**0.95**	**0.98**	**0.69**	**0.99**	**1.00**	**0.99**	0.68	**0.94**	**0.93**
JB	**0.87**	**0.92**	**0.89**	**0.94**	0.58	**0.96**	**0.71**	**0.83**	**0.81**	**0.86**	**0.88**	**0.94**
IM	**0.84**	**0.99**	**0.87**	**0.99**	**0.78**	**0.80**	0.60	**0.83**	**0.80**	**0.90**	**0.85**	**0.99**

Otro aspecto importante de la Tabla 5.7 es que todos los *3*-gramas mejoraron significativamente la exactitud, los *3*-gramas de relaciones sintácticas (sr) mostraron eficiencia similar o mejor a las otras categorías. En general, la exactitud de los modelos sin reducción (Todo) fue ligeramente superior a los modelos con reducción (PCA) y (LSA).

5.2.1 Modelos sin reducción de dimensiones con 2 etapas

La Tabla 5.8 muestra la exactitud promedio de los 9 experimentos en modelos sin reducción de dimensiones. Destacan los resultados de los autores *Booth Tarkington* (BT), *Edgar Rice* (ER), *John Buchan* (JB) e *Iris Murdoch* (IM), al obtener 100% de exactitud en al menos uno de los *3*-gramas. Los altos porcentajes de clasificación parecen una consecuencia del mayor tiempo de separación en años que existen entre las etapas. Los resultados de *Louis Tracy* (LT) y *Mark Twain* (MT) mejoran de forma significativa respecto a los experimentos con tres etapas.

Tabla 5.8 Modelos sin reducción de dimensiones en 2 etapas.

AUTOR	car				pal				pos				sr			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
BT	**0.94**	**0.92**	**0.94**	**0.96**	**1.00**	**0.97**	**0.83**	**0.81**	**1.00**	**0.97**	**0.96**	**0.97**	**0.83**	**0.83**	**0.80**	**0.78**
CD	0.61	0.67	0.67	**0.70**	**0.89**	**0.89**	**0.83**	**0.86**	**0.78**	**0.92**	**0.91**	**0.90**	**0.89**	**0.92**	**0.91**	**0.86**
FM	0.67	**0.72**	**0.76**	**0.76**	**0.78**	**0.72**	0.67	0.66	**0.78**	**0.81**	**0.82**	**0.84**	**0.89**	**0.92**	**0.85**	**0.89**
GM	**0.72**	**0.78**	**0.76**	**0.77**	**0.72**	**0.69**	0.61	**0.74**	0.61	0.47	0.58	0.56	**0.89**	**0.86**	**0.85**	**0.85**
GV	0.50	0.50	0.46	0.45	0.67	0.67	0.63	0.61	0.61	0.67	0.65	**0.74**	**0.83**	**0.78**	**0.80**	**0.79**
LT	**0.72**	0.67	0.65	0.65	0.61	0.64	0.61	0.49	0.61	0.64	0.61	0.49	0.50	0.50	0.56	0.53
MT	0.61	0.58	**0.70**	**0.70**	**0.83**	**0.83**	**0.83**	**0.78**	**0.78**	**0.78**	**0.74**	**0.71**	**0.78**	**0.83**	**0.78**	**0.81**
AC	**0.78**	**0.75**	**0.74**	**0.75**	**1.00**	**0.75**	0.39	0.42	0.67	**0.78**	0.69	0.66	**0.83**	**0.81**	**0.80**	**0.82**
ER	**1.00**	**1.00**	**1.00**	**1.00**	**0.89**	**0.86**	**0.89**	**0.82**	**1.00**	**1.00**	**0.94**	**0.96**	**1.00**	**0.94**	**0.92**	**0.92**
JB	**0.89**	**0.86**	**0.87**	**0.85**	**1.00**	**0.92**	**0.89**	**0.86**	**1.00**	**0.83**	**0.89**	**0.85**	**1.00**	**0.92**	**0.94**	**0.92**
IM	**0.89**	**0.83**	**0.81**	**0.81**	**1.00**	**1.00**	**0.98**	**0.96**	**0.89**	**0.89**	**0.89**	**0.82**	**1.00**	**1.00**	**1.00**	**0.96**

5.2.2 Modelos con reducción de dimensiones en 2 etapas

La Tabla 5.9 muestra los resultados obtenidos en modelos con reducción de dimensiones PCA. Para los autores *Frederick Marryat* (FM) y *George Vaizey* (GV) la reducción de dimensiones con PCA mejoró la exactitud con respecto al modelo sin reducción. Pero en los autores *Mark Twain* (MT), *Arthur Conan* (AC), se produjo una disminución en la exactitud en los distintos *3*-gramas. Y para *Booth Tarkington* (BT) y *Charles Dickens* (CD) también presentaron una disminución en la exactitud en la categoría de relaciones sintácticas (sr).

Tabla 5.9 Modelos con reducción de dimensiones PCA en 2 etapas.

AUTOR	car				pal				pos				sr			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
BT	**1.00**	**0.89**	**0.91**	**0.82**	**0.89**	**0.86**	**0.85**	**0.70**	**0.94**	**0.89**	**0.85**	**0.82**	0.56	0.56	0.54	0.52
CD	0.56	0.53	0.52	0.54	**0.83**	**0.81**	0.68	**0.72**	0.67	0.67	0.68	0.64	0.56	0.58	0.52	0.53
FM	**0.89**	**0.92**	**0.91**	**0.90**	**0.78**	**0.72**	0.59	0.67	**1.00**	**1.00**	**0.96**	**0.95**	**0.83**	**0.86**	**0.87**	**0.89**
GM	0.56	0.67	**0.78**	**0.76**	**0.83**	**0.83**	**0.83**	**0.82**	0.56	**0.75**	**0.72**	**0.75**	0.67	**0.81**	**0.78**	0.68
GV	**1.00**	**0.92**	**0.92**	**0.89**	**0.78**	**0.78**	**0.80**	**0.77**	**1.00**	**0.83**	**0.96**	**0.78**	**0.72**	**0.75**	**0.76**	**0.76**
LT	**0.78**	**0.78**	**0.76**	**0.71**	**0.78**	0.58	0.63	0.61	0.67	**0.72**	0.63	0.57	**1.00**	**1.00**	**0.98**	**0.97**
MT	0.50	0.44	0.48	0.47	0.50	0.39	0.39	0.49	0.56	0.56	0.54	0.56	0.50	0.42	0.37	0.38
AC	0.44	0.44	0.46	0.50	0.44	0.44	0.43	0.38	0.50	0.58	0.52	0.54	0.50	0.58	0.56	0.57
ER	**1.00**	**0.98**	**0.98**	**0.97**	**0.78**	**0.72**	0.65	0.61	**1.00**	**1.00**	**0.98**	**0.99**	**1.00**	**1.00**	**1.00**	**1.00**
JB	**0.72**	0.61	0.52	0.46	**1.00**	**0.97**	**0.94**	**0.93**	**0.78**	**0.78**	0.67	0.61	**0.89**	**0.86**	**0.80**	**0.77**
IM	**0.83**	**0.78**	**0.76**	**0.74**	**0.83**	**0.78**	**0.80**	**0.79**	0.61	0.61	0.61	0.58	**0.83**	**0.83**	**0.83**	**0.83**

La Tabla 5.10 muestra los resultados obtenidos en modelos con reducción de dimensiones LSA. Los mejores resultados se observan en *3*-gramas de relaciones sintácticas (*sr*). Para el autor *Mark Twain* (MT), la reducción con este método produjo un incremento en la exactitud. Llama la atención que, para este mismo autor, la reducción con PCA disminuyó la exactitud (Véase Tabla 5.9).

Tabla 5.10. Modelos con reducción de dimensiones LSA en 2 etapas.

AUTOR	car				pal				pos				sr			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
BT	**0.72**	**0.92**	**0.94**	**0.96**	0.56	**0.78**	**0.81**	**0.78**	**0.89**	**0.92**	**0.92**	**0.95**	**0.67**	**0.72**	**0.69**	**0.74**
CD	0.56	0.64	**0.67**	**0.71**	0.50	0.64	**0.83**	**0.83**	**0.72**	**0.89**	**0.91**	**0.92**	**0.78**	**0.92**	**0.91**	**0.86**
FM	0.61	**0.69**	**0.76**	**0.76**	0.61	**0.72**	**0.70**	0.67	**0.72**	**0.78**	**0.82**	**0.84**	**0.89**	**0.92**	**0.85**	**0.88**
GM	0.56	**0.78**	**0.78**	**0.78**	0.56	0.64	0.63	0.57	0.61	0.44	0.59	0.53	0.67	**0.83**	**0.82**	**0.85**
GV	0.50	0.50	0.48	0.45	0.33	0.50	0.56	0.63	0.61	0.61	0.57	0.65	**0.83**	**0.81**	**0.80**	**0.78**
LT	0.56	0.64	0.65	0.63	0.50	0.50	0.57	0.60	0.61	**0.72**	0.65	0.66	0.44	0.50	0.56	0.52
MT	**0.67**	0.58	0.65	**0.70**	**0.72**	**0.72**	**0.81**	**0.70**	**0.72**	**0.75**	**0.70**	**0.73**	**0.67**	**0.75**	**0.80**	**0.79**
AC	0.61	**0.69**	**0.69**	**0.67**	0.56	0.56	0.41	0.43	**0.67**	**0.69**	**0.76**	**0.75**	**0.89**	**0.83**	**0.83**	**0.85**
ER	**1.00**	**0.97**	**1.00**	**1.00**	0.61	0.56	**0.78**	**0.77**	**1.00**	**0.97**	**0.91**	**0.89**	**1.00**	**0.94**	**0.87**	**0.90**
JB	**0.72**	**0.86**	**0.85**	**0.82**	**0.83**	**0.86**	**0.87**	**0.86**	**0.94**	**0.83**	**0.89**	**0.85**	**1.00**	**0.92**	**0.94**	**0.91**
IM	**0.78**	**0.81**	**0.81**	**0.81**	**0.72**	**1.00**	**0.96**	**0.92**	**0.89**	**0.83**	**0.87**	**0.81**	**1.00**	**1.00**	**1.00**	**0.96**

5.3 Pruebas de significancia

Las técnicas estadísticas incluyen pruebas de significancia que se aplican con el fin de determinar cuál es la probabilidad de que la diferencia entre dos o más grupos no se deba al azar. El grado de probabilidad o el nivel de significación de un análisis está reflejado por el *p-valor*. Las pruebas de significancia siempre parten de una hipótesis que asume lo opuesto a lo que se pretende demostrar con el análisis estadístico. A esta hipótesis se le conoce como *hipótesis nula*. Para confirmar la presencia de cambios significativos en el estilo, se realizó una prueba de hipótesis mediante la prueba *Wilcoxon Signed Ranks* (Wilcoxon 1945) con el siguiente planteamiento:

Hipótesis nula (H_0): no existe cambio de estilo de escritura.
Hipótesis alternativa (H_1): existe cambio de estilo de escritura.

Demšar (2006) aportó dos argumentos a favor de este tipo de prueba: desde el punto de vista estadístico es más segura ya que no asume distribuciones normales y que los valores atípicos tienen menos efecto sobre la media. Si $p \leq 0.05$, las diferencias entre los datos se pueden considerar significativas y, por consiguiente, se descarta la hipótesis nula. Las pruebas se llevaron a cabo con los niveles de significancia $p < 0.05$, para indicar cambios significativos y $p < 0.01$ cambios muy significativos. Cuanto más pequeño es el valor de *p* más evidencia existe para rechazar H_0.

La prueba de significancia se realizó con los autores del Grupo 1, ya que los cambios de estilo detectados en el grupo 2 son más que evidentes. La Tabla 5.11 muestra los resultados de la prueba con $p < 0.05$ en textos de novelas completas y medias novelas. Para facilitar la interpretación, se utilizó la siguiente notación: si $p < 0.05$, la leyenda **1** indica se cumplió la condición de que no existe evidencia suficiente para rechazar H_0 y que se acepta H_1. La leyenda **0** indica que no hay evidencia suficiente para rechazar a H_0.

Tabla 5.11. Prueba de significancia con *p*<0.05 para *3*-gramas del Grupo 1.

Autor	Novelas completas				Medias novelas			
	car	pal	pos	sr	car	pal	pos	sr
BT	1	1	1	0	1	1	1	0
CD	1	1	1	1	1	1	1	1
FM	1	1	1	1	1	1	1	1
GM	1	1	1	1	1	1	1	1
GV	1	1	1	1	1	1	1	1
LT	1	1	1	1	1	1	1	1
MT	1	1	1	1	1	1	1	0

La prueba arrojó que el autor *Booth Tarkington* (BT) no presentó cambios significativos en *3*-gramas de relaciones sintácticas (sr) en novelas completas y medias novelas. El autor *Mark Twain* (MT) no mostró cambios significativos en medias novelas. De acuerdo con planteamiento, el resto de los autores presentó cambios significativos en el estilo de escritura.

La Tabla 5.12 muestra los resultados de la prueba con $p < 0.01$. Aquí, el autor *Booth Tarkington* (BT) no presentó cambios muy significativos en 3-gramas de relaciones sintácticas (sr) en ambos textos ni en la categoría de etiquetas POS (pos) en novelas completas. Por su parte, el autor *Mark Twain* (MT) no presentó cambios muy significativos en *3*-gramas de relaciones sintácticas (sr) en ambos tamaños de texto. El resto de los autores presentan cambios muy significativos.

Tabla 5.12 Prueba de significancia con *p*<0.01 para *3*-gramas del Grupo 1.

Autor	Novelas completas				Medias novelas			
	car	pal	pos	sr	car	pal	pos	sr
BT	1	1	0	0	1	1	1	0
CD	1	1	1	1	1	1	1	1
FM	1	1	1	1	1	1	1	1
GM	1	1	1	1	1	1	1	1
GV	1	1	1	1	1	1	1	1
LT	1	1	1	1	1	1	1	1
MT	1	1	1	0	1	1	1	0

Los resultados de la prueba de significancia coinciden con los observados a lo largo de los experimentos del Grupo 1: los autores *Booth Tarkington* y *Mark Twain* fueron los que presentaron la exactitud más baja con respecto a la línea base de 33%.

5.4 Experimentos con *n*-gramas de palabras

Adicionalmente, se realizaron experimentos con *n*-gramas para conocer el porcentaje de clasificación correcta con este tipo de característica estilométrica. Los n-gramas evaluados fueron de longitud $n = \{1,2,3,4\}$ en textos de novelas completas (1) y medias novelas (2), ya que son los textos que mejores resultados aportan. La Tabla 5.13 muestra los promedios de exactitud del Grupo 1. La columna N representa el tamaño del n-grama. Las entradas con el símbolo $+$ indican combinaciones de n-gramas de distintos tamaños. Los porcentajes de exactitud de autores como *Charles Dickens* (CD), *Frederick Marryat* (FM) y *Louis Tracy* (LT) son de al menos 80% en la mayoría de los casos. El resto de los autores superan el 50% de exactitud. La combinación de n-gramas de distintos tamaños tiene la finalidad de observar una mejora en la métrica. Sin embargo, esto no se cumplió en todos los casos. Los resultados de 1-gramas de palabras fueron iguales o superiores al resto de n-gramas y sus combinaciones.

Tabla 5.13 Experimentos con *n*-gramas de palabras y PCA del Grupo 1.

N	BT		CD		FM		GM		GV		LT		MT	
	1	2	1	2	1	2	1	2	1	2	1	2	1	2
1	0.49	**0.53**	**0.86**	**0.86**	**0.89**	**0.83**	**0.65**	**0.61**	**0.63**	**0.65**	**0.77**	**0.69**	**0.59**	**0.62**
2	**0.54**	**0.62**	**0.83**	**0.80**	**0.84**	**0.72**	**0.64**	**0.64**	**0.56**	**0.56**	**0.84**	**0.78**	**0.58**	**0.59**
3	**0.53**	**0.57**	**0.75**	**0.72**	**0.72**	**0.61**	**0.51**	**0.56**	0.44	0.39	**0.70**	**0.59**	**0.63**	**0.64**
4	0.25	0.36	**0.75**	**0.72**	**0.80**	**0.80**	0.49	0.43	**0.89**	**0.48**	**0.53**	**0.56**	**0.51**	**0.52**
1+2	**0.53**	**0.58**	**0.81**	**0.86**	**0.88**	**0.80**	**0.63**	**0.65**	**0.62**	**0.54**	**0.80**	**0.75**	**0.60**	**0.63**
1+3	**0.52**	**0.54**	**0.78**	**0.83**	**0.89**	**0.80**	**0.60**	**0.64**	**0.58**	**0.62**	**0.77**	**0.71**	**0.53**	**0.66**
1+4	0.49	**0.53**	**0.83**	**0.87**	**0.89**	**0.83**	**0.65**	**0.62**	**0.65**	**0.64**	**0.77**	**0.69**	**0.57**	**0.60**
2+3	**0.56**	**0.62**	**0.81**	**0.78**	**0.81**	**0.69**	**0.65**	**0.63**	0.48	**0.54**	**0.84**	**0.78**	**0.62**	**0.67**
2+4	**0.54**	**0.61**	**0.84**	**0.80**	**0.84**	**0.70**	**0.65**	**0.63**	**0.57**	**0.56**	**0.84**	**0.78**	**0.57**	**0.64**
3+4	**0.52**	**0.57**	**0.73**	**0.71**	**0.74**	**0.63**	**0.53**	**0.58**	**0.51**	0.44	**0.70**	**0.62**	**0.67**	**0.66**
1+2+3+4	**0.53**	**0.57**	**0.81**	**0.83**	**0.85**	**0.80**	**0.63**	**0.65**	**0.62**	**0.59**	**0.85**	**0.76**	**0.60**	**0.61**

Los resultados de la Tabla 5.13 también muestran cambios significativos en el estilo de escritura. Pero se debe recordar que el análisis de estilo basado en palabras puede no ser concluyente debido a la influencia temática.

5.5 Experimentos con 1-gramas de palabras

Para corroborar los cambios detectados, se realizó un experimento adicional con los autores de Grupo 2. La Tabla 5.14 permite observar los porcentajes de exactitud de las características propuestas con respecto al vocabulario del autor. Los resultados superiores de lista de palabras sugieren que el uso de palabras (sea de función o de contenido semántico) cambió a lo largo del tiempo. Otro fenómeno ya observado en los experimentos anteriores es que, en algunos casos la reducción de dimensiones tiene un efecto negativo sobre la exactitud.

Tabla 5.14 Exactitud de *1*-gramas vs *3*-gramas del Grupo 2.

1-gramas de palabras				*3*-gramas											
Autor	todo	PCA	LSA	Todo				PCA				LSA			
				car	pal	pos	sr	car	pal	pos	sr	car	pal	pos	sr
BT	**0.97**	**0.88**	**0.98**	**0.94**	**0.90**	**0.98**	**0.81**	**0.90**	**0.83**	**0.88**	0.54	**0.89**	**0.73**	**0.92**	**0.70**
CD	0.60	0.54	0.63	0.66	**0.87**	**0.88**	**0.89**	0.54	**0.76**	0.66	0.55	0.64	**0.70**	**0.86**	**0.87**
FM	0.60	**0.91**	0.60	**0.73**	**0.71**	**0.81**	**0.89**	**0.90**	**0.69**	**0.98**	**0.86**	**0.71**	0.68	**0.79**	**0.88**
GM	0.64	0.61	0.62	**0.76**	**0.69**	0.55	**0.86**	**0.69**	**0.83**	**0.69**	**0.73**	**0.72**	0.60	0.54	**0.79**
GV	0.67	**0.92**	0.66	0.48	0.64	0.67	**0.80**	**0.93**	**0.78**	**0.89**	**0.75**	0.48	0.50	0.61	**0.80**
LT	**0.86**	**0.72**	**0.80**	0.67	0.59	0.59	0.52	**0.76**	0.65	0.65	**0.99**	0.62	0.54	0.66	0.50
MT	**0.72**	0.46	0.68	0.65	**0.82**	**0.75**	**0.80**	0.47	0.44	0.55	0.42	0.65	**0.74**	**0.73**	**0.75**
AC	**0.81**	0.48	**0.75**	**0.76**	0.64	**0.70**	**0.81**	0.46	0.42	0.54	0.55	0.67	0.49	**0.72**	0.85
ER	**0.98**	**0.95**	**0.96**	**1.00**	**0.87**	**0.98**	**0.95**	**0.98**	**0.69**	**0.99**	**1.00**	**0.99**	0.68	**0.94**	**0.93**
JB	**0.96**	**0.75**	**0.96**	**0.87**	**0.92**	**0.89**	**0.94**	0.58	**0.96**	**0.71**	**0.83**	**0.81**	**0.86**	**0.88**	**0.94**
IM	**0.79**	0.46	**0.73**	**0.84**	**0.99**	**0.87**	**0.99**	**0.78**	**0.80**	0.60	**0.83**	**0.80**	**0.90**	**0.85**	**0.99**

De manera general, el porcentaje de clasificación correcta de *1*-gramas de palabras indica cambios considerables en el estilo de escritura. Este diagnóstico se vuelve confiable al cotejarlo con los resultados de clasificación de los *3*-gramas de distintos tipos.

5.6 Análisis del tamaño de los textos

La Figura 5.2 muestra que, en modelos sin reducción de dimensiones la exactitud más alta se logra en textos de novelas completas (1) y *3*-gramas de relaciones sintácticas (sr).

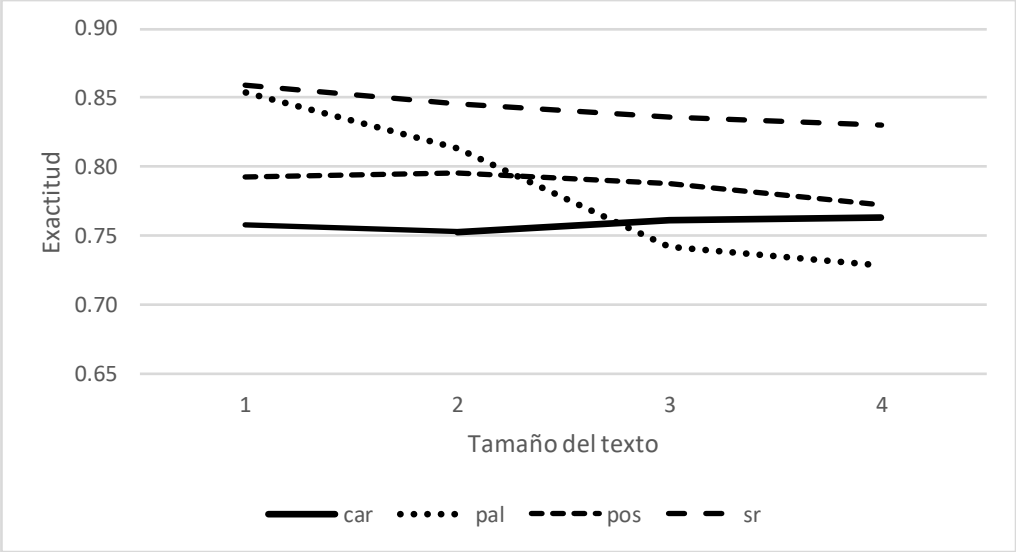


Figura 5.2 Promedio de exactitud en modelos sin reducción.

La Figura 5.3 muestra que en modelos con reducción de dimensiones PCA, los 3-gramas de etiquetas POS (pos) y relaciones sintácticas (sr) obtienen mejores resultados en medias novelas (2). Mientras que los *3*-gramas de caracteres (car) y palabras (pal) en textos de novelas completas (1). Se observa que conforme disminuye el tamaño de bloque se registra un descenso gradual de la exactitud.

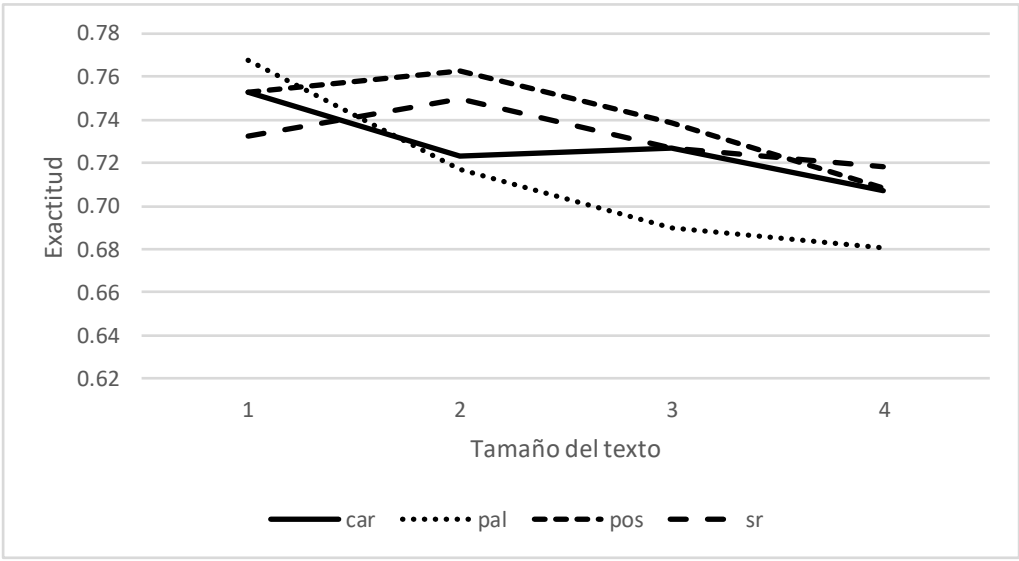


Figura 5.3 Promedio de exactitud en modelos con reducción PCA.

La Figura 5.4 muestra que en modelos con reducción de dimensiones LSA, los distintos *3*-gramas logran porcentajes similares, siendo los más altos los de *3*-gramas de relaciones sintácticas (sr), con al menos 80% de exactitud. Con esta técnica de reducción de dimensiones, las novelas completas (1) muestran porcentajes ligeramente inferiores al resto de los bloques. Aparentemente, en la reducción con LSA la exactitud aumenta conforme disminuye el tamaño de los textos.

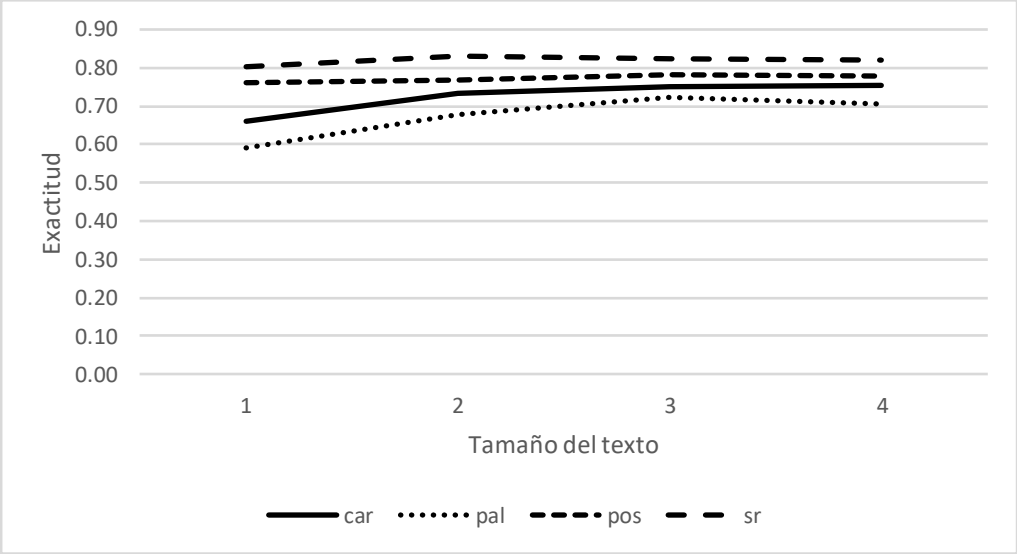


Figura 5.4 Promedio de exactitud en modelos con reducción LSA.

Un último experimento realizado en el grupo 1 consistió en entrenar con las novelas de autor y evaluar con las novelas de un autor distinto. Se espera que la exactitud máxima apenas supere el 33%. Ya que los datos que se están evaluados no provienen del mismo autor. Los resultados que se obtuvieron permitieron confirmar lo esperado. Prácticamente todos los experimentos presentaron una exactitud máxima de 33%. Véase el Anexo D.

Conclusiones y trabajos futuros

6.1 Conclusiones

En esta tesis se abordó el problema de la detección de cambio de estilo de escritura a través del tiempo utilizando novelas escritas por autores de largas carreras literarias. En este contexto, el cambio de estilo se refiere a la variación en la frecuencia de uso de las características estilométricas presentes en los textos. Las características aplicadas en la investigación se denominan n-gramas, constituidos por elementos de distinta naturaleza: caracteres, palabras, etiquetas POS y relaciones sintácticas. Dichas características se utilizaron con un enfoque de aprendizaje automático supervisado.

La propuesta supone la existencia de al menos dos etapas. De forma que al comparar los patrones de uso de *3*-gramas entre las etapas, los algoritmos de aprendizaje automático determinen a qué etapa pertenece una muestra de novela. Cuando más alto sea el porcentaje de clasificación correcta, más evidente es el cambio de estilo detectado. Para validar la propuesta de solución, se creó una muestra de 11 autores y derivado de los experimentos realizados se llegaron a las siguientes conclusiones:

El estilo de escritura cambia a través del tiempo. Los cambios en las frecuencias de uso de los *3*-gramas permitió crear modelos de predicción confiables, ya que el menos dos de cada tres muestras fueron clasificadas de forma correcta. Esta conclusión se deriva de los porcentajes observados de al menos 66% hasta 100% en las pruebas con tres y dos etapas. Los resultados demuestran que, la variación estadísticamente significativa expresada en cambios de la frecuencia de uso de *3*-gramas, pudo identificar rasgos distintivos del estilo del autor a través de las etapas. Este cambio se manifiesta de manera más clara en la escritura cuando existe al menos 10 años de diferencia entre las etapas evaluadas.

Dentro de las características estilométricas propuestas, algunas de ellas ya han sido utilizadas en la detección de cambio de estilo de escritura, principalmente las

relacionadas a características léxicas (palabras), etiquetas POS y aquellas que utilizan información sintáctica. Sin embargo, los trabajos identificados en el estado del arte no utilizan los *n*-gramas sintácticos de relaciones de dependencias como un recurso para caracterizar el estilo de un autor. Los resultados obtenidos con este tipo de n-gramas muestran que son una opción viable, ya que su desempeño fue igual y en muchos casos, mejor que los otros n-gramas. Además, son robustos a los cambios de tópico del documento. Los n-gramas sintácticos también pueden constituirse con palabras y etiquetas POS. Estos factores les permiten identificar patrones de uso que no son visibles a nivel superficial del texto.

En cuanto a la cantidad de información presente en los textos, los mejores resultados de clasificación se obtuvieron utilizando los textos de las novelas completas o en su defecto, media novelas. De acuerdo con las novelas y las características estilométricas aquí evaluadas, se requieren de al menos 1,000 oraciones para obtener modelos confiables. En problemas de la vida real, esto puede ser una limitante pues no siempre se dispone de documentos con mucho contenido. Estimar el valor exacto de este parámetro es difícil, ya que depende de la extensión del texto, el tipo de característica utilizada y las frecuencias que estas registran.

La reducción de dimensiones debe aplicarse con reservas, ya que no siempre se obtiene una mejora en la métrica aplicada. La reducción de dimensiones por selección de *n*-gramas frecuentes no resultó favorable, ya que la mayor parte de los 3-gramas que poblaron los modelos fueron de baja frecuencia. Es probable que establecer un umbral de corte elimina 3-gramsa que poseen información importante acerca del estilo de escritura del autor. En lo que respecta a la reducción de dimensiones por extracción, los algoritmos Análisis de Componentes Principales y Análisis Semántico Latente no eliminan características. En su lugar, estos algoritmos transforman el conjunto de características a uno nuevo de menores dimensiones. Estas técnicas adolecen del mismo problema de la selección por frecuencia: establecer cuál es el número apropiado de dimensiones del nuevo modelo. La reducción excesiva puede disminuir el desempeño de un modelo de aprendizaje. Hacen falta realizar otros experimentos para averiguar si la

naturaleza discreta de los datos influye en el proceso de reducción de dimensiones. Con base en los resultados obtenidos, se recomienda que, para el cambio de estilo de escritura, deben considerarse todas las características del modelo y utilizar la reducción de dimensiones para fines de representación gráfica.

6.2 Trabajos futuros

Como en cualquier otro proyecto de investigación, existen diversas líneas de investigación que quedan abiertas y en las que es posible continuar trabajando. A continuación, se presentan algunos trabajos futuros que pueden desarrollarse como productos de esta investigación o por exceder los alcances de la misma, no han podido ser tratados con la suficiente profundidad.

- Evaluar la exactitud de los *n*-gramas sintácticos de palabras y *n*-gramas sintácticos de etiquetas POS en la detección de cambio de estilo de escritura. Los *n*-gramas sintácticos no siguen el orden lineal de las palabras en el texto, esta peculiaridad puede ayudar a descubrir patrones distintos a los identificados con los *n*-gramas de relaciones sintácticas.

- Explorar el uso de herramientas de visualización de datos multidimensionales como *glueviz* u *orange3*. Esto proporcionará una mejor comprensión del comportamiento o la evolución de las características a través del tiempo.

- Diseñar una estrategia para el aprovechamiento de características estilométricas con baja frecuencia de uso. Utilizar un enfoque distinto al aprendizaje automático dará una nueva perspectiva a estas características, que por lo general son consideradas como irrelevantes.

- Realizar un análisis exhaustivo sobre los patrones sintácticos que obtienen los *n*-gramas obtenidos del árbol de dependencias. Por ejemplo: la construcción sintáctica *ccomp[nsubj,aux]* aparece con mucha frecuencia. Identificar las oraciones que contienen tal construcción e identificar la

similitud entre ellas puede llevar el análisis de estilo a un nivel abstracto y no como un mero recuento de frecuencias.

- Realizar experimentos con Aprendizaje Profundo (*Deep Learning*). La idea central sobre este trabajo es el de permitir que el Deep Learning identifique las características estilométricas más representativas del estilo de un autor.

6.3 Publicaciones realizadas

La publicación realizada con este trabajo de investigación fue:

- Germán Ríos-Toledo, Noé Alejandro Castro Sánchez, Grigori Sidorov, Juan-Pablo Posadas-Durán. (2019). Identificación de cambios en el estilo de escritura literaria con el aprendizaje automático. *Onomázein*. DOI: 10.7764/onomazein.46.04

Referencias

Addellatif Rahmoun, Zakaria Elberrichi. 2007. “Experimenting N-Grams in Text Categorization.” *The international Arab Journals of Information Tecnology* 4: 377–85.

Alzahrani, Salha, Naomie Salim, and Ajith Abraham. 2012. “Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods.” *EEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews.* 42: 133–49.

Baayen, Harald, Hans Van Halteren, and Fiona Tweedie. 1996. “Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution.” *Literary and Linguistic Computing* 11(3): 121–32.

Barrón-Cedeño, A., and P. Rosso. 2009. “Based on n -Grams Comparison.” *European Conference on Information Retrieval*: 696–700.

Binongo, José Nilo G. 2003. “Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution.” *Chance* 16(2): 9–17.

Brennan, Michael, Sadia Afroz, and Rachel Greenstadt. 2012. “Adversarial Stylometry.” *ACM Transactions on Information and System Security* 15(3): 1–22.

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

Can, Fazli, and Jon M. Patton. 2004. “Change of Writing Style with Time.” *Computers and the Humanities* 38(1): 61–82.

Chambers, J. K. (2007). Sociolinguistics. *The Blackwell encyclopedia of sociology*.

Chaski, Carole E. 2006. “Empirical Evaluations of Language-Based Author Identification Techniques.” *Forensic Linguistics* 8(1): 1–65.

Comelles, Elisabet, Victoria Arranz, and Irene Castellón. 2010. “Constituency and Dependency Parsers." *Procesamiento del lenguaje natural*, *45*, 59-66.

Cunningham, P´adraig. 2007. “Dimension Reduction.” *Machine Learning for Multimedia Content Analysis* 1: 15–35. http://link.springer.com/10.1007/978-0-387-69942-4_2.

Demšar, Janez. 2006. “Statistical Comparisons of Classifiers over Multiple Data Sets.” *Journal of Machine Learning Research* 7: 1–30.

Durán, JPF Posadas. 2017. “Detección Automática de Plagio Usando Información Sintáctica.” http://tesis.ipn.mx/handle/123456789/21664.

Escalante, Hugo Jair, Thamar Solorio, and Manuel Montes-y-Gomez. 2011. “Local Histograms of Character N -Grams for Authorship Attribution.” *Computational Linguistics*: 288–98. http://www.aclweb.org/anthology/P11-1030.

Forsyth, Rs. 1999. “Stylochronometry with Substrings, or: A Poet Young and Old.” *Literary and Linguistic Computing* 14(4): 467–78.

Garc, V. 2009. “Index of Balanced Accuracy : A Performance Measure.” *Evaluation*: 441–48.

Garrard, Peter, Lisa M. Maloney, John R. Hodges, and Karalyn Patterson. 2005. “The Effects of Very Early Alzheimer’s Disease on the Characteristics of Writing by a Renowned Author.” *Brain* 128(2): 250–60.

Genkin, Alexander, and David D. Lewis. 2005. “Author Identification on the Large Scale.” *Proceedings of the Meeting of the Classification Society of North America*: 1–20.

Golcher, Felix. 2007. “A New Text Statistical Measure and Its Application to Stylometry.” *Proceedings of Corpus Linguistics. University of Birmingham*: 1–26.

Grieve, Jack. 2007. “Quantitative Authorship Attribution: An Evaluation of Techniques.” *Literary and Linguistic Computing* 22(3): 251–70.

Hirst, Graeme, and Vanessa Wei Feng. 2012. “Changes in Style in Authors with Alzheimer’s Disease.” *English Studies* 93(3): 357–70.

Holmes, DavidI., Michael Robertson, and Roxanna Paez. 2001. “Stephen Crane and the New-York Tribune: A Case Study in Traditional and Non-Traditional Authorship Attribution.” *Computers and the Humanities* 35(3): 315–31.

Holmes, Janet. 2008. *An Introduction to Sociolinguistcs*. Fourth. ed. Pearson Longman.

Houvardas, John, and Efstathios Stamatatos. 2006. “N-Gram Feature Selection for Authorship Identification.” *Artificial Intelligence Methodology Systems and Applications* 4183: 77–86.

John, George H., and Pat Langley. 2013. “Estimating Continuous Distributions in Bayesian Classifiers.” : 338–45. http://arxiv.org/abs/1302.4964.

Juola, Patrick. 2007. “Authorship Attribution.” *Foundations and Trends® in Information Retrieval* 1(3): 233–334.

Kaster, Andreas, Stefan Siersdorfer, and Gerhard Weikum. 2005. “Combining Text and Linguistic Document Representations for Authorship Attribution.” *SIGIR Workshop Stylistic Analysis of Text for Information Access STYLE* 1(Pt 1): 27–35.

Kešelj, Vlado, Fuchun Peng, Nick Cercone, and Calvin Thomas. 2003. “N-Gram-Based Author Profiles for Authorship Attribution.” *Pacific Association for Computational Linguistics*: 255–64.

Klaussner, Carmen. 2017. “Elements of Style Change.” University of Dublin.

Labov, W. 1972. *Sociolinguistic Patterns*. University of Pennsylvania Press, Incorporated. https://books.google.com.mx/books?id=hD0PNMu8CfQC.

Lancashire, Ian, and Graeme Hirst. 2009. “Vocabulary Changes in Agatha Christie’s Mysteries as an Indication of Dementia: A Case Study.” *Presented at the 19th Annual Rotman Research Institute Conference, Cognitive Aging: Research and Prac- tice, 8–10 March 2009, Toronto* (March): 1–5.

Le, Xuan. 2010. “Longitudinal Detection of Dementia Through Lexical and Syntactic

Changes in Writing."

Luyckx, Kim, and Walter Daelemans. 2008. "Using Syntactic Features to Predict Author Personality from Text." *Proceedings of Digital Humanities*: 146–49.

Manning, Christopher D., and Prabhakar Raghavan. 2009. "An Introduction to Information Retrieval" ed. A Cannon-Bowers E Salas. *Online*: 1. http://dspace.cusat.ac.in/dspace/handle/123456789/2538.

Markov, Ilia et al. 2017. "CIC-FBK Approach to Native Language Identification."

Mosteller, Frederick, and David L. Wallace. 1966. "Inference and Disputed Authorship: The Federalist." *Journal of the American Statistical Association*, *58*(302), 275-309.

N-Grams:, Syntactic Dependency-Based. 2013. "Syntactic Dependency-Based n-Grams: More Evidence of Usefulness in Classification." *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 7816 LNCS(PART 1): 13–24.

Pennebaker, James W, and Lori D Stone. 2003. "Words of Wisdom: Language Use over the Life Span." *Journal of personality and social psychology* 85(2): 291–301.

Pol, Marta Sanchez. 2005. "A Stylometry-Based Method to Measure Intra- and Inter-Authorial Faithfulness for Forensic Applications."

Posadas-Durán, Juan Pablo, Grigori Sidorov, Ildar Batyrshin, and Elibeth Mirasol-Meléndez. 2015. "Author Verification Using Syntactic N-Grams." *CEUR Workshop Proceedings* 1391(Cic): 8–11.

Rea, Alethea, and William Rea. 2016. "How Many Components Should Be Retained from a Multivariate Time Series PCA?" : 1–49. http://arxiv.org/abs/1610.03588.

Rosenberg, Sheldon, and Leonard Abbeduto. 1987. "Indicators of Linguistic Competence in the Peer Group Conversational Behavior of Mildly Retarded Adults." *Applied Psycholinguistics* 8(01): 19–32.

Rude, Stephanie, Eva-Maria Gortner, and James Pennebaker. 2004. “Language Use of Depressed and Depression-Vulnerable College Students.” *Cognition & Emotion* 18(8): 1121–33.

Sapkota, Upendra et al. 2014. “Cross-Topic Authorship Attribution: Will out-of the Topic Data Help?” *The 25th International Conference on Computational Linguistics (COLING 2014)*.

Sidorov, Grigori. 2013. *Construccion No Lineal de N-Gramas En La Linguistica Computacional*. primera ed. Mexico D.F.: Kronos DIgital S.A. de C.V.

Sidorov, Grigori, Ildar Batyrshin, and Juan Pablo Posadas-Durán. 2014. “Complete Syntactic N-Grams as Style Markers for Authorship Attribution.” *Human-Inspired Computing and Its Applications - 13th Mexican International Conference on Artificial Intelligence, MICAI 2014* (Cic): 9–17.

Sidorov, Grigori, Francisco Velasquez, and Efstathios Stamatatos. 2012. “Syntactic Dependency-Based N-Grams as Classification Features.” 7630(Cic): 1–11.

Spassova, Maria Stefanova. 2009. “El Potencial Discriminatorio de Las Secuencias de Categorías Gramaticales En La Atribución Forense de Autoría de Textos En Español.”

Stamatatos, Efstathios. 2006. “Authorship Attribution Based on Feature Set Subspacing Ensembles.” *International Journal on Artificial Intelligence Tools* 15(05): 823–38.

Stamatatos, Efstathios. 2009. “A Survey of Modern Authorship Attribution Methods.” *Journal of the American Society for Information Science and Technology* 60: 538–56.

Strunk, Jan, and Tibor Kiss. 2006. “Unsupervised Multilingual Sentence Boundary Detection.” *Journal of Computational Linguistics* (1990): 1–38.

Tabata, T. 1994. “‘Dickens’ Narrative Style: A Statistical Approach to Chronological Variation.” *Revue Informatique et Statistique dans les Sciences Humaines* 30: 165–

82.

Turell, Maria Teresa, and Núria Gavaldà. 2013. “Towards an Index of Idiolectal Similitude (or Distance) in Forensic Authorship Analysis.” *Journal of Law and Policy* XXI(2): 495–514.

Turney, Peter D., and Patrick Pantel. 2010. “From Frequency to Meaning: Vector Space Models of Semantics.” *Journal of Artificial Intelligence Research* 37: 141–88.

Uzuner, Özlem, and Boris Katz. 2005. “A Comparative Study of Language Models for Book and Author Recognition.” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 3651 LNAI: 969–80.

Velázquez, Francisco Antonio Castillo. 2014. “Detección de Autoría de Texto Usando Información Sintáctica.” Instituto Politécnico Nacional.

Wilcoxon, Frank. 1945. “Individual Comparisons by Ranking Methods.” *Biometrics Bulletin* 1(6): 80. http://www.jstor.org/stable/10.2307/3001968?origin=crossref.

Williams, Kristine, Frederick Holmes, Susan Kemper, and Janet Marquis. 2003. “Written Language Clues to Cognitive Changes of Aging: An Analysis of the Letters of King James VI/I.” *The journals of gerontology. Series B, Psychological sciences and social sciences* 58(1): P42–44.

Yoon, Su Youn, and Suma Bhat. 2012. “Assessment of ESL Learners’ Syntactic Competence Based on Similarity Measures.” *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.* (pp. 600-608). Association for Computational Linguistics.

Anexo A

Etiquetas POS (Penn Treebank)

CC: coordinating conjunction	**PP$**: possessive pronoun
CD: cardinal number	**RB**: adverb
DT: Determiner	**RBR**: adverb, comparative
EX: Existencial there	**RBS**: adverb, superlative
FW: foreign word	**RP**: particle
IN: preposition/subordinating conjunction	**SYM**: symbol (mathematical or scientific)
JJ: adjective	**TO**: to
JJR: adjective, comparative	**UH**: interjection
JJS: adjective, superlative	**VB**: verb, base form
LS: list item marker	**VBD**: verb, past tense
MD: modal	**VBG**: verb, gerund/present part
NN: noun, singular or mass	**VBN**: verb, past participle
NNS: noun, plural	**VBP**: verb, non-3rd ps. sing. present
NNP: proper noun, singular	**VBZ**: verb, 3rd. ps. sing. present
NNPS: proper noun, plural	**WDT**: wh-determiner
PDT: predeterminer	**WP**: wh-pronoun
POS: possessive ending	**WP$**: possessive wh-pronoun
PRP: personal pronoun	**WRB**: wh-adverb

Anexo B

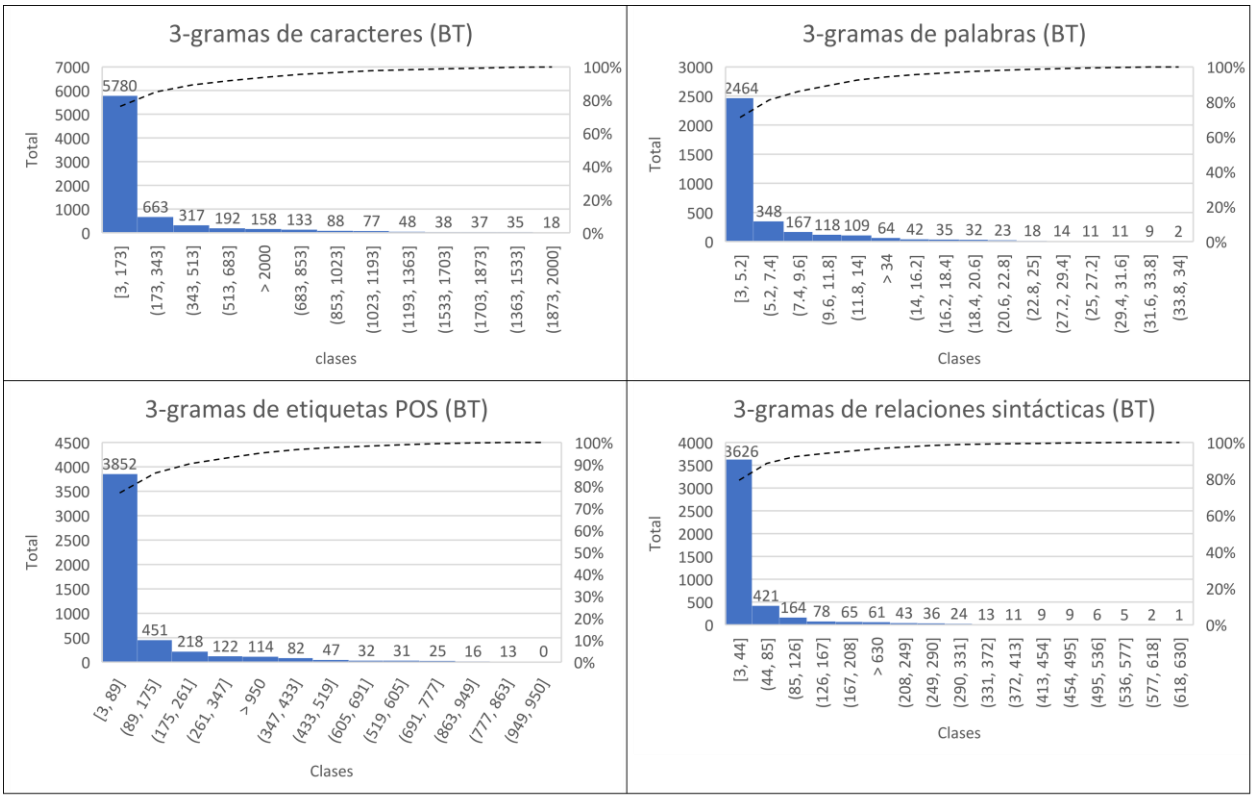
Relaciones gramaticales del Stanford Parser

abbrev: abbreviaton modifier	**nn**: noun compound modifier
acomp: adjectival complement	**nsubj**: nomimal subject
advcl: adverbial clause modifier	**nsubjpass**: passive nominal subject
advmod: adverbial modifier	**num**: numeric modifier
agent: agent	**number**: element of compound number
amod: adjectival modifier	**parataxis**: parataxis
appos: appositional modifier	**partmod**: participial modifier
attr: attributive	**pcomp**: prepositional complement
aux: auxiliary	**pobj**: object of a preposition
auxpass: passive auxiliary	**poss**: possession modifier
cc: coordination	**possessive**: possessive modifier
ccomp: clausal complement	**preconj**: preconjunct
complm: complementizer	**predet**: predeterminer
conj: conjunct	**prep/prepc**: prepositional modifier
cop: copula	**prt**: phrasal verb particle
csubj: clausal subject	**punct**: punctuation
csubjpass: clausal passive subject	**purpcl**: purpose clause modifier
det: determiner	**quantmod**: quantifier phrase modifier
dobj: direct object	**rcmod**: relative clause modifier
expl: expletive	**ref**: referent
infmod: infinitival modifier	**rel**: relative
iobj: indirect object	**tmod**: temporal modifier
mark: marker	**xcomp**: open clausal complement
measure: measure-phrase modifier	**xsubj**: controlling subject
neg: negation modifier	

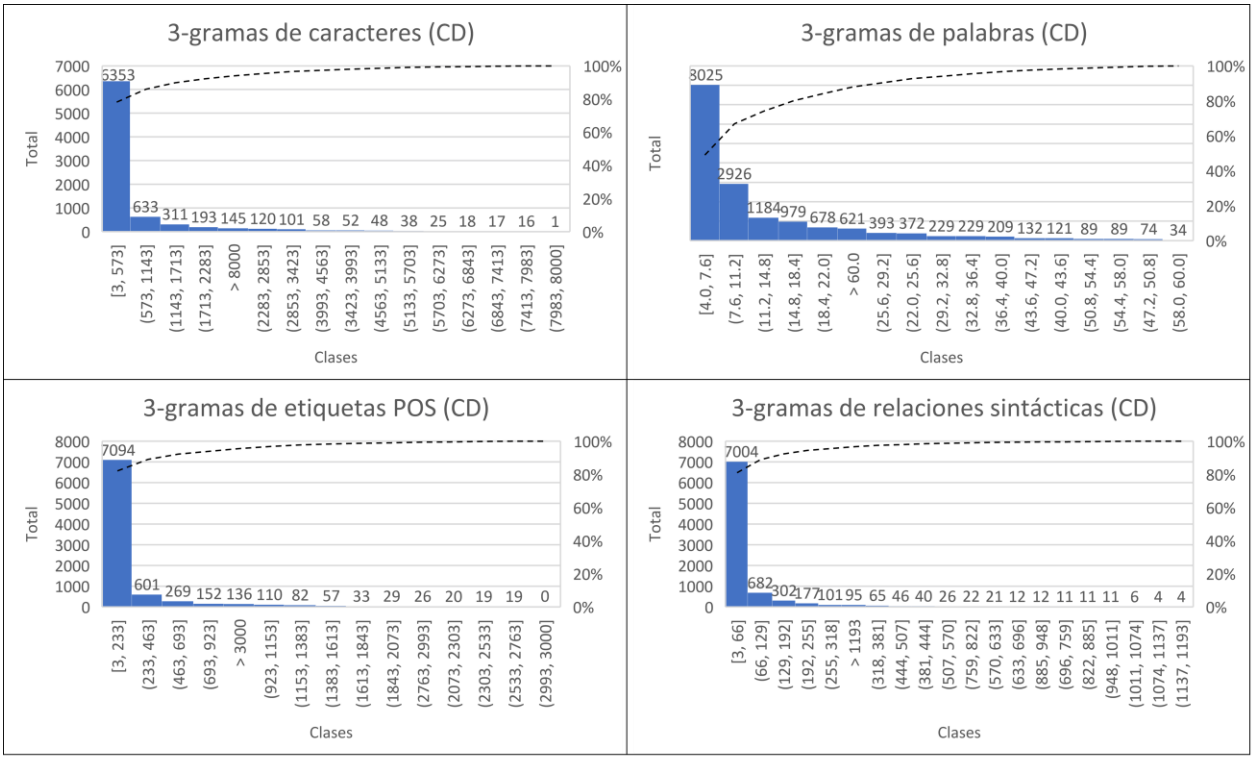
Anexo C

Frecuencias de *3*-gramas utilizando en textos de novelas completas

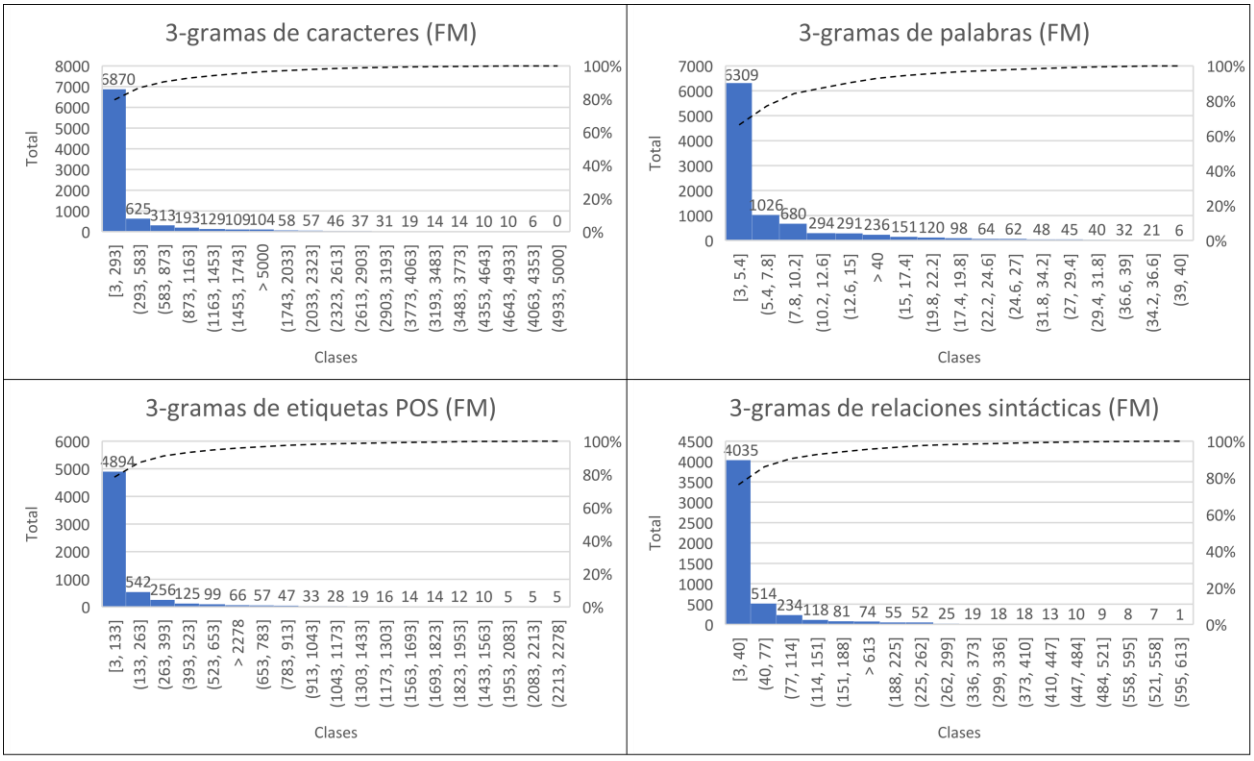
Frecuencias de 3-gramas de Booth Tarkington (BT)



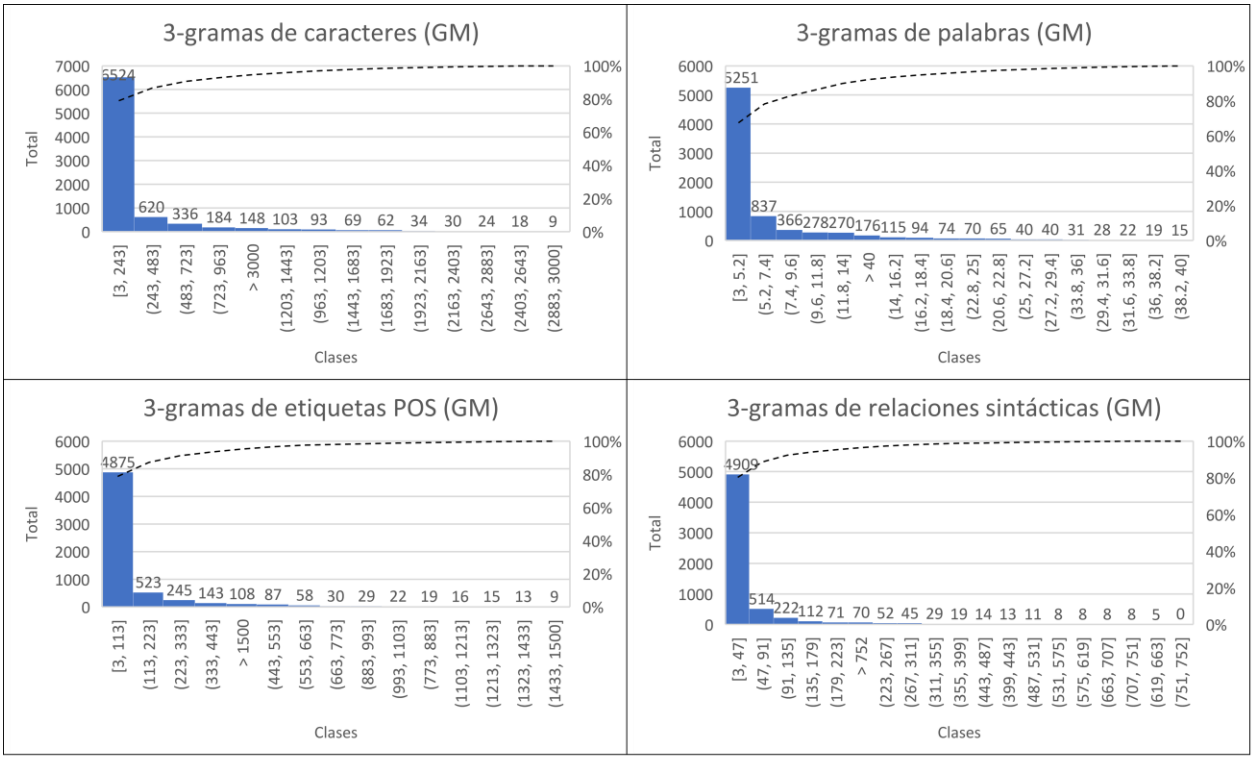
Frecuencias de 3-gramas de Charles Dickens (CD)



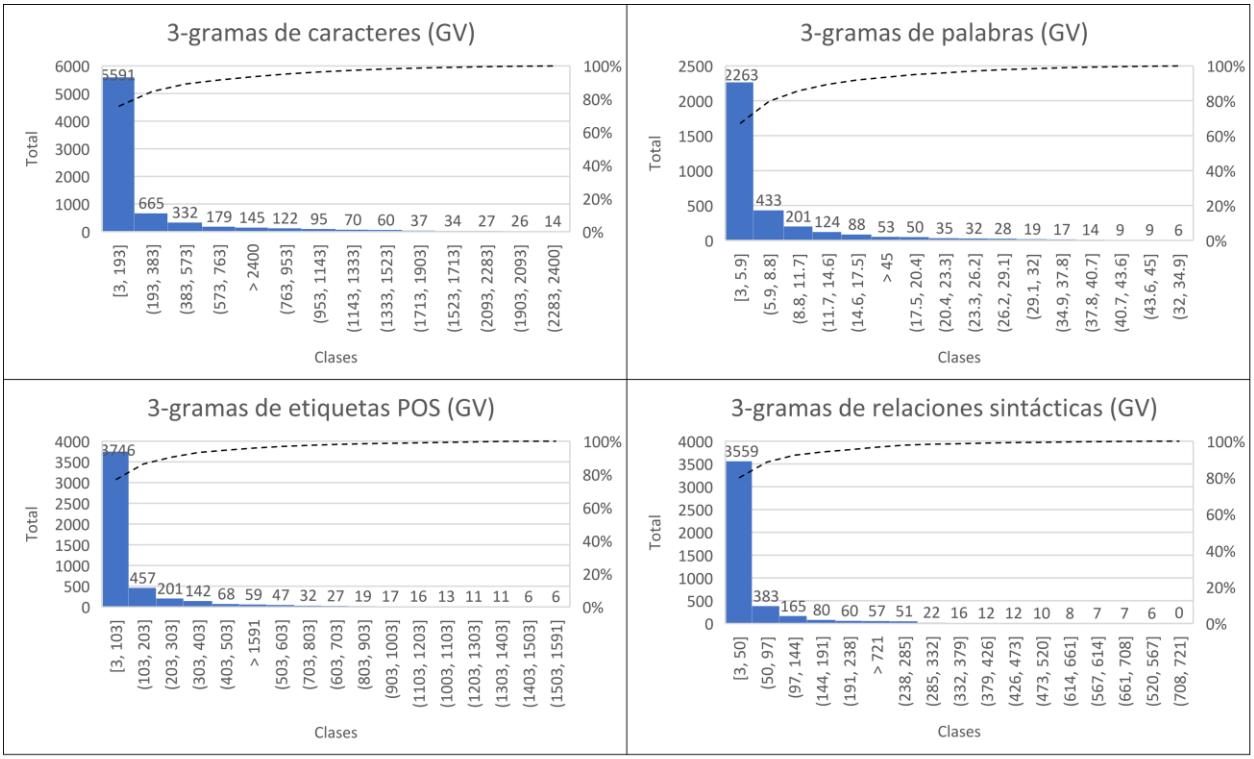
Frecuencias de 3-gramas de Frederick Marryat (FM)



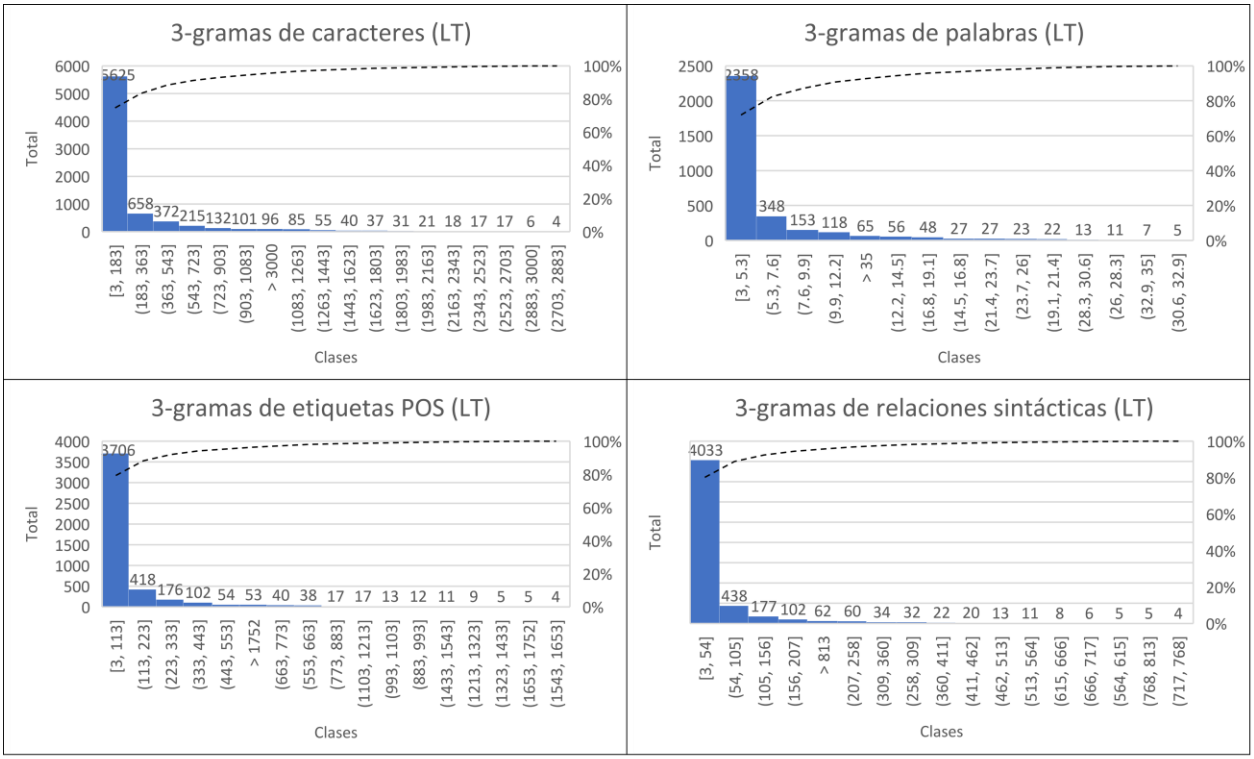
Frecuencias de 3-gramas de George Macdonald (GM)



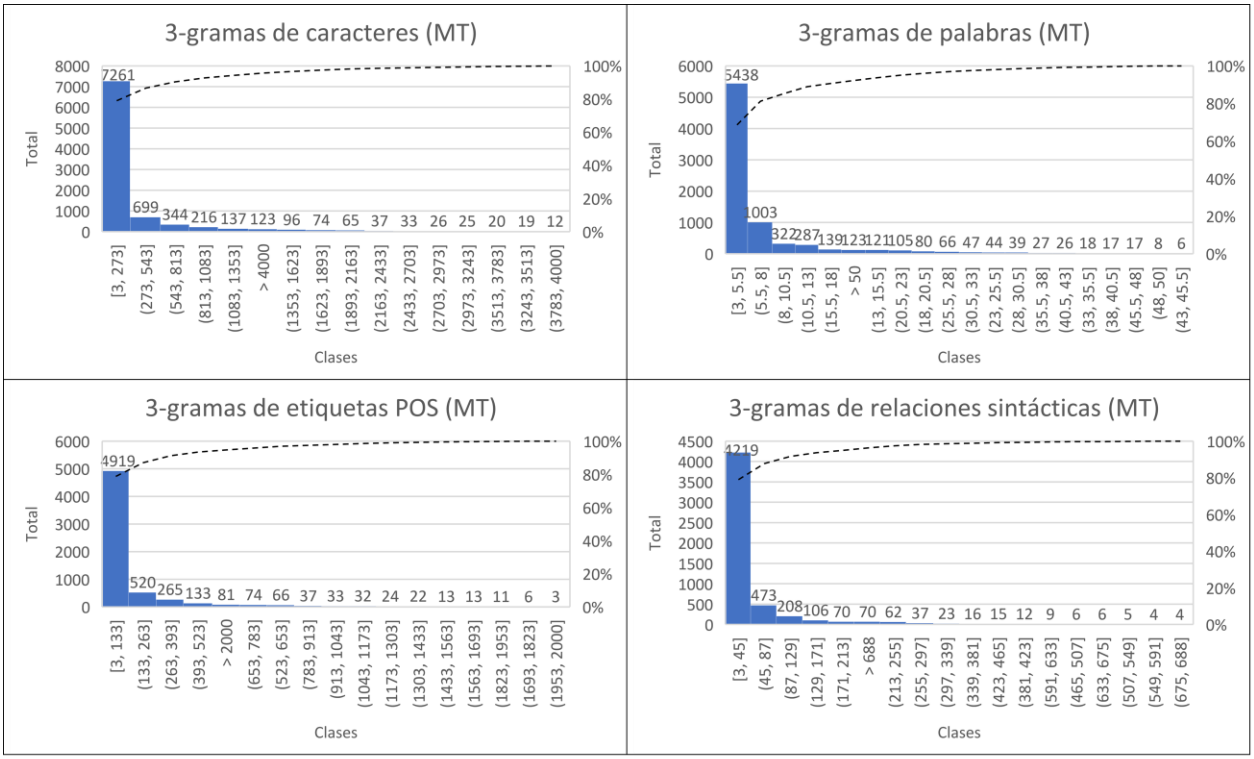
Frecuencias de 3-gramas de George Vaizey (GV)



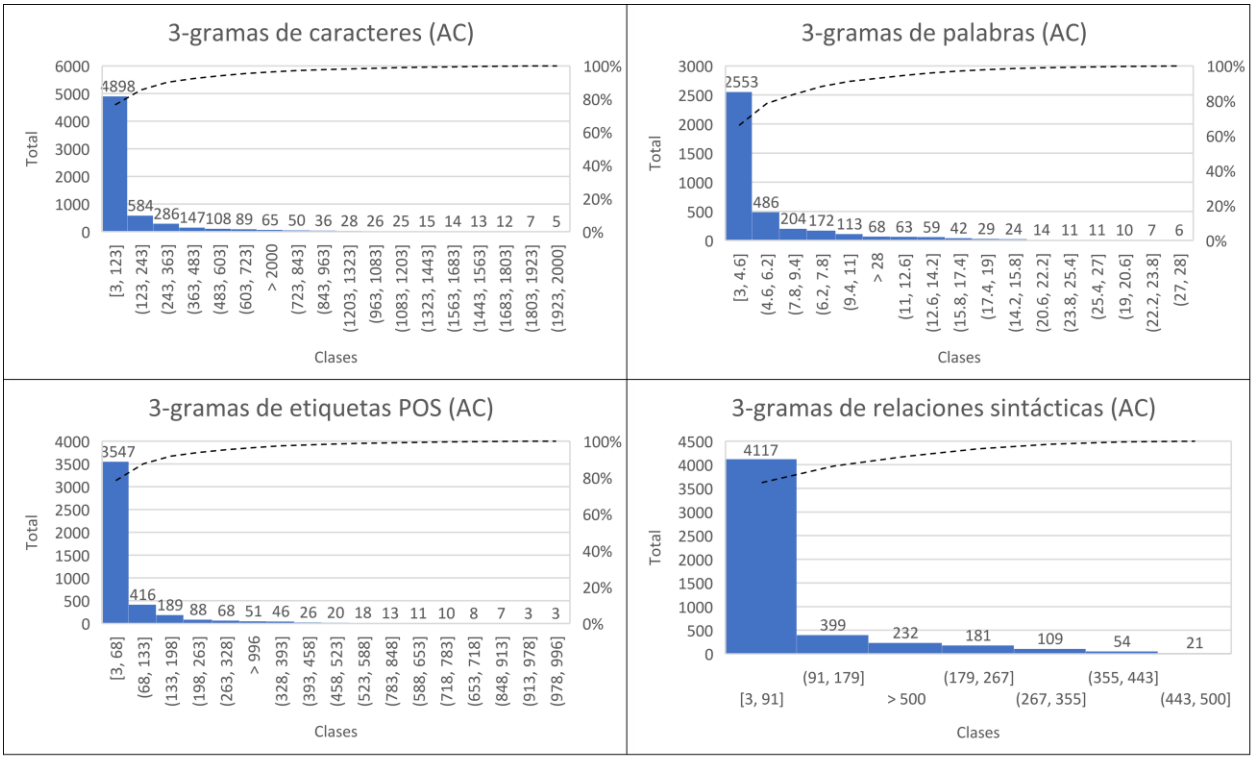
Frecuencias de 3-gramas de Louis Tracy (LT)



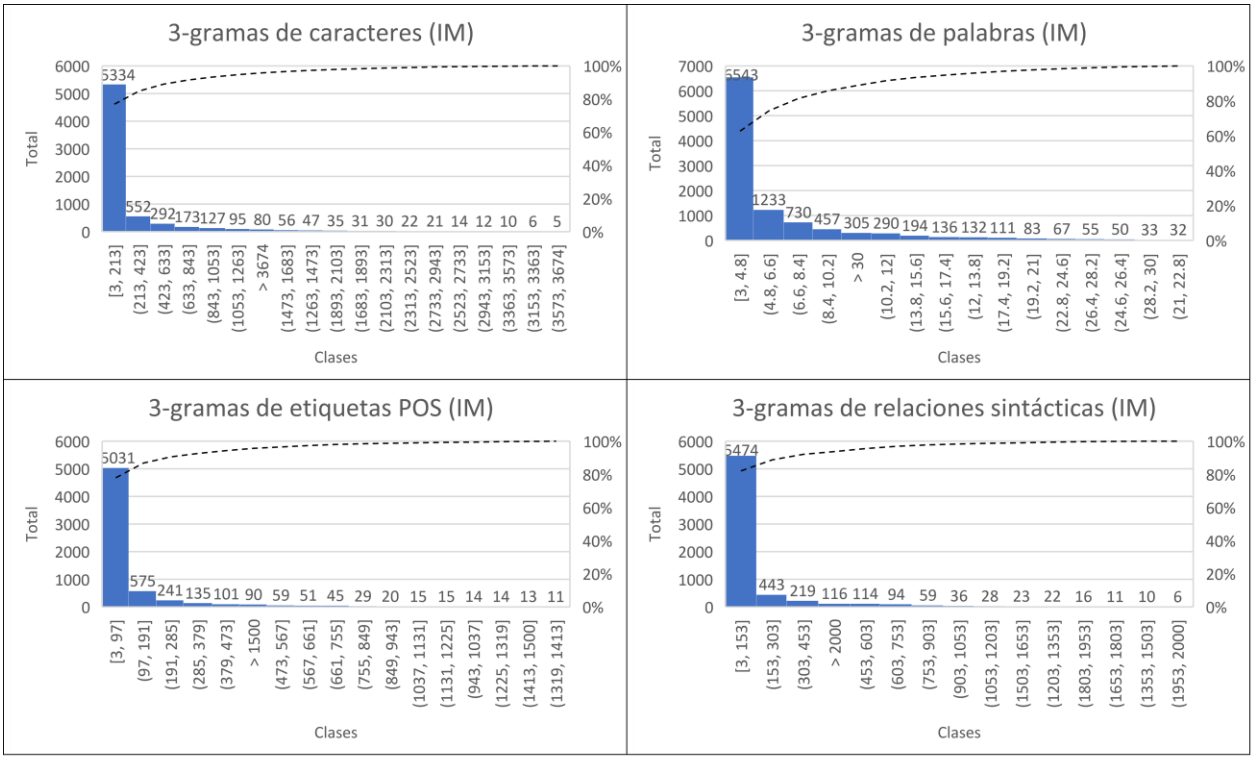
Frecuencias de 3-gramas de Mark Twain (MT)



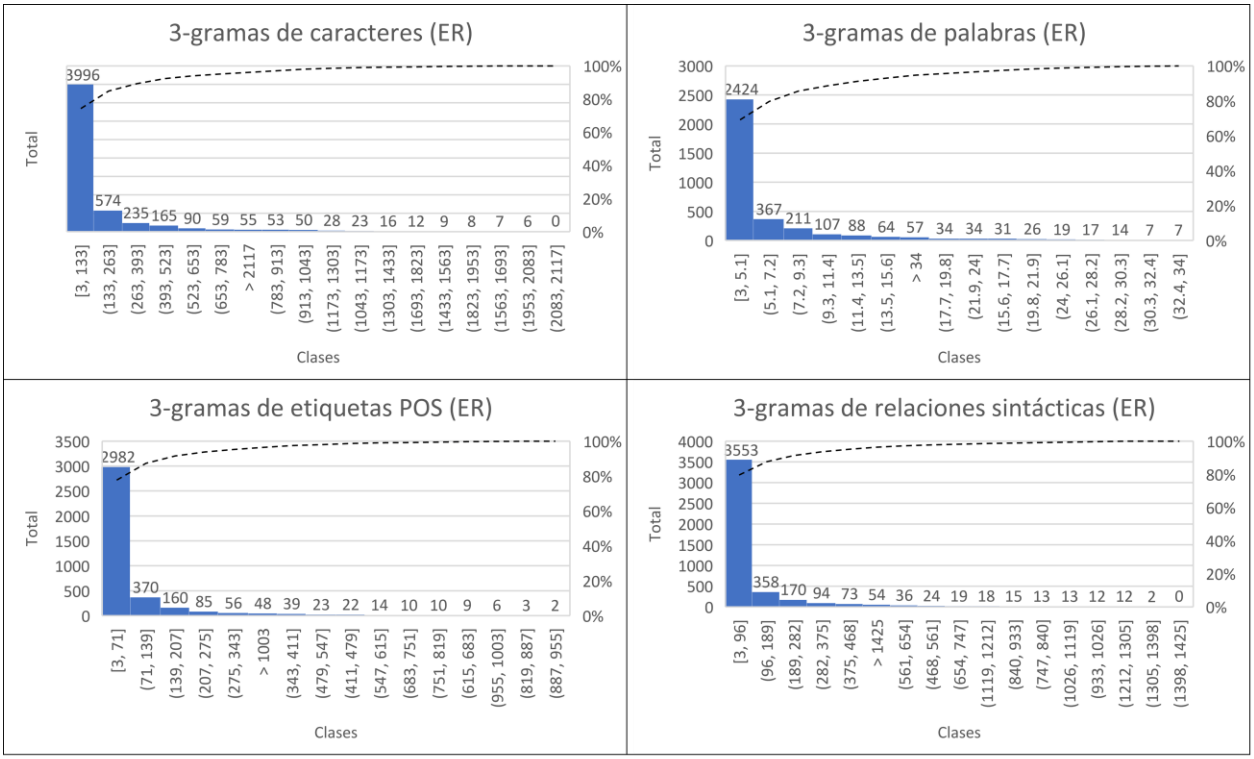
Frecuencias de 3-gramas de Arthur Conan (AC)



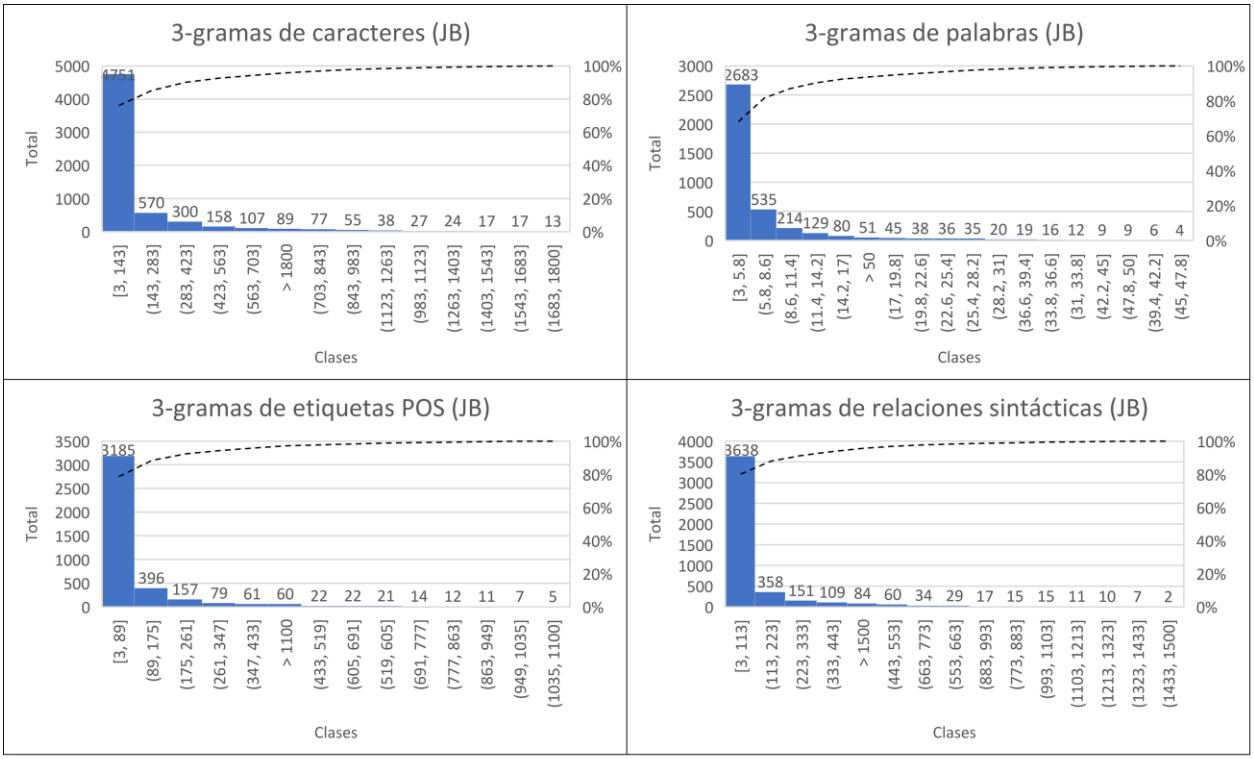
Frecuencias de 3-gramas de Iris Murdoch (IM)



Frecuencias de 3-gramas de Edgar Rice (ER)



Frecuencias de 3-gramas de John Buchan (JB)



Anexo D

Aquí se muestran los resultados de algunos autores en las pruebas de clasificación con 3 etapas. La leyenda T hace referencia al tamaño del texto.

T	EXPERIMENTOS EN 3-GRAMAS DE CARACTER DEL AUTOR CD																											Prom
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	
1	1	0.67	0.67	1	0.67	0.33	1	0.67	0.67	1	0.67	0.67	1	0.67	0.67	0.67	0.67	0.67	0.67	0.33	0.67	0.67	0.33	0.67	0.33	0.33	0.67	0.67
2	0.83	0.67	0.50	1	0.67	0.50	1	0.67	0.67	1	0.67	0.5	1	0.67	0.67	0.83	0.67	0.67	0.67	0.33	0.67	0.67	0.33	0.67	0.50	0.17	0.67	0.66
3	0.78	0.67	0.56	1	0.56	0.56	0.89	0.67	0.78	0.89	0.67	0.44	1	0.67	0.67	0.89	0.56	0.67	0.67	0.33	0.67	0.67	0.33	0.67	0.44	0.22	0.67	0.65
4	0.67	0.67	0.50	1	0.67	0.5	1	0.67	0.75	0.83	0.67	0.42	1	0.67	0.67	0.83	0.50	0.67	0.67	0.33	0.67	0.67	0.33	0.67	0.42	0.08	0.67	0.64

T	EXPERIMENTOS CON 3-GRAMAS POS DE FREDERICK MARRYAT																											Prom
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	
1	0.67	1	0.67	0.67	1	0.67	0.67	0.67	0.67	1	1	1	0.67	1	1	0.67	0.67	0.67	0.67	0.33	1	0.33	1	0.67	0.67	0.67	0.67	0.75
2	0.67	0.83	0.33	0.83	0.83	0.67	0.67	0.67	0.67	1	1	1	0.83	0.83	0.83	0.67	0.67	0.67	1	0.83	1	0.67	0.83	0.83	0.67	0.67	0.67	0.77
3	0.67	0.89	0.33	0.67	0.89	0.56	0.56	0.67	0.67	0.89	0.89	1	0.78	0.78	0.78	0.56	0.67	0.67	0.89	0.78	1	0.67	0.89	0.78	0.56	0.67	0.67	0.73
4	0.67	0.75	0.42	0.67	0.83	0.75	0.58	0.67	0.67	0.92	0.92	1	0.83	0.83	0.83	0.58	0.67	0.67	0.92	0.75	1	0.67	0.75	0.75	0.58	0.67	0.67	0.74

T	EXPERIMENTOS CON 3-GRAMAS SR DE GEORGE MACDONALD																											Prom
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	
1	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	1	1	1	1	1	1	1	1	1	0.78
2	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.83	0.83	0.83	1	1	0.83	1	1	1	0.75
3	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.89	0.89	0.89	0.89	0.89	0.78	1	1	1	0.75
4	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.58	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.83	0.92	0.83	0.92	1	0.92	1	1	1	0.76

T	EXPERIMENTOS CON 3-GRAMAS POS DE GEORGE VAIZEY																										Prom
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	
1	0.67	1	1	1	1	1	0.67	0.67	0.67	1	1	1	0.67	0.67	0.67	1	1	0.67	0.67	1	1	1	1	1	0.67	0.67	0.67	0.85
2	0.67	0.83	0.83	0.67	0.83	1	0.67	0.5	0.83	0.67	0.83	1	0.5	0.5	0.67	0.83	0.67	0.67	0.83	0.67	0.83	0.83	0.83	1	0.83	0.5	0.67	0.75
3	0.67	0.67	0.56	0.78	1	1	0.67	0.56	0.67	0.67	0.78	0.78	0.56	0.67	0.67	0.67	0.78	0.67	0.78	0.78	0.56	0.78	0.89	1	0.78	0.78	0.78	0.74
4	0.67	0.83	0.67	0.75	0.75	0.83	0.67	0.67	0.58	0.67	0.92	0.67	0.67	0.67	0.75	0.58	0.67	0.58	0.83	0.75	0.67	0.83	0.92	0.92	0.75	0.75	0.67	0.73

T	EXPERIMENTOS CON 3-GRAMAS SR DE LOUIS TRACY																										Prom
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	
1	1	1	1	0.67	0.67	0.67	0.33	1	1	0.67	1	1	0.67	0.67	0.67	0.67	1	1	1	1	1	0.67	0.67	0.33	0.67	1	1	0.82
2	0.67	0.83	0.83	0.5	0.5	0.67	0.5	0.67	1	0.67	0.83	0.83	0.5	0.67	0.67	0.67	0.83	1	0.67	0.83	1	0.5	0.67	0.33	0.67	0.83	0.83	0.71
3	0.78	0.89	0.89	0.67	0.56	0.67	0.44	0.67	0.89	0.78	0.89	0.89	0.78	0.78	0.67	0.78	0.89	0.89	0.78	0.89	0.89	0.67	0.67	0.33	0.56	0.78	0.78	0.75
4	0.75	0.75	0.75	0.58	0.5	0.58	0.58	0.75	0.75	0.67	0.83	0.83	0.58	0.67	0.67	0.67	0.83	1	0.75	0.75	0.92	0.58	0.67	0.42	0.75	0.83	0.83	0.71

Anexo E

Este anexo muestra la exactitud obtenida cuando se entrena con textos de un autor y se prueba con textos de un autor diferente. Estos resultados corresponden al Grupo 1. Observe que, en la mayor parte de los experimentos, la exactitud apenas supera el 33%, la línea base para 3 etapas.

Entrenamiento con textos de Booth Tarkington (BT)																
Autores	car				pal				pos				sr			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
CD	0.33	0.33	0.33	0.36	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.37	0.42
FM	0.33	0.33	0.33	0.33	0.33	0.33	0.37	0.39	0.33	0.33	0.33	0.33	0.33	0.39	0.37	0.36
GM	0.22	0.33	0.30	0.28	0.33	0.44	0.52	0.44	0.44	0.33	0.26	0.31	0.22	0.22	0.22	0.25
GV	0.33	0.22	0.26	0.28	0.22	0.22	0.15	0.25	0.11	0.17	0.19	0.08	0.33	0.33	0.30	0.33
LT	0.33	0.33	0.33	0.33	0	0.22	0.26	0.42	0.33	0.33	0.33	0.39	0.22	0.28	0.26	0.25
MT	0.33	0.33	0.33	0.33	0.33	0.50	0.44	0.53	0.22	0.33	0.33	0.31	0.22	0.44	0.41	0.36

Entrenamiento con textos de Charles Dickens (CD)																
Autores	car				pal				pos				sr			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
BT	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33
FM	0.33	0.33	0.33	0.36	0.56	0.56	0.56	0.53	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33
GM	0.33	0.33	0.33	0.33	0.33	0.44	0.52	0.50	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33
GV	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33
LT	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.36	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33
MT	0.33	0.39	0.33	0.36	0.33	0.33	0.30	0.42	0.33	0.39	0.37	0.36	0.33	0.33	0.33	0.33

Entrenamiento con textos de Frederick Marryat (FM)																
Autores	car				pal				pos				sr			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
BT	0.33	0.33	0.33	0.33	0.56	0.39	0.44	0.36	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.36
CD	0.33	0.33	0.33	0.33	0.33	0.33	0.26	0.33	0.44	0.33	0.37	0.39	0.33	0.33	0.33	0.33
GM	0.56	0.56	0.56	0.56	0.56	0.44	0.56	0.42	0.67	0.72	0.63	0.69	0.56	0.56	0.56	0.5
GV	0.17	0.17	0.19	0.28	0.11	0.17	0.22	0.33	0.33	0.22	0.3	0.28	0.22	0.28	0.30	0.25
LT	0.33	0.33	0.33	0.33	0.44	0.33	0.48	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.41	0.36
MT	0.33	0.28	0.26	0.28	0.22	0.22	0.26	0.33	0.44	0.33	0.3	0.28	0.33	0.22	0.22	0.28

Entrenamiento con textos de George Macdonald (GM)																
Autores	car				pal				pos				sr			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
BT	0.33	0.33	0.33	0.33	0.33	0.33	0.3	0.28	0.44	0.44	0.44	0.42	0.22	0.22	0.26	0.22
CD	0.67	0.5	0.56	0.56	0.56	0.56	0.56	0.5	0.33	0.33	0.33	0.33	0.43	0.28	0.22	0.31
FM	0.44	0.44	0.44	0.44	0.44	0.44	0.44	0.44	0.44	0.33	0.44	0.47	0.22	0.33	0.26	0.31
GV	0.33	0.33	0.33	0.33	0.22	0.28	0.22	0.39	0.22	0.28	0.3	0.22	0.33	0.33	0.33	0.33
LT	0.33	0.28	0.3	0.31	0.33	0.33	0.33	0.5	0.33	0.33	0.33	0.31	0.56	0.61	0.59	0.58
MT	0.44	0.56	0.44	0.5	0.44	0.39	0.44	0.42	0.56	0.67	0.52	0.58	0.44	0.44	0.48	0.44

Entrenamiento con textos de George Vaizey (GV)																
Autores	car				pal				pos				sr			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
BT	0.22	0.22	0.3	0.28	0.33	0.33	0.3	0.33	0	0.06	0.15	0.17	0.33	0.33	0.33	0.33
CD	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.22	0.22	0.22	0.25
FM	0.33	0.33	0.30	0.33	0.22	0.28	0.37	0.42	0.22	0.28	0.26	0.25	0.33	0.33	0.33	0.25
GM	0	0	0	0.08	0.22	0.06	0.11	0.17	0.11	0.11	0.07	0.06	0.11	0.11	0.11	0.11
LT	0.22	0.22	0.26	0.22	0.33	0.33	0.39	0.22	0.33	0.33	0.26	0.22	0.22	0.22	0.22	0.19
MT	0.44	0.39	0.41	0.36	0.22	0.17	0.19	0.25	0.33	0.39	0.33	0.31	0.22	0.28	0.3	0.28

Entrenamiento con textos de Louis Tracy (LT)																
Autores	car				pal				pos				sr			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
BT	0.44	0.44	0.3	0.31	0.44	0.39	0.41	0.53	0.44	0.33	0.37	0.36	0.33	0.33	0.33	0.33
CD	0.33	0.33	0.33	0.33	0.33	0.33	0.26	0.33	0.44	0.61	0.33	0.33	0.33	0.28	0.26	0.33
FM	0.33	0.33	0.33	0.36	0.44	0.44	0.41	0.36	0.44	0.44	0.41	0.47	0.33	0.5	0.59	0.53
GM	0.67	0.44	0.56	0.42	0.22	0.39	0.41	0.36	0.56	0.61	0.56	0.5	0.67	0.61	0.59	0.58
GV	0.22	0.17	0.07	0.11	0.11	0.17	0.37	0.28	0.22	0.22	0.22	0.14	0.11	0.11	0.11	0.19
MT	0.22	0.22	0.26	0.22	0.33	0.17	0.3	0.36	0.11	0.22	0.26	0.28	0.33	0.33	0.3	0.33

Entrenamiento con textos de Mark Twain (MT)																
Autores	car				pal				pos				sr			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
BT	0.67	0.67	0.56	0.58	0.56	0.44	0.52	0.56	0.33	0.33	0.33	0.42	0.44	0.39	0.3	0.36
CD	0.33	0.33	0.33	0.33	0.44	0.44	0.41	0.5	0.44	0.39	0.48	0.56	0	0.06	0.15	0.22
FM	0.11	0.22	0.19	0.22	0.33	0.28	0.22	0.22	0	0.06	0.04	0.17	0.11	0.11	0.15	0.11
GM	0.67	0.67	0.67	0.67	0.44	0.56	0.56	0.5	0.67	0.61	0.59	0.64	0.44	0.5	0.48	0.42
GV	0.33	0.33	0.33	0.33	0.11	0.28	0.15	0.25	0.22	0.11	0.15	0.14	0.22	0.33	0.33	0.28
LT	0.44	0.44	0.44	0.42	0.33	0.44	0.41	0.36	0.33	0.28	0.3	0.36	0.33	0.33	0.37	0.31