Clasificación Automática de Textos considerando el Estilo de Redacción

Por

ROSA MARIA COYOTL MORALES

Tesis sometida como requisito parcial para obtener el grado de

Maestra en Ciencias *en la especialidad de* ***Ciencias Computacionales***

en el

Instituto Nacional de Astrofísica, Óptica y Electrónica.
INAOE

Supervisada por:

DR. LUIS VILLASEÑOR PINEDA
Coordinación de Ciencias Computacionales, INAOE

DR. MANUEL MONTES Y GÓMEZ
Coordinación de Ciencias Computacionales, INAOE

Tonantzintla, Pue.

2007

A mi mamá y papá, Ana Teresa y Clemente,

por todo su cariño, motivación, comprensión y apoyo.

"Gracias por creer en mi".

A mis hermanos, Martin y Diana

por todo su cariño, motivación y apoyo.

A mis cuñandos, Laura y Marco

por su cariño y apoyo.

Agradecimientos

A mis asesores Dr. Luis Villaseñor Pineda y Dr. Manuel Montes y Gómez mi más sincero agradecimiento por su apoyo constate, sus comentarios acertados y sus consejos que me acompañaron a lo largo de mis estudios de maestría en el INAOEP.

A mis sinodales, Dr. Aurelio López López, Dr. Jesús Ariel Carrasco Ochoa y Dra. Angélica Muñoz Meléndez por sus observaciones y comentarios.

Al INAOE, por todas las facilidades proporcionadas durante mi estancia académica.

A mis compañeros de la maestría por su amistad y por darme tantos momentos de alegría.

A CONACYT por el apoyo económico a través de la beca No. 189686.

Resumen

En la actualidad existe una inmensa cantidad de información disponible en formato electrónico. Toda esta información es improductiva si no se dispone con mecanismos apropiados para su acceso, clasificación y análisis. En particular, la *clasificación automática de textos* consiste en colocar un documento dentro de un grupo de clases previamente definidas. La mayor parte del trabajo en esta área se ha enfocado en la clasificación de textos por su tema o tópico. Sin embargo, un documento también puede ser clasificado de acuerdo a su *estilo* (clasificación no-temática). En la clasificación no-temática se consideran tareas tales como la clasificación de opiniones, la detección de plagio, la atribución de autoría, la clasificación por género, etc. El objetivo principal de esta tesis es proponer métodos que permitan determinar los rasgos léxicos que hacen posible caracterizar el estilo de escritura de los documentos. Los métodos descritos consideran la caracterización de los documentos a través un conjunto de secuencias de palabras que combinan tanto palabras de contenido como funcionales. La utilidad de este tipo de caracterización se demuestra mediante su aplicación en las tareas de *atribución de autoría* y *clasificación por género*.

Abstract

Nowadays there is a large amount of information available in digital format. All this information is useless if we do not have adequate mechanisms for its access, classification and analysis. In particular, *text classification* concerns the automatic assignment of free text documents to one or more predefined categories. Most work in this field focuses on categorizing documents by their topic. However, a document can be also classified by its *written style* (non-topic classification). Basically, non-topic classification considers tasks such as sentiment classification, plagiarism detection, authorship attribution, genre classification, etc. The main objective of this thesis is to propose methods for determining the lexical features that allow characterizing the written style of documents. The proposed methods consider the characterization of documents by sets of word sequences that combine content and functional words. The usefulness of this kind of characterization is demonstrated by its application in the tasks of *authorship attribution* and *genre classification*.

Contenido

Resumen ..... **i**

Abstract ..... **iii**

Contenido ..... **v**

Índice de Tablas ..... **vii**

Índice de Figuras ..... **ix**

Capítulo 1 ..... **1**

Introducción **1**
1.1 Motivación 1
1.2 Descripción del Problema 2
1.3 Objetivos de la Tesis 3
1.4 Estructura de la Tesis 4

Capítulo 2 ..... **5**

Antecedentes **5**
2.1 Clasificación Automática de Textos 5
2.1.1 Representación de los Documentos 7
2.1.2 Reducción de Dimensionalidad 9
2.1.3 Algoritmo de Aprendizaje 10
2.1.4 Medidas de Evaluación 12
2.2 Clasificación por Estilo 14
2.2.1 Atribución de Autoría 16
2.2.2 Clasificación por Género 18
2.3 Trabajo Relacionado 19
2.3.1 Caracterización Estilométrica 19
2.3.2 Caracterización Sintáctica 20
2.3.3 Caracterización Léxica 20

Capítulo 3 23

Corpus y Resultados de Referencia 23
3.1 Corpus 24
3.2 Resultados de Referencia 27
3.2.1 Atribución de Autoría 27
3.2.2 Clasificación por Género 30

Capítulo 4 35

Métodos Propuestos 35
4.1 Secuencias Frecuentes Maximales 36
4.2 Método Básico 37
4.2.1 Resultados del Método Básico 39
4.3 Método Iterativo 40
4.3.1 Resultados del Método Iterativo 43

Capítulo 5 49

Conclusiones y Trabajo Futuro 49
5.1 Conclusiones 49
5.2 Trabajo Futuro 50

Bibliografía 55

Publicaciones 53

Apéndice A 61

Resultados al aplicar los método propuesto en la clasificación tematica 61
A.1 Resultados de Referencia 62
A.2 Resultados con los métodos propuestos 64

Índice de Tablas

Tabla 2-1 Ejemplo 13
Tabla 3-1 Corpus poetas 24
Tabla 3-2 Ejemplos de poemas con y sin presencia de anacronismos 25
Tabla 3-3 Fragmentos de dos poemas de nuestro corpus 25
Tabla 3-4 Corpus género 26
Tabla 3-5 Etiquetas de los signos de puntuación 26
Tabla 3-6 Método estilométrico 27
Tabla 3-7 Método bolsa de palabras 28
Tabla 3-8 Método bolsa de palabras y palabras funcionales 29
Tabla 3-9 Método uni-gramas y bi-gramas 29
Tabla 3-10 Método uni-grams, bi-gramas y tri-gramas 30
Tabla 3-11 Método estilométrico 30
Tabla 3-12 Método bolsa de palabras 31
Tabla 3-13 Método bolsa de palabras y palabras funcionales 31
Tabla 3-14 Método uni-gramas y bi-gramas 32
Tabla 3-15 Método uni-gramas, bi-gramas y tri-gramas 32
Tabla 4-1 Resultados del método básico en el corpus poetas 39
Tabla 4-2 Resultados método básico en el corpus género 40

Tabla 4-3 Construcción del conjunto de características en el corpus poetas.....43
Tabla 4-4 Ejemplos de SFM con mayor GI para el corpus de poetas 44
Tabla 4-5 Resultados del método iterativo en el corpus poetas 45
Tabla 4-6 Construcción del conjunto de características en el corpus género.....45
Tabla 4-7 Resultados del método iterativo en el corpus género.....46
Tabla 4-8 Resutados experimentales: Corpus poetas 46
Tabla 4-9 Resultados experimentales: Corpus género 47
Tabla A-1 Corpus desastres…….……61
Tabla A-2 Método estilométrico..……....62
Tabla A-3 Método bolsa de palabras……...62
Tabla A-4 Método de palabras y palabras funcionales…..……...63
Tabla A-5 Método uni-gramas y bi-gramas……..……...63
Tabla A-6 Método uni-gramas, bi-gramas y tri-gramas……..63
Tabla A-7 Resultado del método básico …….64
Tabla A-8 Construcción de conjunto de características...…….....65
Tabla A-9 Resultados del método iterativo…..……...65
Tabla A-10 Resultados experimentales: Corpus desastres……...……...66

Índice de Figuras

Figura 2-1 Un paradigma de aprendizaje. 6
Figura 2-2 Ejemplos de características estilométricas 19
Figura 4-1 Proceso de caracterización básica y clasificación 37
Figura 4-2 Proceso de caracterización iterativa y clasificación 41

Capítulo 1

Introducción

1.1 Motivación

Un hecho bien conocido es que existe una enorme cantidad de documentos en línea gracias a las posibilidades de la Web. Este hecho ha motivado investigaciones diversas alrededor de este gran cúmulo de información. La Recuperación de Información [Henzinger, M., 2000], la Minería de Texto [Hernández J. et al., 2004], la Extracción de Información [Téllez A., 2005], la Búsqueda de Respuestas [Vicedo J. et al., 2003] son algunas de tantas de estas líneas de investigación. Una de estas líneas es la llamada Clasificación o Categorización de Textos [Sebastiani F., 2005]. En ella, una máquina determina la categoría de un texto, de entre varias posibles, de acuerdo a ciertas características presentes en dicho texto. En particular, la Categorización de Textos ha sido ampliamente estudiada y ha alcanzado resultados sorprendentes al distinguir el tema o tópico [Diederich J. et al., 2003; Joachims T., 1998; Kaster A. et al., 2005; Sebastiani F., 1999] En este caso, se desea determinar a partir del texto, sin ningún tipo de anotación, la categoría de dicho texto. Por ejemplo, en el caso de una nota periodística se desea determinar si ésta pertenece a la sección de política, negocios o deportes de la edición diaria de un periódico. Decimos que es sorprendente pues los métodos actuales no basan esta categorización en el

"entendimiento" del texto, es decir, no se busca llevar a una representación semántica el documento. Estos métodos determinan la categoría del documento únicamente con base en sus palabras, sus repeticiones y las combinaciones entre ellas [Diederich J. et al., 2003; Fürnkranz J., 1998; Peng F. et al., 2004, Sebastiani F., 1999].

Por otro lado, es claro que la información presente en un texto es de muy variada naturaleza. Un texto conlleva otro tipo de información no-temática: su estilo. Esta información puede brindarnos elementos para determinar su autor, para categorizar el texto por su género literario, para determinar el nivel del escritor[1] o para determinar el tipo de lector esperado de dicho documento, entre muchas otras

De esta manera, existen dos grandes tipos de categorización de textos: temática, interesada en el que; y la no-temática, interesada en el cómo fue escrito un texto.

El presente trabajo se orienta a la clasificación no-temática, y para ello busca entre los elementos léxicos de los documentos aquellos atributos que conduzcan a una adecuada identificación de los documentos.

En la siguiente sección se describe y acota el problema abordado por esta tesis. Después, en la sección 1.2 se presentarán los objetivos de la tesis, por último en la sección 1.3 se expondrá brevemente la organización de la tesis.

1.2 Descripción del Problema

La falta de herramientas que se encaminen a identificar aquellos documentos que son más convenientes para un usuario y el gran incremento en la cantidad de información, dificultan la búsqueda de material adecuado. Una manera de mejorar la búsqueda consiste en considerar el perfil del usuario. Un analista y un estudiante de educación media, a pesar de investigar el mismo tópico, buscan documentos distintos. De ahí la importancia de clasificar los documentos no sólo por su contenido temático sino por su estilo de redacción. Por supuesto, los estilos de redacción son muchos y

[1] Es decir, el grado de educación que tiene: no-escolarizado, escolarizado, especialista, etc.

variados de ahí la enorme tipología que existe al respecto. El presente trabajo no pretende abarcar toda esta problemática y se limita a dos casos específicos: la clasificación por género y la atribución de autoría.

Ahora bien, la clasificación por género se puede definir como "un agrupamiento de documentos que son estilísticamente consistentes" [Finn A. & Kushmerick N., 2003]. Por otro lado, la atribución de autoría es la tarea de identificar el autor de un texto dado [Sebastiani F., 2005], es decir, se pretende caracterizar el estilo de redacción de un autor. De manera concreta este trabajo pretende encontrar las características estilísticas que permiten la clasificación por género o por autor. Como es de imaginar, muy diferentes esquemas se han planteado para esta caracterización [Argamon S. & Sterling S., 2003; Argamon S. & Levitan., 2005; Chaski C., 2005; Corney M. et al., 2002; de Vel O. et al., 2001; Diederich J. et al., 2003; Finn A. & Kushmerick N., 2003; Fürnkranz J., 1998; Kaster, A. et al., 2005; Keselj V. et al., 2003; Luyckx K. et al., 2004; Malyutov M., 2004; Peng F. et al., 2004; Stamatatos E. et al., 2000; Stamatatos E. et al., 2001; Zhao Y. & Zobel J., 2005]. Es interés de esta tesis limitar la naturaleza de estas características al nivel léxico, ya que no se desea depender de costosas herramientas lingüísticas, alcanzado un método lo más general posible.

1.3 Objetivos de la Tesis

Objetivo General

Proponer un método que permita encontrar los rasgos que caracterizan la escritura de un texto.

Objetivos Específicos

- Abordar las problemáticas de identificación del autor e identificación del género de un texto

- Proponer un método general limitándonos al nivel léxico de los documentos, es decir, mantener el método lo más independiente posible del uso de costosas herramientas lingüísticas.

1.4 Estructura de la Tesis

En el capítulo 2 se presenta los conceptos básicos relacionados al contenido de la tesis, los cuales incluyen nociones sobre la clasificación de textos, así como una breve discusión sobre los conceptos de estilo, género y atribución de autoría. Por otro lado, se presenta una revisión del trabajo relacionado más relevante y reciente que se ha realizado con respecto a la clasificación por estilo. Se analizan las características propuestas por los distintos trabajos.

En el capítulo 3 se presentan los datos experimentales recopilados para el desarrollo de esta tesis, así como los resultados obtenidos con otros métodos. Estos resultados son usados para realizar el análisis comparativo del método propuesto.

En el capítulo 4 se presenta una primera versión del método propuesto, se explican las ideas centrales y se presentan sus resultados experimentales al aplicarlo tanto en clasificación por género como en atribución de autoría.

En el capítulo 5 se presenta una versión revisada del método propuesto y se muestran y discuten los resultados.

Finalmente, en el capítulo 6 se ofrecen las conclusiones y el trabajo futuro que se desprende del presente trabajo.

Capítulo 2

Antecedentes

En este capítulo introducimos los conceptos básicos cuyo conocimiento resulta imprescindible para abordar los siguientes capítulos. En la sección 2.1 se describen las nociones básicas de la clasificación automática de textos. En la sección 2.2 se describe la tarea de clasificación de textos por estilo, centrándose en la clasificación por género y en la atribución de autoría. Y finalmente en la sección 2.3 se presenta un estudio del trabajo relacionado más relevante.

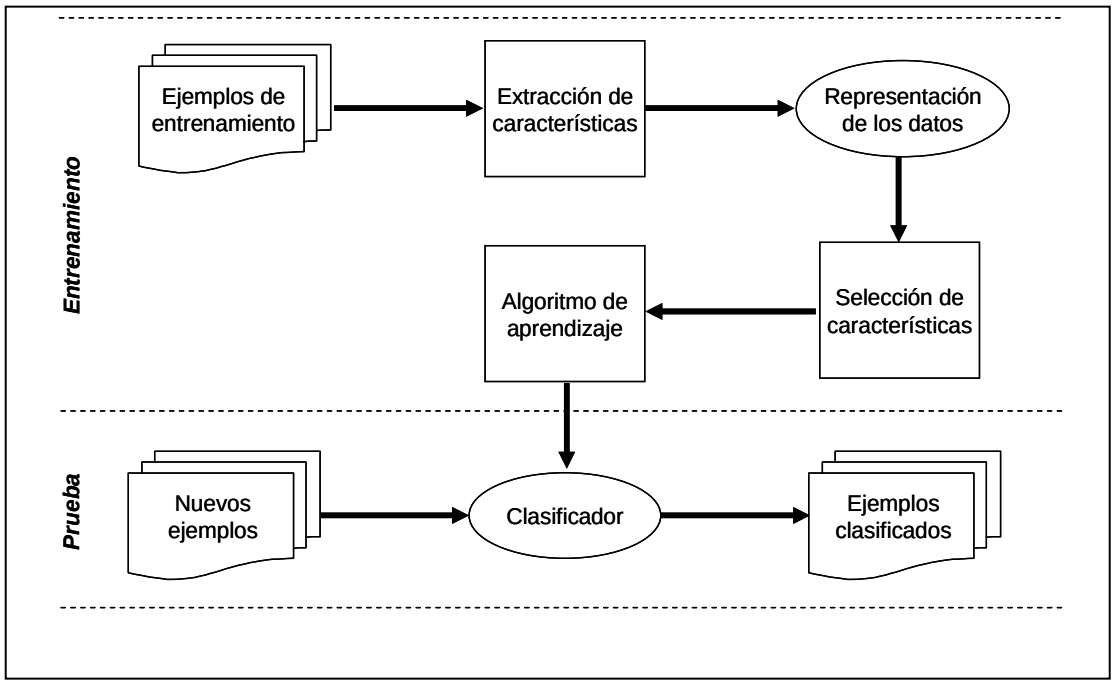
2.1 Clasificación Automática de Textos

La clasificación automática de textos tiene sus orígenes en la Recuperación de información (*Information Retrieval*) y últimamente ha recibido más atención debido al incremento en la cantidad de información disponible en formato electrónico. Es por esta razón, que cada vez es mayor la necesidad de herramientas que ayuden a satisfacer las necesidades del usuario en cuanto a la información que busca, y además encontrar ésta en un tiempo adecuado.

El objetivo de la clasificación automática de texto es categorizar documentos dentro de un número fijo de categorías predefinidas en función de su contenido. Un mismo documento puede pertenecer a una, varias, todas o ninguna de las categorías

dadas [Joachims T., 1998]. Cuando se utiliza aprendizaje automático, el objetivo es aprender a clasificar a partir de ejemplos que permitan hacer la asignación a la categoría automáticamente.

Figura 2-1 Un paradigma de aprendizaje[2].



En la Figura 2-1 podemos observar el paradigma de aprendizaje inductivo, el cual intenta aprender conceptos a través de ejemplos de éstos. El clasificador construido utiliza los conceptos aprendidos para clasificar nuevos ejemplos. El aprender de ejemplos es un problema conocido como aprendizaje supervisado, ya que parte de una serie de clases o categorías diseñadas *a priori*, en las cuales hay que distribuir a cada uno de los documentos.

En resumen, la construcción de un clasificador automático de texto comienza con la recopilación y clasificación manual de un conjunto de documentos (documentos de entrenamiento), después se llevan los documentos a una representación adecuada para que finalmente se puedan aplicar distintos algoritmos

[2] http://www.exa.unicen.edu.ar/catedras/ayrdatos/slides/clasificacion1p.pdf

de clasificación y así obtener el clasificador. En las siguientes secciones se describen las técnicas utilizadas en este trabajo de tesis para la construcción del clasificador.

2.1.1 Representación de los Documentos

Para llevar a cabo la clasificación automática de texto se tiene que representar cada documento de los ejemplos de entrenamiento, de manera que a esa representación se le pueda aplicar el algoritmo de clasificación. La representación más utilizada es el *modelo vectorial*, ésta es manejada ampliamente por los sistemas de recuperación de información.

Este modelo consiste en representar la colección de documentos como una matriz de palabras o términos por documentos [Ass K. & Eikvil L., 1999]. Es decir, cada texto o documento d_j es representado por medio de un vector $\vec{d}_j = \left(w_{1j},...,w_{|\tau|j}\right)$ de términos w, donde τ es el conjunto de palabras del vocabulario de la colección y w_{ij} representa un valor numérico que expresa en qué grado el documento d_j posee el término t_i. Frecuentemente, el conjunto τ es el resultado de filtrar las palabras del vocabulario con respecto a una lista de palabras vacías, éstas son palabras frecuentes que no contienen información semántica (de ahí el nombre de palabras vacías). Ejemplos de palabras vacías (también llamadas palabras funcionales) son las preposiciones, conjunciones, artículos, etc. Otra estrategia es el uso de un lematizador el cual tiene como objetivo eliminar afijos de una palabra de tal manera que aparezca sólo su raíz léxica[4]. Esto se realiza con la finalidad de que las palabras que tienen el mismo significado conceptual sean representadas por su raíz léxica, por ejemplo, caminar, caminará, caminó, caminando se representa por *camin*. Con respecto al peso del término w_{ij} se tiene distintas

[4] Otra posibilidad es el uso de un truncador como el propuesto por [Porter, 1980]

maneras de calcularlo. A continuaron se da una breve descripción de tres tipos de pesado.

Ponderado booleano: Asigna el peso de 1 si la palabra t_i ocurre en el documento d_j y 0 en caso contrario.

$$W_{ij} = \begin{cases} 1 & si \quad t_i \quad aparece \quad en \quad d_j \\ 0 & en \quad otro \quad caso \end{cases}$$

Ponderado por frecuencia de término: Asigna el número de veces que el término i ocurre en el documento d_j, denotado como f_{ij}.

$$W_{ij} = f_{ij}$$

Este cálculo se debe a que si un término aparece muchas veces en un documento, se supone que es importante en ese documento.

Ponderado tf-idf: Asigna el peso de la palabra i en el documento j en proporción a el número de ocurrencias de la palabra en el documento y en proporción inversa al número de documentos en la colección para los cuales ocurre la palabra al menos una vez.

$$W_{ij} = f_{ij} * \log\left(\frac{N}{n_i}\right)$$

Donde N es el número de documentos en la colección y n_i es el número de documentos en los que el término i aparece.

En el presente trabajo sólo se reportan los experimentos utilizando el ponderado booleano. Cabe mencionar que se realizaron experimentos con las otras dos medidas, sin embargo, estas últimas no obtuvieron resultados sobresalientes en comparación con la representación booleana.

2.1.2 Reducción de Dimensionalidad

El modelo vectorial lleva a espacios de términos de alta dimensión al representar el conjunto de entrenamiento, por lo tanto existe la necesidad de reducir el conjunto original de características o términos [Ass K. & Eikvil L., 1999]. Para llevar a cabo esta reducción de dimensionalidad se hace una selección de un subconjunto de características, con la finalidad de encontrar los términos con mayor poder discriminante. Existen varias técnicas para reducir la dimensionalidad de la matriz de características, por ejemplo ganancia de información (*information gain*). Esta técnica [Yang Y. & Pedersen J., 1997] consiste en medir el número de bits de información obtenida para predecir la categoría por medio de la presencia o ausencia de una palabra en el documento. La definición formal de ganancia de información es la siguiente:

Dado $c_1,...,c_k$ el conjunto de posibles clases. La ganancia de información de una palabra o término t_i es definida como:

$$IG(t_i) = -\sum_{k=1}^{M} P(c_k)\log P(c_k) + P(t_i)\sum_{k=1}^{M} P(c_k|t_i)\log P(c_k|t_i) + P(\bar{t_i})\sum_{k=1}^{M} P(c_k|\bar{t_i})\log P(c_k|\bar{t_i})$$

donde M es el número de clases, $P(c_k)$ es la probabilidad de la clase c_k, $P(t_i)$ es la probabilidad de seleccionar un documento que contienen el término t_i, $P(c_k|t_i)$ es la probabilidad condicional de que un documento con el término t_i pertenezca a la

categoría c_k, $P(\overline{t_i})$ es la probabilidad de seleccionar un documento que no contienen el término t_i y finalmente $P(c_k|\overline{t_i})$ es la probabilidad condicional de que un documento con el término t_i no pertenezca a la categoría c_k .

A partir del cálculo de la ganancia de información de cada término es posible identificar aquellos términos con mayor poder discriminativo. Usualmente se seleccionan aquellos términos que sobrepasan un cierto umbral. De manera particular en nuestros experimentos usamos todos los términos con IG > 0.

2.1.3 Algoritmo de Aprendizaje

Como se había mencionado antes, en la clasificación automática de textos se parte de una serie de clases o categorías prediseñadas, en las cuales hay que colocar cada uno de los documentos. El proceso de calcular patrones en base a los documentos preclasificados se conoce como entrenamiento. Uno de los algoritmos más utilizados para el cálculo de estos patrones es el denominado Naive Bayes. Este clasificador es de tipo probabilístico, el cual se basa en el cálculo de distribuciones de probabilidad en función de datos observados. Enseguida se describirá el algoritmo Naive Bayes de una manera formal.

Dado un documento d y un conjunto de clases predefinidas $\{c_1, c_2, ..., c_k\}$, el clasificador Naive Bayes primero calcula la probabilidad a posteriori de que el documento pertenezca a cada clase particular c_k, es decir, $P(c_k|d)$ y entonces asigna el documento a la clase o clases con las probabilidades más altas. La probabilidad a posteriori es calculada aplicando el teorema de Bayes:

$$P(c_k|d) = \frac{P(d|c_k)P(c_k)}{P(d)} \quad (1)$$

El denominador $P(d)$ en la formula (1) es independiente de las clases; por lo tanto, puede ser ignorado. Por ende:

$$P(c_k|d) = P(d|c_k)P(c_k) \quad (2)$$

En Naive Bayes, se asume independencia de ocurrencia entre los l términos del vocabulario c_k, es decir, para $d = t_{1,} t_2,...,t_l$ de ahí que podemos calcular $P(c_k|d)$ de la siguiente forma:

$$P(c_k|d) = P(c)\prod_{i=1}^{l} P(t_i|c_k) \quad (3)$$

Así la fórmula 3 puede ser expresada como sigue:

$$P(c_k|d) = \arg\max_{c \in C} \prod_{i=1}^{l} P(t_i|c_k)P(c_k) \quad (4)$$

En general, $P(c_k|d)$ representa la probabilidad condicional de que dada una nueva instancia se asigne la clase o categoría con mayor probabilidad, dados los atributos o características que describen la instancia.

En la fórmula 4, $P(c_k)$ puede ser estimado a partir de los ejemplos pertenecientes a la clase c_k:

$$P(c_k) = \frac{N_k}{N} \quad (5)$$

Donde N es el número total de documentos de entrenamiento y N_k es el número de documentos en la clase c_k. Por su parte $P(t_i|c_k)$ es usualmente calculado de la siguiente manera:

$$P(t_i|c_k) = \frac{1 + count(t_i, c_k)}{m + N_k} \quad (6)$$

donde $count(t_i, c_k)$ es el número de veces que la palabra t_i ocurre dentro de los documentos de entrenamiento de la clase c_k, y m es el tamaño del vocabulario. Esta estimación usa Laplace (agrega uno) para resolver el problema de la probabilidad cero. Con esta estimación lo que se pretende es que todas las configuraciones posibles incluidas las no vistas, tengan una probabilidad asociada, ya que la configuración que no esté en la colección de datos tendrá una probabilidad cero.

En conclusión, la tarea de aprendizaje en el clasificador Naive Bayes consiste en construir una hipótesis por medio de estimar las probabilidades $P(c_k)$ y $P(t_i|c_k)$ en términos de los ejemplos de entrenamiento pertenecientes a la clase c_k.

2.1.4 Medidas de Evaluación

Para evaluar un sistema de clasificación de texto se utilizan las medidas de precisión y recuerdo[5] (*precision and recall*), que son medidas comunes en el área de recuperación de información. La precisión es la probabilidad de que un documento clasificado en la clase "c_i" corresponda realmente a esa clase. El recuerdo es la probabilidad de que un documento que pertenece a la clase "c_i" es clasificado dentro de esa clase [Hernández J. et al., 2004; Lewis D., 1991]. Así, la precisión se puede ver como una medida de la corrección del sistema, mientras que el recuerdo da una medida de cobertura o completitud.

Para calcular estas medidas en un conjunto de prueba, se considera el problema de forma binaria. En este contexto, la siguiente tabla resume el comportamiento de un sistema según los casos de aciertos y errores:

[5] No existe un común acuerdo entre los diferentes autores sobre este término en español. Entre las posibles acepciones encontramos: alcance, cobertura, evocación y recubrimiento.

Tabla 2-1 Ejemplo

	predicción positiva	**predicción negativa**	**total de predicciones**
Clase positiva	*a*	*b*	*a*+*b*
Clase negativa	*c*	*d*	*c*+*d*
	a + *c*	*b*+*d*	*a*+*b*+*c*+*d* =*n*

En la Tabla 2-1, cada celda representa el número de predicciones positivas y negativas. Así, $a + d$ son los aciertos del sistema y $c + b$ son los errores, y la suma de las cuatro celdas ($a + b + c + d$) equivale al número total de predicciones binarias. Los valores de esta tabla permiten estimar las medidas de precisión y recuerdo según las siguientes expresiones:

$$precisión = \frac{a}{a+c}$$

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

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

El recuerdo refleja cuantos de todos los documentos de una clase son clasificados en ella.

Describir el comportamiento de un clasificador de textos con dos medidas no es práctico para comparar sistemas. Para ello es común utilizar la medida F_β que se define como:

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

para $\beta = 1$, es la medida armónica de la precisión y el recuerdo. En la medida, β es un parámetro que controla la importancia relativa entre las dos medidas. Es común usar el valor 1, que da igual importancia a las dos medidas.

Otra medida que es empleada en este trabajo de tesis es la exactitud, la cual representa el porcentaje de las predicciones que son correctas.

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

Existen medidas que toman en cuenta la precisión y recuerdo de la colección completa. Estas son las siguientes

Micropromedio *(Microaverage):* consiste en calcular la efectividad considerando el conjunto completo de predicciones *n* como un sólo grupo de muestras.

Macropromedio *(Macroaverage):* consiste en calcular un promedio de efectividad considerando cada clase como un grupo de muestra distinto.

2.2 Clasificación por Estilo

Cada vez un mayor número de investigadores procedentes de diferentes campos, sobre todo informáticos y documentalistas, se están interesando en el desarrollo de herramientas y métodos para la clasificación automatizada de textos. La clasificación automática de textos es el área que se encarga de ubicar cada documento a la clase que pertenece. Tradicionalmente, la clasificación automática de textos se ha enfocado en categorizar documentos en función de su contenido. Como se vio en la sección anterior, el primer paso para construir un sistema automático de clasificación es determinar las clases a las que puede pertenecer un texto, es decir, la caracterización deseada. Por supuesto, esta caracterización está en función del problema a resolver. Un ejemplo de esto serían las secciones de un periódico, donde las notas periodísticas se agrupan en función de su contenido: política, negocios, deportes, etc. Lo que es más, dentro de la sección de política podrían existir subsecciones, una de política nacional y otra de internacional. Así, dependiendo del

problema es necesario determinar la tipología de los textos. Será posteriormente a la definición de la tipología de textos, que se planteará el problema de qué criterios seguir para obtener la caracterización más apropiada que ayude a identificar la clase de un texto. Así, el primer paso de nuestra tarea es determinar las clases posibles para categorizar un documento por su estilo. Para lograr esto necesitamos responder qué se entiende por estilo de un texto y especificar los estilos de texto de nuestro interés. Los párrafos subsecuentes discuten estos conceptos.

El concepto de *estilo* no es fácil de definir, es un término ambiguo, asociado a múltiples nociones, utilizado en diferentes disciplinas, y además con una fuerte carga coloquial. El estilo es la manera en que una persona actúa, lo cual marca la acción en sí misma con una inscripción única [Bruce D., 2000]. Para un experto, su estilo es una forma natural de ejecutar una actividad cualquiera. Por ejemplo, un escritor después de un periodo de duro trabajo y experimentación, desarrolla su propio estilo. En general, cada persona adquiere su propio estilo a lo largo de su vida, de acuerdo a las experiencias que ha tenido en su vida laboral, intelectual y personal.

En nuestro caso, limitándonos al estilo de un texto y dado que está fuera del alcance de esta tesis definir formalmente el concepto de estilo, adoptaremos la siguiente definición: *"estilo es el conjunto de distintos aspectos o rasgos que caracterizan la escritura de un texto"* [Alcaraz E. & Martínez A., 1997]. En nuestro caso, estos rasgos son los que identifican a un documento como elemento de una clase. Desafortunadamente, estos rasgos no son claros y en algunos casos desconocidos (i.e. ¿cuáles son los rasgos que identifican a los textos escritos por Octavio Paz?). Precisamente este es el problema que aborda esta tesis al proponer un método que identifica rasgos orientados a determinar el estilo de un texto. En este trabajo, únicamente abordaremos dos tareas de categorización por estilo y partiremos de ellas en la búsqueda de un método de caracterización para la categorización por estilo.

Estas dos tareas son la clasificación por género y la atribución de autoría. A continuación se describen los detalles de estas dos tareas.

2.2.1 Atribución de Autoría

Como se mencionó en párrafos anteriores, el estilo no es sólo un ingrediente en una pieza de escritura, sino un reflejo que proviene desde el autor. Es decir, es dependiente de cómo el autor transfiere una idea a lenguaje escrito [Bruce D., 2000]. En consecuencia, el escrito final está fuertemente relacionado al autor. En otras palabras, el estilo es indicado por características que revelan la elección del autor de un modo de expresión, es decir, la elección específica de palabras, estructuras sintácticas, estrategias del discurso o combinaciones. Además, existen variaciones que influyen en el texto tales como la educación, el estatus social, la personalidad del autor, la audiencia y la época en que fue escrito el documento. Todas estas variaciones son independientes de la temática del texto, pero son determinantes para la correcta categorización de dicho texto.

Como se expuso, existen varios factores que influyen para clasificar correctamente un texto y más aún, hay una infinidad de rasgos que caracterizan el estilo de escritura de un autor y que son difíciles de identificar, pues "el estilo de un autor se refiere a las particulares condiciones de apropiación y actualización de los enunciados"[6]. Sin embargo, éste puede ser variable debido a la diferencia en temas o género y también al desarrollo de cada autor a través del tiempo. La principal tarea en la atribución de autoría es identificar las características, las cuales deben ser invariantes y éstas deben ayudar a discriminar a un autor de otros. En contraste con las tareas de clasificación por contenido, en este caso no es claro cómo determinar el conjunto de características que deben ser utilizadas para identificar un autor. Así, el desafío principal de esta tarea es la caracterización apropiada de los documentos que capture el estilo de escritura de los autores.

[6] http://dghispanos.org/blog/archives/115/

En particular para el caso de obras literarias los autores intentan recorrer y transcribir la vivencia íntima de un ser humano. Así por ejemplo, un escritor desea acceder al lector, por eso persigue la comunicación, el placer estético y la emoción mediante la poesía. El poeta transmite todo esto por medio de palabras, de frases y de la composición conjunta. En fin, lo que busca es ofrecer una poesía que se integre en el ser humano: en su memoria, por el metro y la rima; en su inteligencia, por las informaciones transmitidas y en su corazón, por el ritmo, la fuerza expresiva que estimula y estremece. Así todas estás características que imprime el escritor o escritores en sus poemas a través de las palabras, de los signos de puntuación, la composición entre estos, etc.; es el estilo propio de redacción que tiene cada autor.

En general, el estilo de un escritor no es exclusivamente su sensibilidad literaria, su capacidad creativa para escribir novelas, cuentos o hacer poesías. Todas las personas, en cualquier redacción que realicen, tienen un estilo propio. Se podría definir como un equilibrio entre el orden y el movimiento. Por una parte el estilo depende de la organización y jerarquía de las ideas en el texto, de la coherencia que se desprende del mismo; pero también de la capacidad de interesar, agradar y dar vida a las ideas, dar un ritmo adecuado al tipo de mensaje.

En resumen la atribución de autoría es un problema donde un conjunto de documentos con autores conocidos es utilizado para entrenamiento de modelos, para posteriormente determinar automáticamente el autor de un texto anónimo. De esta manera, sólo capturamos rasgos relevantes de un autor que permiten identificarlo contra un conjunto de autores determinado.

Existen trabajos que tratan de identificar el estilo de un autor en diferentes contextos. Por ejemplo, en la clasificación de correos electrónicos [de Vel O. et al., 2001; Argamon S. et al., 2003], en la detección de plagio [Zhao, Y. et al., 2005] o, en el análisis forense de un texto, que intenta determinar el autor en relación a una investigación criminal [de Vel O. et al., 2001].

2.2.2 Clasificación por Género

Determinar una taxonomía de género de textos es un proceso subjetivo, por lo tanto, existen infinidad de propuestas. Así la gente puede estar en acuerdo o desacuerdo sobre las características o atributos comunes que constituyen un género [Finn A. & Kushmerick N., 2003; Diederich J. et al., 2003]. Identificar, por lo tanto, las características que conforman un género no es una tarea resuelta.

Por otro lado, dado que nuestro interés es enfocarnos en el estilo de redacción dejaremos atrás todas aquellas tipificaciones que consideran el contenido del texto (p.e. ciencia-ficción, policiales, etc.) para centrarnos en aquellas más cercanas a la forma de expresión (p.e. cuento, novela, poesía, etc.). En este último caso, el género de un documento es considerado ortogonal al contenido temático de éste. Ya que los documentos que tratan el mismo tema pueden ser de diferentes géneros, y de manera similar, los documentos que son del mismo género pueden ser de diferentes temas. Por ejemplo, un artículo divulgativo y uno especializado, a pesar del tratar sobre el mismo tema, no pertenecen al mismo género, porque el especializado será más formal y planteará directamente la información que se desea transmitir, mientras que el divulgativo entrará menos en el detalle y empleará recursos explicativos y expresiones menos técnicas para que pueda ser comprendido por los lectores.

En conclusión, el término género es ampliamente utilizado para referirse a diferentes dimensiones de un texto. En nuestro caso particular, género es una abstracción basada en un agrupamiento natural de documentos escritos en un estilo similar independiente del tema tratado.

En la siguiente sección se presentan los trabajos relacionados con esta problemática.

2.3 Trabajo Relacionado

En esta sección se presentan los trabajos centrales que se han realizado con la temática de clasificación de textos por estilo. Este panorama general se orienta a los métodos de caracterización que se han propuesto en la literatura, principalmente los utilizados en atribución de autoría y clasificación por género.

2.3.1 Caracterización Estilométrica

Los primeros intentos en la caracterización de documentos por estilo provienen de esfuerzos de análisis literarios. Básicamente éstos se enfocaron exclusivamente en el uso de medidas estilométricas. Características como la longitud de las palabras o de las oraciones, así como la riqueza del vocabulario han sido algunas de las medidas utilizadas [Corney M. et al., 2002; de Vel O. et al., 2001]. A pesar de que intuitivamente la amplitud del vocabulario y la frecuencia de uso de tales palabras parecerían ser elementos básicos característicos de cada autor, este tipo de características no son suficientes. Al parecer esto se debe a que, por un lado, existen importantes variaciones aún para el mismo autor dependiendo el tipo de texto; y por otro lado, estas características son en extremo sensibles al tamaño del documento, perdiendo gran parte de su significado para textos pequeños. Ejemplos de estas medidas pueden observarse en la Figura 2-2.

Figura 2-2 Ejemplos de características estilométricas

promedio = número de palabras / número de oraciones
riqueza del vocabulario = número de palabras / total del vocabulario
hapax = número de palabras cuya ocurrencia en el documento es uno / total de vocabulario
riqueza de palabras = número de oraciones / número de palabras
riqueza de oraciones = número de oraciones / total del vocabulario
palabras en mayúsculas = (palabras que comienzan con mayúsculas – número de oraciones) / número de oraciones
promedio de las palabras = total de caracteres / número de palabras

2.3.2 Caracterización Sintáctica

Otro intento es la caracterización de los textos por un conjunto de marcadores de estilo (*style markers*[7]). Estos marcadores de estilo van más allá de las simples medidas estilométricas sobre las palabras e integran información sobre la estructura del lenguaje empleado. Para ello, es necesario realizar un análisis complejo usando analizadores morfológicos y sintácticos (i.e. taggers, parsers) [Chaski C., 2005; Finn A. & Kushmerick N., 2003; Luyckx K. et al., 2004; Malyutov M., 2004; Stamatatos E. et al., 2001]. Así un texto se caracteriza por la presencia y frecuencia de ciertas estructuras sintácticas. Desafortunadamente, esta caracterización es costosa y en algunos casos imposible dada la inexistencia de tales herramientas para el idioma en cuestión. Inclusive hay que recordar que dichas herramientas no son del todo confiables dado que no son capaces de tratar cualquier frase del idioma, introduciendo errores en el análisis.

Cabe recordar, que nuestro trabajo busca una manera lo más general posible de caracterizar un texto por su estilo, el uso de marcadores sintácticos es demasiado específico quedando fuera del interés de esta tesis.

2.3.3 Caracterización Léxica

Este enfoque incluye por lo menos tres métodos diferentes. En el primero, la caracterización se realiza usando exclusivamente un conjunto de palabras funcionales, ignorando las palabras de contenido [Argamon S. & Sterling S., 2003; Argamon S. & Levitan., 2005; Stamatatos E. et al., 2000; Stamatatos E. et al., 2001; Zhao Y. & Zobel J., 2005]. Por palabras funcionales entendemos aquellas palabras que sin tener un significado propio son utilizadas como conectores para integrar un mensaje.

[7] Por ejemplo, el uso del impersonal para caracterizar documentos técnicos.
[9] Son los términos más frecuentes en una colección de documentos que incluyen palabras vacías y algunas veces tanto signos de puntuación como palabras con una frecuencia alta.

Ejemplos de estas palabras son las preposiciones, los artículos, las conjunciones, etc. A pesar de que las palabras funcionales no parecen ser marcas de estilo confiables, ya que son muy frecuentes y ocurren en todo texto, el uso y frecuencia de estas palabras es característico del estilo de los autores. La clasificación basada en este tipo de caracterización trabaja apropiadamente pero es muy sensible al tamaño de los documentos. En este caso, la longitud de los documentos no sólo influye en la frecuencia de ocurrencia de las palabras funcionales sino también en su posible presencia.

Un segundo método usa la representación tradicional de bolsa de palabras[9] considerando únicamente las palabras de contenido [Diederich J. et al., 2003; Finn A. & Kushmerick N., 2003; Kaster, A. et al., 2005], es decir, el enfoque tradicional usado para la clasificación temática. Este método produce resultados aceptables siempre y cuando exista una fuerte correlación entre los temas y los autores.

Finalmente, un tercer método considera *n*-gramas, es decir, secuencias de *n* palabras sucesivas. Este método intenta capturar la estructura del lenguaje de los textos por medio de simples secuencias de palabras en contraste de las complejas estructuras sintácticas [Fürnkranz J., 1998; Keselj V. et al., 2003; Peng F. et al., 2004;]. De esta manera, el propósito es obtener una adecuada caracterización de los textos sin llevar a cabo un costoso análisis sintáctico. Desafortunadamente este tipo de caracterización nos lleva a una explosión combinatoria, por lo que comúnmente se emplean secuencias de una, dos o a lo más tres palabras (uni-gramas, bi-gramas y tri-gramas). En conclusión, los n-gramas permiten un mejor tratamiento de las frases como “Bill Gates” o “White House”, pues la representación anterior (*bolsa de palabras*) separaba estas palabras y su estructura se pierde.

El enfoque propuesto en este trabajo se ubica en este último esquema, limitando la representación de los documentos a su nivel léxico. Esto se debe a que se trata de una caracterización simple pero general, la cual puede aplicarse a cualquier tipo de texto, sin importar su dominio e incluso el idioma. El principal problema de esta caracterización es que cada documento será representado por un enorme número de atributos. Como podrá verse en el próximo capítulo, mientras más atributos se

tienen el rendimiento del clasificador disminuye. Así, el presente trabajo se aboca a encontrar los atributos a nivel léxico más representativos.

Capítulo 3

Corpus y Resultados de Referencia

En esta sección se describen los corpus utilizados así como los resultados experimentales obtenidos de los métodos de referencia.

Para alcanzar una adecuada evaluación comparativa a nivel de la caracterización se fijaron ciertas condiciones en todos los experimentos: a) se utilizó un ponderado booleano, pues con base en experimentos anteriores llegamos a concluir que este tipo de representación es el más adecuado; b) se aplicó el método de ganancia de información con umbral mayor a cero (IG>0) para reducir la dimensionalidad del espacio original de atributos; c) se utilizó Naive Bayes como algoritmo de clasificación, ya que en experimentos previos este algoritmo resultó ser el más apropiado; d) se aplicó un esquema de validación cruzada en 10 pliegues para la evaluación del clasificador. Este esquema consiste en dividir aleatoriamente el conjunto de entrenamiento en diez partes, conservando en cada partición la proporción original de las clases, posteriormente cada parte es mantenida una vez y el esquema de aprendizaje entrena sobre las nueve partes restantes, entonces la exactitud del clasificador es calculada sobre la parte conservada fuera del proceso de entrenamiento. Así, el proceso es ejecutado un total de diez veces sobre diferentes conjuntos de entrenamiento. Por último, los diez estimados de exactitud son promediados para producir una completa estimación de la misma.

3.1 Corpus

En esta sección se describen las condiciones y los corpus recolectados para nuestra tarea. Para la tarea de atribución de autoría se recopiló un corpus de poemas. Este corpus fue recopilado a partir de la Web, y consiste de 353 poemas escritos por cinco autores mexicanos. La Tabla 3-1 resume algunas estadísticas del corpus. Es importante mencionar que los poemas recolectados son documentos muy cortos (176 palabras en promedio).

Tabla 3-1 Corpus poetas

Poetas	Número de Documentos	Tamaño del Vocabulario	Número de Frases	Promedio de Palabras por Documento	Promedio de Frases por Documento
Efraín Huerta	48	3831	510	236.5	22.3
Jaime Sabines	80	3955	717	155.8	17.4
Octavio Paz	75	3335	448	162.6	27.2
Rosario Castellanos	80	4355	727	149.3	16.4
Rubén Bonifaz	70	4769	720	178.3	17.3

Se cuidó que se tratará de poetas contemporáneos para evitar identificar el estilo de un texto de acuerdo a las palabras anacrónicas[10] que aparecen en éste. Es sabido que escritos de diferentes épocas que pueden ser clasificados fácilmente por la presencia de estos anacronismos. Para ilustrar dicho problema se comparan los fragmentos de dos poemas, en los cuales podemos observar la presencia de anacronismos (Tabla 3-2).

[10] Error de cronología que consiste en atribuir a una época elementos pertenecientes a otra.

Tabla 3-2 Ejemplos de poemas con y sin presencia de anacronismos

Sor Juana Inés de la Cruz (1648 - 1695)	**Rosario Castellanos** (1925 - 1974)
Hombres necios que ***acusáis*** a la mujer sin razón sin ver que ***sois*** la ocasión de lo mismo que ***culpáis***	Y entonces supe: yo no estaba allí Ni en ninguna parte Ni había estado nunca ni estaría

Dos ejemplos de poemas recolectados se muestran en la Tabla 3-3.

Tabla 3-3 Fragmentos de dos poemas de nuestro corpus

Jaime Sabines	**Efraín Huerta**
Pequeña del amor, tú no lo sabes, tú no puedes saberlo todavía, no me conmueve tu voz ni el ángel de tu boca fría, ni tus reacciones de sándalo en que perfumas y expiras, ni tu mirada de virgen crucificada y ardida.	Éste es un amor que tuvo su origen y en un principio no era sino un poco de miedo y una ternura que no quería nacer y hacerse fruto. Un amor bien nacido de ese mar de sus ojos, un amor que tiene a su voz como ángel y bandera, un amor que huele a aire y a nardos y a cuerpo húmedo, un amor que no tiene remedio, ni salvación, ni vida, ni muerte, ni siquiera una pequeña agonía.

Un segundo corpus fue recolectado para la tarea de clasificación por género. La Tabla 3-4 muestra algunas estadísticas de este corpus. Este corpus fue recolectado de la Web y se incluyeron tres clases: novelas, noticias y poemas. El conjunto de noticias consta de 210 documentos recolectados de diferentes periódicos publicados en México. Éste contiene noticias relacionadas con desastres naturales como: incendio forestal, huracán, inundación, sequía y sismo. Cabe mencionar que para el caso de novelas se consideraron únicamente fragmentos de ellas. El corpus de género consta de 594 documentos en total.

Tabla 3-4 Corpus género

	Documentos	Tamaño del Vocabulario	Número de Frases	Promedio de palabras por Documento	Promedio de Frases por Documento
Novelas	27	47,903	3,236	1,774.19	119.87
Noticias	210	59,877	2,044	285.13	9.73
Poemas	353	62,712	3,122	177.65	8.84

A ambos corpus se les aplicó un proceso de normalización. Éste consistió en convertir todas las palabras a minúsculas, omitir el título de cada uno de los documentos y convertir los signos de puntuación a etiquetas previamente definidas, la Tabla 3-5 muestra estas etiquetas.

Tabla 3-5 Etiquetas de los signos de puntuación

Etiquetas	*Signo(s) de puntuación*
<PUNTOSS>	…
<PUNTO>	.
<COMA>	,
<DOSPUNTOS>	:
<PUNTO Y COMA>	;
<SIA>	¿
<SIC>	?
<SAA>	¡
<SAC>	!
<PA>	(
<PC>	)
<GUION>	—, -
<COMILLA>	", «, »,“,”

3.2 Resultados de Referencia

Dado que no existe un conjunto de datos estándar para compararnos con respecto a la clasificación por estilo, se seleccionaron y aplicaron diferentes métodos conocidos para tener una base de comparación. Cabe recordar que la presente tesis busca determinar el alcance de la información léxica de ahí que no se consideren métodos de referencia utilizando información sintáctica. Dentro de los métodos que se eligieron están: caracterización estilométrica, y tres caracterizaciones léxicas (bolsa de palabras, bolsa de palabras más palabras funcionales y n-gramas). A continuación, se muestran los resultados de estos métodos al aplicarlos a los corpus recolectados.

3.2.1 Atribución de Autoría

En las siguientes secciones, se muestran los resultados de los experimentos realizados con los métodos de referencia.

La Tabla 3-6 muestra los resultados al utilizar medidas estilométricas como atributos para representar los documentos. Bajo este enfoque podemos observar que no se obtienen buenos resultados (47.93% de exactitud). Suponemos que estos resultados se deben principalmente al reducido tamaño de los documentos.

Tabla 3-6 Método estilométrico

Total Atributos	**Autor**	**Precisión**	**Recuerdo**	**Medida-F**
10	EfrainH	0.46	0.35	0.40
Atributos IG	JaimeS	0.60	0.11	0.19
10	OctavioP	0.46	0.47	0.46
Exactitud	RosarioC	0.48	0.75	0.58
47.93%	RubenB	0.63	0.89	0.73
	Promedio	**0.53**	**0.51**	**0.47**

En contraste al usar la caracterización por bolsa de palabras, se obtienen resultados superiores. En la Tabla 3-7 se muestran los resultados, con este método se obtuvo una exactitud de 73.09%. Hay que recordar que este método alcanza exactitudes mayores al 90% en la clasificación temática. Así que suponemos que la exactitud alcanzada se debe básicamente a una distinción temática sin considerar en absoluto el estilo de los poetas. Por lo tanto, es claro que no es suficiente utilizar sólo las palabras de contenido, sino que también hay que poner especial interés en la estructura que tiene cada uno de los poemas.

Tabla 3-7 Método bolsa de palabras

Total Atributos	**Autor**	**Precisión**	**Recuerdo**	**Medida-F**
9,809	EfrainH	0.91	0.79	0.84
Atributos IG	JaimeS	0.76	0.68	0.72
193	OctavioP	0.88	0.69	0.78
Exactitud	RosarioC	0.52	0.76	0.62
73.09%	RubenB	0.84	0.76	0. 80
	Promedio	**0.78**	**0.74**	**0.75**

Debido a los resultados alcanzados con el método anterior se deseaba conocer el impacto de incluir las palabras funcionales. Así, que aplicamos el método de bolsa de palabras incluyendo las palabras funcionales. Esta nueva caracterización (Tabla 3-8) no tuvo ningún impacto en la clasificación. Es claro que el método no permite capturar la estructura del texto, pues sólo se consideran las palabras en su forma aislada.

Tabla 3-8 Método bolsa de palabras y palabras funcionales

Total Atributos	Autor	Precisión	Recuerdo	Medida-F
1,040	EfrainH	0.89	0.77	0.82
Atributos GI	JaimeS	0.75	0.61	0.68
212	OctavioP	0.90	0.73	0.81
Exactitud	RosarioC	0.54	0.83	0.65
73.09%	RubenB	0.81	0.73	0.77
	Promedio	**0.78**	**0.73**	**0.75**

El siguiente paso trata con caracterizaciones basadas en las combinaciones de palabras más usadas por un autor. Trabajos previos en autoría han demostrado la utilizada de los n-gramas al obtener resultados superiores al de bolsa de palabras [Diederich J. et al., 2002]. La Tabla 3-9, muestra los resultados al utilizar uni-gramas y bi-gramas como elementos para la caracterización. Con este método se obtuvo una mayor exactitud que en los métodos anteriores. Así, los n-gramas permiten describir con mejor precisión la estructura de los textos.

Tabla 3-9 Método uni-gramas y bi-gramas

Total Atributos	Autor	Precisión	Recuerdo	Medida-F
45,245	EfrainH	0.93	0.79	0.85
Atributos GI	JaimeS	0.82	0.73	0.77
455	OctavioP	1.00	0.73	0.86
Exactitud	RosarioC	0.57	0.86	0.68
78.75%	RubenB	0.91	0.83	0.87
	Promedio	**0.85**	**0.79**	**0.81**

Sin embargo, al usar n-gramas de mayor grado en este caso incluyendo hasta tri-gramas, se produce una leve caída en la exactitud, como se observa en la Tabla

3-10. Esto se debe principalmente a la gran cantidad de atributos agregados y además dichos atributos no son relevantes en la clasificación.

Tabla 3-10 Método uni-grams, bi-gramas y tri-gramas

Total Atributos	Autor	Precisión	Recuerdo	Medida-F
91,013	EfrainH	0.93	0.79	0.85
Atributos GI	JaimeS	0.81	0.73	0.76
590	OctavioP	1.00	0.64	0.78
Exactitud	RosarioC	0.54	0.88	0.67
76.77%	RubenB	0.91	0.81	0.86
	Promedio	**0.84**	**0.77**	**0.78**

3.2.2 Clasificación por Género

De igual manera que al primer corpus se aplicaron los diferentes métodos de referencia para permitir la evaluación comparativa del método propuesto.

En primer lugar, se aplicó una caracterización estilométrica. Recordemos que esta caracterización considera la longitud de las palabras y la riqueza del vocabulario, entre otras (ver sección 2.3.1).

En la Tabla 3-11 se muestran los resultados obtenidos con este método. En los resultados se puede apreciar que la exactitud es del 97.46%, lo cual muestra que el método de caracterización es muy bueno. Al parecer este resultado se debe en gran medida a las diferencias de tamaño entre los documentos de cada clase.

Tabla 3-11 Método estilométrico

Total Atributos	Género	Precisión	Recuerdo	Medida-F
10	Novelas	0.75	0.78	0.76
Atributos GI	Noticias	0.99	0.97	0.98
10	Poemas	0.97	0.99	0.99
Exactitud				
97.46%	**Promedio**	**0.90**	**0.91**	**0.91**

Utilizando el método de caracterización de bolsa de palabras se obtiene una exactitud del 97.80%. Los resultados experimentales se muestran en la Tabla 3-12. Como puede observarse se utilizaron 3,754 atributos después de aplicar ganancia de información contrastando con los 10 atributos del método anterior.

Tabla 3-12 Método bolsa de palabras

Total Atributos	Género	Precisión	Recuerdo	Medida-F
22,391	Novelas	0.89	0.89	0.89
Atributos GI	Noticias	1.00	0.97	0.98
3,754	Poemas	0.99	0.99	0.98
Exactitud				
97.80%	**Promedio**	**0.96**	**0.95**	**0.95**

Cuando se agregaron las palabras funcionales a la bolsa de palabras los resultados no cambiaron significativamente, como se observa en Tabla 3-13. En este caso, al aplicar ganancia de información se obtiene un total de 3,969 atributos.

Tabla 3-13 Método bolsa de palabras y palabras funcionales

Total Atributos	Género	Precisión	Recuerdo	Medida-F
22,817	Novelas	0.89	0.89	0.89
Atributos GI	Noticias	1.00	0.97	0.99
3,969	Poemas	0.98	0.99	0.98
Exactitud				
97.97%	**Promedio**	**0.96**	**0.95**	**0.95**

Finalmente, al utilizar n-gramas lo que se intenta es capturar la estructura del los textos por medio de simples secuencias de palabras. Ya que este método toma en cuenta tanto palabras funcionales como de contenido, debido a que en este caso se incluyeron los signos de puntuación y las palabras funcionales al obtener los n-gramas.

Los resultados que se obtuvieron se muestran en la Tabla 3-14. Como podemos observar se obtuvo un total de 121,238 atributos usando tanto uni-gramas como bi-gramas. De la misma manera se aplicó la técnica de ganancia de información para reducir la dimensionalidad. Por lo tanto, se redujo a 10, 807 atributos que es menos del 9% del conjunto de atributos original.

Ahora bien, el resultado que se obtuvo con respecto a la exactitud es igual al método anterior. En este caso se puede observar claramente que el número de atributos es mucho mayor que en los métodos anteriores. Aunque al comparar los resultados de la Medida-F se puede apreciar que es ligeramente mayor (0.95 del método anterior contra 0.96 utilizando uni-gramas y bi-gramas).

Tabla 3-14 Método uni-gramas y bi-gramas

Total Atributos	**Género**	**Precisión**	**Recuerdo**	**Medida-F**
121,238	Novelas	0.93	0.89	0.91
Atributos GI	Noticias	1.00	0.97	0.98
10,807	Poemas	0.97	0.99	0.98
Exactitud				
97.97%	**Promedio**	**0.97**	**0.95**	**0.96**

El último método, al considerar tri-gramas, obtuvo los mejores resultados entre los diferentes métodos. Aunque la desventaja es la dimensionalidad de la matriz, pues aumenta por más de 3,000 atributos con respecto al anterior.

Tabla 3-15 Método uni-gramas, bi-gramas y tri-gramas

Total Atributos	**Género**	**Precisión**	**Recuerdo**	**Medida-F**
154,836	Novelas	0.92	0.89	0.91
Atributos GI	Noticias	1.00	0.97	0.97
14,055	Poemas	0.98	0.99	0.99
Exactitud				
98.14%	**Promedio**	**0.97**	**0.95**	**0.96**

En resumen, el método tradicional de bolsa de palabras tuvo una exactitud del 97.80% contra el 98.14% del método de n-gramas. Bajo un marco costo-beneficio podríamos considerar el caso tradicional de bolsa de palabras como el más adecuado. Cabe mencionar, que a pesar de que el método estilístico tiene un menor costo –sólo son necesarios 10 atributos– no tiene el mismo comportamiento en las colecciones de género y autoría. El excelente comportamiento del método estilístico en el caso de género se debe, en gran medida, a las diferencias de tamaño entre los documentos de cada clase. Ya que podemos observar que los resultados obtenidos al aplicar el método al corpus de poetas sólo alcanza un 47.93%, donde los tamaños de los documentos entre clases son muy cercanos.

Capítulo 4

Métodos Propuestos

Los métodos de caracterización expuestos en el capítulo 3 tienen ciertas desventajas. En el caso de la caracterización sintáctica se necesitan recursos lingüísticos apropiados para el lenguaje que se está tratando, y se tiene que lidiar con herramientas incompletas cuyo análisis sintáctico además de parcial puede presentar errores importantes. Por otro lado, en caracterización léxica podemos encontrar dos desventajas importantes: 1) la bolsa de palabras no toma en cuenta el orden de aparición de las palabras en el texto, y pierde la estructura del texto; 2) los n-gramas, a pesar de conservar la estructura, llevan a una explosión combinatoria y en consecuencia una alta dimensionalidad en el espacio de atributos.

En este trabajo de tesis se presenta un nuevo método para la clasificación de documentos por estilo. Este método caracteriza los documentos por un conjunto de secuencias relevantes que combinan palabras funcionales, de contenido y signos de puntuación. La idea es usar estas secuencias para clasificar los documentos, en vista de que éstas expresan las colocaciones léxicas[11] más significativas utilizadas por el autor. Tradicionalmente, las secuencias se extraen aplicando un cálculo general de n-gramas. En contraste, nosotros proponemos descubrirlos por medio del cálculo de secuencias frecuentes maximales.

[11] Nos referimos a combinaciones frecuentes de unidades léxicas (palabras), de manera particular, dada la naturaleza de este trabajo usaremos este término para combinaciones léxicas inmediatamente contiguas.

En la siguiente sección se define el concepto de secuencia frecuente maximal, y posteriormente se describe el método básico propuesto para la clasificación de textos por estilo.

4.1 Secuencias Frecuentes Maximales

A continuación se definen las secuencias frecuentes maximales (SFM). Asumiendo que D es un conjunto de textos (por texto nos referimos a un documento completo o incluso a una sola oración), donde cada texto consiste de una secuencia de palabras. Tenemos la siguientes definiciones [Ahonen-Myka H., 2002].

Definición 1. Una secuencia $p = a_1 \ldots a_k$ es una subsecuencia de una secuencia q si todos los elementos a_i, $1 \leq i \leq k$, ocurren en q y además ocurren en el mismo orden que en p. Si una secuencia p es una subsecuencia de una secuencia q, entonces se dice que p ocurre en q.

Definición 2. Una secuencia p es *frecuente* en D si p es una subsecuencia de al menos σ textos de D, donde σ es un umbral de frecuencia predefinido.

Definición 3. Una secuencia p es una *secuencia frecuente maximal* en D si no existe alguna otra secuencia p' en D tal que p es una subsecuencia de p' y p' es frecuente en D.

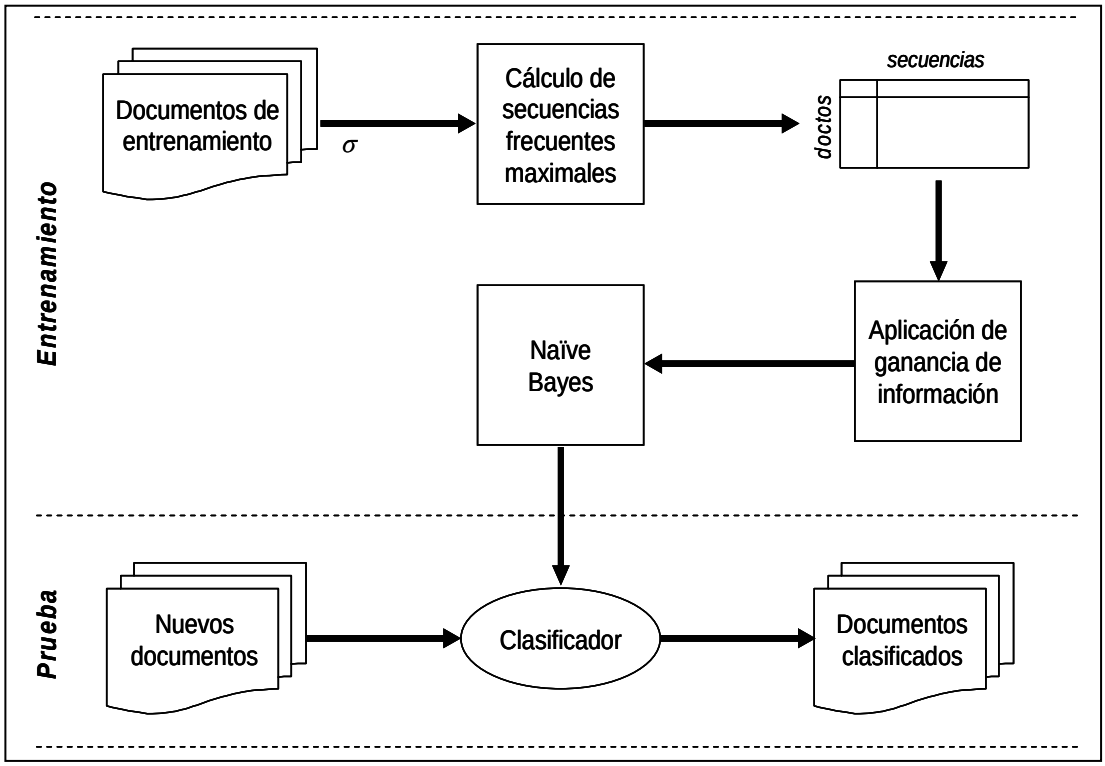
Una vez introducidas las secuencias frecuentes maximales, el problema de minería puede declararse formalmente como sigue: Dada una colección de textos D y un valor entero arbitrario σ tal que $1 \leq \sigma \leq |D|$, enumerar todas las secuencias frecuentes maximales en D.

Es importante mencionar que la implementación de un método para minería basado en secuencias frecuentes maximales no es una tarea trivial dada su complejidad computacional. El algoritmo usado en nuestros experimentos está descrito en [García R. et al., 2006].

4.2 Método Básico

El método básico se basa en secuencias frecuentes maximales para representar los documentos. La Figura 4-1 muestra el proceso de caracterización y clasificación de los documentos. La caracterización consiste en tres pasos principales: (i) Se calculan las secuencias frecuentes maximales dado un umbral de frecuencia σ; (ii) Se utilizan estas secuencias para representar cada uno de los documentos considerando un pesado booleano; y (iii) Se aplica ganancia de información para reducir el conjunto de atributos.

Figura 4-1 Proceso de caracterización básica y clasificación



En el Método 4-1 se muestra el método básico propuesto; incluyendo la fase de creación del clasificador y su aplicación. Como puede observarse el método depende de la adecuada definición del umbral de frecuencia σ. Se espera que los distintos valores de σ generen diferentes conjuntos de secuencias de palabras, y por consiguiente produzcan diferentes tasas de clasificación.

Por ejemplo, los valores bajos de σ permiten extraer secuencias grandes y favorecen la tasa de precisión, mientras los valores altos de σ tienden a generar muchas secuencias cortas que contribuyen al recuerdo. Desafortunadamente, el valor de σ más adecuado es influenciado por el tamaño de la colección de documentos, y por lo tanto debe ser determinado empíricamente para cada situación en particular. Por esta razón es que en la siguiente sección se propone un método iterativo que trata de definir una condición de paro, evitando una búsqueda exhaustiva de las secuencias.

Método 4-1 Método básico para la clasificación por estilo

D_T es el conjunto de documentos etiquetados que será usado para entrenar
d es un documento anónimo

ENTRENAMIENTO

1. Inicializar el umbral de frecuencia σ
2. Extraer todas las secuencias frecuentes maximales de D_T correspondientes al umbral de frecuencia dado
3. Representar las instancias de entrenamiento usando las secuencias frecuentes maximales obtenidas como características utilizando el pesado booleano.
4. Realizar el entrenamiento con el algoritmo de aprendizaje dado.

CLASIFICACIÓN

1. Construir la representación de *d* de acuerdo al conjunto de características extraídas en la etapa de entrenamiento.
2. Clasificar la nueva instancia con el clasificador entrenado.

4.2.1 Resultados del Método Básico

En las siguientes tablas se muestran los resultados al aplicar el método básico a las colecciones de documentos. Cabe mencionar que el número de secuencias que se muestra en cada una de las tablas es el resultado al aplicar la técnica de ganancia de información.

En la Tabla 4-1 se muestran los resultados obtenidos al aplicar el método básico al corpus de poetas. Estos resultados superan a los reportados por el método estilométrico y el de bolsa de palabras. Sin embargo, no son mejores que los alcanzados con n-gramas (78.75%). Sin embargo, es interesante observar que en el mejor caso utilizando SFM con σ =3 se tiene un número menor de atributos que con n-gramas, 203 y 455 respectivamente.

Tabla 4-1 Resultados del método básico en el corpus poetas

σ	Número de Secuencias	Promedio de palabras por secuencia	Exactitud	Precisión Promedio	Recuerdo Promedio
2	141	2.59	68.60%	0.76	0.69
3	203	2.32	77.30%	0.82	0.77
4	225	2.26	77.30%	0.82	0.77
5	195	1.67	77.10%	0.81	0.77
6	156	1.59	75.40%	0.79	0.75
7	129	1.57	74.80%	0.78	0.74
8	124	1.50	74.20%	0.76	0.74
9	105	1.46	71.40%	0.73	0.71
10	94	1.45	70.50%	0.72	0.70

Los resultados obtenidos al aplicar el método básico a nuestro corpus de género pueden observarse en la Tabla 4-2. En este caso, los resultados alcanzados no logran ser mayores a los obtenidos con el método tradicional (bolsa de palabras). Los

resultados son prácticamente iguales (97.80% bolsa de palabras y palabras funcionales contra 97.46% del método básico) pero obtenidos con un número menor de atributos.

Tabla 4-2 Resultados método básico en el corpus género

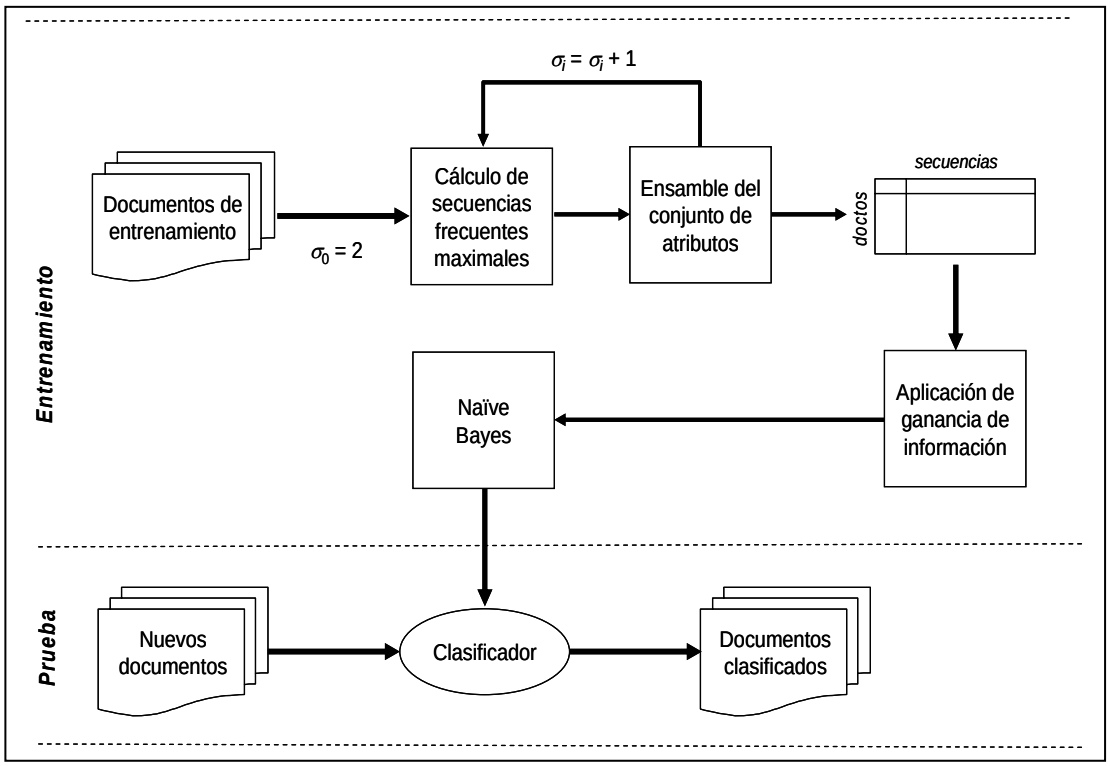
σ	Número de Secuencias	Promedio de palabras por secuencia	Exactitud	Precisión Promedio	Recuerdo Promedio
2	5,895	2.26	96.44%	0.96	0.92
3	5,605	2.10	97.12%	0.96	0.93
4	5,275	2.06	97.29%	0.95	0.93
5	4,628	1.96	97.29%	0.95	0.93
6	4,026	1.89	97.29%	0.95	0.94
7	3,580	1.84	96.78%	0.95	0.94
8	3,199	1.79	97.12%	0.95	0.94
9	2,884	1.76	97.46%	0.95	0.96
10	2,609	1.73	97.12%	0.94	0.94

Como se mencionó anteriormente, el gran problema del método es determinar el umbral de frecuencia adecuado. De ahí, la búsqueda de un segundo método de caracterización más general. Este método iterativo se presenta en la siguiente sección.

4.3 Método Iterativo

El método iterativo plantea construir un conjunto de características que combine las secuencias frecuentes maximales extraídas con diferentes valores de σ. La idea es construir un conjunto de características a través de un proceso iterativo, incrementando a cada paso el valor de σ.

Figura 4-2 Proceso de caracterización iterativa y clasificación



En la Figura 4-2 se ilustra este proceso de caracterización. El proceso inicia calculando las SFM con un umbral mínimo de frecuencia de 2 y este proceso se repite incrementando el umbral mientras las nuevas SFM calculadas sean secuencias de tamaño mayor a 1. Es decir, se calcularán todas las colocaciones frecuentes entre dos o más palabras. Con esto la caracterización de los documentos se basa en secuencias de dos o más palabras que capturan las colocaciones más representativas del corpus. En el Método 4-2 se describe el método para el cálculo de esta caracterización y su aplicación en la clasificación.

Método 4-2 Método para realizar la clasificación de textos usando la caracterización iterativa

---

D_T es el conjunto de documentos previamente etiquetados para realizar el entrenamiento
d es un documento anónimo

ENTRENAMIENTO

1. Inicializar el valor del umbral de frecuencia $\sigma = 2$
2. Inicializar el conjunto de las características $F_1 = \varnothing$
3. HACER
   a. Extraer todas las secuencias frecuentes maximales de D_T correspondientes al umbral de frecuencia σ. Nombrar el conjunto de secuencias obtenidas S_σ
   b. Integrar las secuencias obtenidas al conjunto de características, es decir, $F_\sigma = F_{\sigma\text{-}1} \cup S_\sigma$
   c. Incrementar el umbral de frecuencia, es decir, $\sigma = \sigma + 1$

   MIENTRAS ($S_{\sigma\text{-}1}$ contenga al menos una secuencia de dos o más palabras no incluida en $F_{\sigma\text{-}2}$)
4. Representar las instancias de entrenamiento usando las secuencias de F_σ utilizando el pesado booleano
5. Realizar el entrenamiento con el algoritmo de aprendizaje dado

CLASIFICACIÓN

1. Construir la representación de *d* de acuerdo al conjunto de características F_σ calculado durante el entrenamiento.
2. Clasificar la nueva instancia con el clasificador entrenado.

En resumen, al construir de esta manera el conjunto de características se busca conservar la naturaleza secuencial del texto y esto nos permite capturar las estructuras estilísticas utilizadas por un autor.

4.3.1 Resultados del Método Iterativo

Al aplicar el método iterativo al corpus de poetas se ensambló un conjunto de 425 características. La Tabla 4-3 muestra algunos datos relacionados con la construcción de dicho conjunto.

Tabla 4-3 Construcción del conjunto de características en el corpus poetas

σ	Secuencias extraídas	Secuencias agregadas	Longitud promedio de las secuencias agregadas	Número de características
2	141	141	2.58	141
3	203	100	1.71	241
4	225	80	1.76	321
5	195	53	1.74	374
6	156	23	1.35	397
7	129	13	1.46	410
8	124	12	1.25	422
9	105	3	1	425

La Tabla 4-4 muestra algunas de las secuencias obtenidas con el método básico (usando un umbral tres) y las obtenidas con el método iterativo, en particular para el corpus de poemas. Es claro que la unión de SFM contendrá un mayor número de secuencias que nos permitan capturar la estructura de los poemas y con ello obtener mejores resultados. También podemos observar en dicha tabla los atributos con mayor ganancia de información en orden ascendente en ambos casos, sin embargo a pesar de que la unión contiene secuencias obtenidas con un umbral tres en algunos casos la ganancia de información es igual (sereno) y en otros casos varia (del_alba_<COMA>). Con lo anterior podemos concluir que al unir las secuencias con diferentes umbrales se obtiene un conjunto más representativo (425 secuencias después de aplicar GI) que el utilizar un conjunto de secuencias con un solo umbral (203 secuencias después de aplicar GI). Finalmente, es claro que el conjunto final de secuencias obtenidas con el método iterativo solo son secuencias frecuentes y no maximales, ya que habrá secuencias que contengan a otras secuencias.

La Tabla 4-5 muestra los resultados alcanzados con el método iterativo. De estos resultados, es claro que este método supera todos los métodos de caracterización mostrados anteriormente. Además, dado que el conjunto de características resultante es comparable en tamaño al conjunto de *n*-gramas, los resultados obtenidos muestran que es más adecuado determinar las secuencias de palabras por su frecuencia de ocurrencia y no por su longitud. Básicamente, el método propuesto permite seleccionar las secuencias de palabras más relevantes.

Tabla 4-4 Ejemplos de SFM con mayor GI para el corpus de poetas

SFM (Umbral: 3)	Unión SFM
sereno	sereno
del_alba_<COMA>	sonrisa_de
del_deseo	nardos
nardos	rosas_y
lágrimas_de	del_deseo
de_ternura	del_ansia
dolor_<PUNTO>	del_alba_<COMA>
voces_de	lágrimas_de
claveles	fiebre_<COMA>
fiebre_<COMA>	la_lenta
la_paloma	dolor_<PUNTO>
la_lenta	bronce
moría	de_ternura
el_ruido	la_paloma
y_miel	claveles
tu_sombra_<COMA>	voces_de
de_rosas	deleite
dominio	mi_frente
y_aire	veía
de_fiebre	el_ruido
pieles	y_sangre
de_bronce	desprecio
sencilla	alba_se
toda_la_vida	y_miel
alba_<DOSPUNTOS>	<COMA>_muy

Tabla 4-5 Resultados del método iterativo en el corpus poetas

Poetas	Precisión	Recuerdo
Efraín Huerta	1.00	0.75
Jaime Sabines	0.83	0.83
Octavio Paz	0.95	0.75
Rosario Castellanos	0.65	0.91
Ruben Bonifaz	0.94	0.87
Promedio	0.87	0.82
Exactitud Total	83%	

Por último, se aplicó el método iterativo al corpus de género. En este caso el método iterativo ensambló en total un conjunto de 37,523 características. La Tabla 4-6 muestra algunos datos relacionados con la construcción de dicho conjunto. Como podemos observar, el método terminó al insertar secuencias con umbral de 48.

Tabla 4-6 Construcción del conjunto de características en el corpus género

σ	Secuencias extraídas	Secuencias agregadas	Longitud promedio de las secuencias agregadas	Número de características
2	5,895	5,895	2.26	5,823
3	5,605	3,151	2.03	9,046
4	5,275	1,755	1.99	10,801
5	4,628	1,066	1.86	11,867
6	4,026	640	1.79	12,507
7	3,580	393	1.69	12,900
8	3,199	342	1.61	13,242
9	2,884	246	1.65	13,488
10	2,609	178	1.65	13,666
…	…	…	…	…
47	463	11	1.36	14,673
48	449	3	1.00	14,676

En la Tabla 4-7 podemos observar los resultados de la clasificación al aplicar el método iterativo como método de caracterización. En este caso, los resultados no superan los métodos anteriores pero se mantienen en el mismo orden de clasificación con un total de 14,676 atributos después de aplicar ganancia de información.

Tabla 4-7 Resultados del método iterativo en el corpus género

Corpus	Precisión	Recuerdo
Novelas	0.923	0.889
Noticias	0.995	0.952
Poemas	0.967	0.994
Promedio	0.961	0.945
Exactitud Total	97.50%	

En las siguientes tablas se muestra un resumen de los experimentos realizados contrastando los resultados de los métodos propuestos con los métodos de referencia. Los resultados alcanzados sobre el corpus de poetas se muestran en la Tabla 4-8. Como podemos apreciar el mejor método de caracterización para este corpus es el método iterativo. Básicamente, nuestro método permite seleccionar las colocaciones más relevantes utilizadas por los autores. Como puede observarse en el caso de métodos con atributos basados en colocaciones (i. e. n-gramas o secuencias), el método iterativo obtiene mejores resultados, aun con una menor cantidad de atributos. De ahí, que podemos decir que las colocaciones léxicas capturadas por el método iterativo son más representativas que las calculadas por los otros métodos.

Tabla 4-8 Resultados experimentales: Corpus poetas

Métodos	Atributos	Atributos GI	Precisión Promedio	Recuerdo Promedio	Exactitud
análisis estilométrico	10	10	0.52	0.51	47.93%
bolsa de palabras	9,809	193	0.78	0.74	73.09%
bolsa de palabras + palabras funcionales	10,040	212	0.78	0.73	73.09%
uni-gramas + bi-gramas	45,245	455	0.84	0.79	78.75%
uni-gramas + bi-gramas +tri-gramas	91,013	590	0.84	0.79	78.75%
método básico	4,273	203	0.82	0.77	77.30%
Método iterativo	11,442	425	0.87	0.82	83.00%

En la Tabla 4-9 se presentan los resultados alcanzados con el corpus de género. Como puede observarse, el mejor método de caracterización para este corpus es n-gramas. Ahora bien, dado que la diferencia es mínima entre utilizar bolsa de palabras y n-gramas, y dado el costo computacional en el cálculo de n-gramas, dependerá de la aplicación cuál escoger.

Por otro lado, observando los resultados podemos afirmar que la distinción de las clases en este corpus recae no sobre características estilísticas sino sobre características temáticas.

Tabla 4-9 Resultados experimentales: Corpus género

Métodos	Atributos	Atributos GI	Precisión Promedio	Recuerdo Promedio	Exactitud
análisis estilométrico	10	10	0.91	0.91	97.46%
bolsa de palabras	22,391	3,754	0.96	0.95	97.80%
bolsa de palabras + palabras funcionales	22,817	3,969	0.95	0.95	97.97%
uni-gramas + bi-gramas	121,238	10,807	0.97	0.95	98.14%
uni-gramas + bi-gramas +tri-gramas	154,836	14,055	0.97	0.95	98.14%
método básico	3,478	2,884	0.95	0.96	97.46%
método iterativo	37,523	14,676	0.96	0.95	97.50%

A manera de conclusión podemos afirmar que el método propuesto funciona adecuadamente, en específico, para el caso de atribución de autoría. En este corpus es clara la mejoría que se alcanza contra los otros métodos. Es importante remarcar que dadas las características de este corpus la tarea de categorización se complica. Se trata de textos muy cortos bajo temáticas similares. De ahí la relevancia de los resultados alcanzados con nuestro método. El método de caracterización nos permitió representar los documentos a través de un conjunto de secuencias que combinan palabras funcionales, de contenido y signos de puntuación. Por lo tanto, el método nos permite combinar tanto características estilísticas como temáticas.

Sin embargo, para corpus cuya distinción entre clases no sólo incluye el estilo sino se complementa fuertemente con la temática, como el caso de nuestro corpus de género, el método se mantiene al mismo nivel que los enfoques tradicionales, como la bolsa de palabras.

Capítulo 5

Conclusiones y Trabajo Futuro

5.1 Conclusiones

En este trabajo de tesis se propusieron dos métodos de caracterización para la clasificación de documentos por su estilo. Los métodos pretenden determinar el conjunto de colocaciones propias de un estilo, estas colocaciones se expresan a través de secuencias de elementos léxicos: palabras (tanto de contenido como funcionales) y signos de puntuación.

El primer método caracteriza los documentos a partir de las secuencias frecuentes maximales de la colección, para el cálculo es indispensable determinar a priori un umbral de frecuencia, de esta manera el desempeño del método depende en gran medida de la adecuada definición de dicho umbral. Para ello es necesario experimentar con varios umbrales hasta obtener el conjunto de características más adecuado para la colección de documentos objetivo.

Es precisamente este problema que motiva la propuesta de un segundo método (método iterativo), dicho método ensambla un conjunto de características al combinar las secuencias frecuentes maximales extraídas con diferentes umbrales de frecuencia. Este conjunto es el resultado de un proceso iterativo y de la definición de una condición de paro.

Los resultados alcanzados mostraron la pertinencia del método propuesto para el caso de atribución de autoría, donde se obtuvieron resultados relevantes. Bajo esta problemática fue claro que la distinción recaía en el estilo de cada autor. Por otro lado, al aplicar nuestro método sobre el corpus de género los resultados fueron equiparables a los obtenidos por métodos tradicionales. Debido, en gran parte, a que las clases en este corpus dependen no sólo del estilo sino de la temática.

En conclusión el método propuesto de caracterización es conveniente para categorizaciones por estilo, como se observó en la tarea de atribución de autoría, y se observó que el método también puede ser aplicado a tareas de clasificación temática con resultados similares a las técnicas tradicionales (véase el Apéndice A).

5.2 Trabajo Futuro

Una aportación del método propuesto es su aplicación sobre textos extremadamente cortos. La clasificación de textos cortos, aun de manera temática, es un problema abierto. Uno de los primeros trabajos sería aplicar nuestro método a la clasificación temática de textos cortos.

Otro punto a explorar es la clasificación de correos electrónicos, donde no hay una clara división de si se trata de un problema de clasificación temático o de estilo y en donde nuestro método se comporta satisfactoriamente.

Otro campo donde aplicar nuestro método es la clasificación de opiniones. En este caso se aborda nuevamente una clasificación no temática (se está de acuerdo o en desacuerdo a cierto suceso) donde los términos subjetivos dan pie a clasificar el documento.

También se plantea mejorar el método desde el punto de vista de eficiencia. El método al ensamblar las secuencias a partir de diferentes umbrales de frecuencia deja de tener sentido hablar de secuencias maximales. Es por ello que se buscarán métodos

alternativos que obtengan las secuencias más frecuentes evitando el cálculo repetitivo.

Por último, se plantea utilizar el método propuesto en corpus en otros idiomas, dada la generalidad del enfoque al emplear únicamente elementos léxicos.

Publicaciones

Coyotl-Morales Rosa María, Luis Villaseñor-Pineda, Manuel Montes-y-Gómez and Paolo Rosso. "*Authorship Attribution using Word Sequences"*, 11th Iberoamerican Congress on Pattern Recognition (CIARP 2006). Lecture Notes in Artificial Intelligence Springer, pp. 844-853, 2006.

Bibliografía

[Ahonen H., 2002] Ahonen H. "Discovery of Frequent Word Sequences in Text". Proceedings of the ESF Exploratory Workshop on Pattern Detection and Discovery. London, UK, pp. 180-189, 2002.

[Alcaraz E. & Martínez A., 1997] Alcaraz E. & Martínez A. "Diccionario de Lingüística Moderna". Editorial Ariel, S.A. 1997.

[Argamon S. & Sterling S., 2003] Argamon S. & Sterling S. "Learning Algorithms and Features for Multiple Authorship Discrimination". Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis. Acapulco, Mexico, 2003.

[Argamon S. & Levitan S., 2005] Argamon S. & Levitan S. "Measuring the Usefulness of Function Words for Authorship Attribution". Proceedings of Association for Literary and Linguistic Computing/ Association Computer Humanities. University Of Victoria, Canada, 2005.

[Ass K. & Eikvil L., 1999] Ass K. & Eikvil L. "Text categorization: A survey". Technical Report. Norwegian Computing Center, 1999.

[Bruce D., 1972] Bruce D. "Purposeful Writing". Addison-Wesley publishing company, 1972.

[Cook M., 2003] Cook M. “Experimenting to Produce a Software Tool for Authorship”. This report is submitted in partial fulfilment of the requirement for the degree of Bachelor of Engineering with Honours in Software Engineering, 2003.

[Corney M. et al., 2002] Corney M., de Vel O., Anderson A. & Mohay G. “Gender-Preferential Text Mining of E-mail Discourse”. 18th Annual Computer Security Applications Conference. Las Vegas, Nevada, pp. 9-13, 2002.

[Chaski C., 2005] Chaski C. “Who's at the Keyword? Authorship Attribution in Digital Evidence Investigations”. International Journal of Digital Evidence (IJDE), 2005.

[de Vel O. et al., 2001] de Vel O., Anderson A., Corney M. & Mohay G. “Mining Email Content for Author Identification Forensics”. Special Section on Data Mining for Intrusion Detection and Threat Analysis SIGMOD Record. New York, USA Volume 30 Issue 4, pp. 55-64, 2001.

[de Vel O. et al., 2001] de Vel O., Anderson A., Corney M. & Mohay G. "Multi-Topic E-mail Authorship Attribution Forensics", ACM Conference on Computer Security - Workshop on Data Mining for Security Applications. Philadelphia, PA, USA, 2001

[Diederich J. et al., 2003] Diederich J., Kindermann J., Leopold E. & Paas G. “Authorship Attribution with Support Vector Machines”. Applied Intelligence. Kluwer Academic Pushishers, Hingham, MA, USA. Volume 19 Issue 1, p.p. 109-123, 2003.

[Finn A. & Kushmerick N., 2003] Finn A. & Kushmerick N. "Learning to classify documents according to genre". IJCAI-03 Workshop on Computational Approaches to Style Analysis and Synthesis. Acapulco, Mexico, 2003.

[Fürnkranz J., 1998] Fürnkranz J. "A Study Using n-gram Features for Text Categorization". Technical Report OEFAI-TR-9830, Austrian Institute for Artificial Intelligence. Wien, Austria, 1998.

[García R. et al., 2006] García R., Martínez F. & Carrasco A. "A New Algorithm for Fast Discovery of Maximal Sequential Patterns in a Document Collection". International Conference on Computational Linguistics and text Processing, CICLing-2006. Mexico City, Mexico, pp. 514-523, 2006.

[Henzinger M., 2000] Henzinger M. "Link Analysis in Web Information Retrieval". IEEE Data Engineering Bulletin, Volume 23, Issue 3, pp. 3-8, 2000.

[Hernández J. et al., 2004] Hernández J., Ramírez J. & Ferri C. "Introducción a la minería de Datos". Prentice Hall, Pearson Educación, S.A., Madrid, 2004.

[Joachims T., 1998] Joachims T. "Text Categorization with Support Vector Machines: Learning with many relevant features". Proceedings of {ECML}-98, 10th European Conference on Machine Learning. Chemnitz, Germany, Issue 1398., pp. 137-142, 1998.

[Kaster A. et al., 2005] Kaster A., Siersdorfer S. & Weikum G. "Combining Text and Linguistic Document Representations for Authorship Attribution". Workshop Stylistic Analysis of Text for Information Access, 28th Int. SIGIR MPI, Saarbrücken, Issue 1. pp. 27-35, 2005.

[Keselj V. et al., 2003] Keselj V., Peng F., Cercone N. & Thomas C. “N-gram-based Author Profiles for Authorship Attribution”. Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING'03, Dalhousie University, Halifax, Nova Scotia, Canada, pp. 255-264, 2003.

[Lewis D., 1991] Lewis D. “Evaluating text categorization”. Proceedings of Speech and Natural Language Workshop. Morgan Kaufmann, California, USA, pp. 312-318, 1991.

[Luyckx K. & Daelemans W., 2004] Luyckx K. & Daelemans W. “Shallow text analysis and machine learning for authorship attribution”. Proceedings of the Fifteenth Meeting of Computational Linguistics in the Netherlands (CLIN’04), Utrecht: LOT, pp. 149-160, 2004.

[Malyutov M., 2004] Malyutov M. “Authorship Attribution of Texts: a Review”. Proceedings of the program "Information transfer" held in ZIF. University of Bielefeld, Germany, pp. 362-380, 2004

[Peng F. et al., 2004] Peng F., Schuurmans D. & Wang S. “Augmenting Naïve Bayes Classifiers with Statistical Languages Models”. Information Retrieval. Kluwer Academic Publishers, Pisa, Italy. Volume 7, Issue 3-4, pp. 317-345, 2004.

[Porter M., 1980] Porter M. “An Algorithm for Suffix Stripping”, Program, vol.14, no. 3, pp.130-137, 1980.

[Sebastiani F., 1999] Sebastiani F. “A Tutorial on Automated Text Categorization”. Argentinian Symposium on Artificial Intelligence. Buenos Aires, pp. 7-35, 1999.

[Sebastiani F., 2005] Sebastiani F. "Text Categorization". Enciclopedia of Database Technologies and Applications. Idea Group Publishing, Hershey, US, pp. 683-687, 2005.

[Stamatatos E. et al., 2000] Stamatatos E., Fakotakis N. & Kokkinakis G. "Text Genre Detection Using Common Word Frequencies". Proceedings of the 18th Int. Conference on Computational Linguistics. Saarbruecken, Alemania, pp.808-814 , 2000.

[Stamatatos E. et al., 2001] Stamatatos E., Fakotakis N. & Kokkinakis G. "Computer-Based Authorship Attribution Without Lexical Measures". Computers and the Humanities. Publisher Springer Netherlands, Kluwer, Volume 35, Issue 2, pp. 193-214, 2001.

[Téllez A. et al., 2003] Téllez A., Montes M., Fuentes O. & Villaseñor L. "Clasificación automática de textos de desastres naturales en México". International Conference on Computer Science, CIICC-2003, Oaxtepec, México, pp. 269-263, 2003.

[Téllez A., 2005] Téllez A. "Extracción de Información con Algoritmos de Clasificación". Tesis de Maestría. Instituto Nacional de Astrofísica, Óptica y Electrónica. Puebla, México., 2005.

[Vicedo J. et al., 2003] Vicedo J., Rodríguez H., Peñas A. & Massot M. "Los sistemas de Búsqueda de Respuestas desde una perspectiva actual". Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural, pp. 351-367, 2003.

[Yang Y. & Pedersen J., 1997] Yang Y. & Pedersen J. "A comparative study on feature selection in text categorization". Proceedings of {ICML}-97, 14th International Conference on Machine Learning. Morgan Kaufmann Publishers, San Francisco, US, pp. 412-420, 1997.

[Zhao Y. & Zobel J., 2005] Zhao Y. & Zobel J. “Effective and Scalable Authorship Attribution Using Function Words”. The 2nd Asian Information Retrieval Symposium, Korea, pp.174-190, 2005.

Apéndice A

RESULTADOS AL APLICAR LOS MÉTODO PROPUESTO EN LA CLASIFICACIÓN TEMATICA

El presente apéndice presenta los resultados de aplicar nuestro método a la problemática tradicional de clasificación temática. El objetivo de estos experimentos es comprobar la generalidad del método. Para probar los métodos se utilizó un corpus temático, de noticias sobre desastres naturales, el cual fue recopilado en trabajos previos [Téllez A. et al., 2003]. El corpus consiste de noticias relacionadas con los desastres naturales como: incendio forestal, huracán, inundación, sequía y sismo. El corpus consta de 439 documentos en total. La Tabla A-1 resume algunas estadísticas de éste.

Tabla A-1 Corpus desastres

Desastres	Documentos	Tamaño del Vocabulario	Número de Frases	Promedio de palabras por Documento	Promedio de Frases por Documento
Forestal	92	25,096	903	272.7826	9.8152
Huracán	76	23,676	824	311.5263	10.8421
Inundación	87	25,361	922	291.5057	10.5977
Sequía	41	11,646	367	284.1220	8.9512
Sismo	143	30,410	1,206	212.6573	8.4335

A.1 Resultados de Referencia

En la Tabla A-2 se muestran los resultados alcanzados con el método estilométrico. Como podemos observar, este método no obtiene buenos resultados, pues sólo se llega a un 39.41% de exactitud.

Tabla A-2 Método estilométrico

Total Atributos	**Desastres**	**Precisión**	**Recuerdo**	**Medida-F**
10	Forestal	0.42	0.42	0.42
Atributos GI	Huracán	0.27	0.74	0.39
4	Inundación	0.20	0.02	0.04
Exactitud	Sequía	0.20	0.02	0.04
39.41%	Sismo	0.61	0.52	0.56
	Promedio	**0.34**	**0.34**	**0.29**

Al aplicar bolsa de palabras se obtiene un 97.72% de exactitud (ver Tabla A-3), confirmando la pertinencia del método para la clasificación temática.

Tabla A-3 Método bolsa de palabras

Total Atributos	**Desastres**	**Precisión**	**Recuerdo**	**Medida-F**
12,230	Forestal	0.96	1.00	0.98
Atributos GI	Huracán	0.99	0.96	0.97
531	Inundación	0.96	0.99	0.97
Exactitud	Sequía	0.97	0.90	0.94
97.72%	Sismo	1.00	0.99	0.99
	Promedio	**0.98**	**0.97**	**0.97**

Como era de esperarse al agregar palabras funcionales (ver Tabla A-4) no se mejoran significativamente los resultados. Confirmando que este tipo de palabras no son de gran relevancia para la clasificación temática.

Tabla A-4 Método palabras y palabras funcionales

Total Atributos	Desastres	Precisión	Recuerdo	Medida-F
12,427	Forestal	0.97	1.00	0.98
Atributos GI	Huracán	0.99	0.97	0.98
547	Inundación	0.96	0.99	0.97
Exactitud	Sequía	0.97	0.90	0.94
97.95%	Sismo	1.00	0.99	0.99
	Promedio	**0.98**	**0.97**	**0.97**

Por otro lado, al caracterizar los documentos con uni-gramas y bi-gramas se puede apreciar que la exactitud disminuye con respecto al método de bolsa de palabras (ver Tabla A-5).

Tabla A-5 Método uni-gramas y bi-gramas

Total Atributos	Desastres	Precisión	Recuerdo	Medida-F
65,941	Forestal	0.97	1.00	0.98
Atributos IG	Huracán	0.97	0.91	0.94
1,576	Inundación	0.90	0.98	0.93
Exactitud	Sequía	0.95	0.90	0.93
96.13%	Sismo	1.00	0.97	0.99
	Promedio	**0.96**	**0.95**	**0.95**

Cuando agregamos tri-gramas la situación empeora debido al considerable aumento en el número de características (ver Tabla A-6).

Tabla A-6 Método uni-gramas, bi-gramas y tri-gramas

Total Atributos	Desastres	Precisión	Recuerdo	Medida-F
158,243	Forestal	0.97	1.00	0.98
Atributos GI	Huracán	0.97	0.90	0.93
2,347	Inundación	0.85	0.98	0.91
Exactitud	Sequía	0.92	0.83	0.87
94.76%	Sismo	1.00	0.96	0.98
	Average	**0.94**	**0.93**	**0.93**

A.2 Resultados con los métodos propuestos

La Tabla A-7 muestra los resultados con el método básico. Como se observa los resultados no logran ser mejores a los obtenidos con el método de bolsa de palabras, 96.81% y 97.72% respectivamente.

Tabla A-7 Resultados método básico

σ	Número de Secuencias	Promedio de palabras por secuencia	Exactitud	Precisión Promedio	Recuerdo Promedio
2	244	2.65	88.15%	0.89	0.86
3	643	2.82	93.85%	0.91	0.93
4	915	2.79	95.44%	0.96	0.93
5	900	2.74	95.44%	0.96	0.94
6	796	2.58	95.90%	0.96	0.95
7	709	2.48	95.67%	0.96	0.95
8	635	2.31	96.81%	0.97	0.96
9	583	2.26	96.13%	0.96	0.95
10	523	2.19	96.36%	0.96	0.95

Al aplicar el método iterativo se ensambló un conjunto de 2,455 características. La Tabla A-8 muestra algunos datos relacionados con la construcción de dicho conjunto. Como se puede observar el método terminó con umbral $\sigma = 30$.

Tabla A-8 Construcción del conjunto de características

σ	Secuencias extraídas	Secuencias agregadas	Longitud promedio de las secuencias agregadas	Número de características
2	244	244	2.65	244
3	643	484	2.80	728
4	915	427	2.80	1,155
5	900	281	2.57	1,436
6	796	209	2.33	1,645
7	709	144	2.56	1,789
8	635	123	2.33	1,912
9	583	77	2.38	1,989
10	523	63	2.16	2,052
…	…	…	…	…
29	193	6	2.33	2,453
30	175	2	1.00	2,455

La Tabla A-9 muestra los resultados alcanzados con el método iterativo. A pesar de que el método no supera la exactitud del método tradicional se obtienen resultados satisfactorios (96% de exactitud).

Tabla A-9 Resultados del método iterativo

Desastres	Precisión	Recuerdo
Forestal	0.979	1
Huracan	0.986	0.908
Inundación	0.878	0.989
Sequía	0.923	0.878
Sismo	1	0.965
Promedio	0.9532	0.948
Exactitud Total	96%	

En la Tabla A-10 se contrastan los resultados de los métodos propuestos con los resultados de los métodos de referencia. Como puede observarse la diferencia es poca demostrando la generalidad del método en la clasificación de textos por estilo como temática.

Tabla A-10 Resultados Experimentales: Corpus desastres

Método	Atributos	Atributos GI	Precisión promedio	Recuerdo promedio	Exactitud
análisis estilométrico	10	4	0.34	0.35	39.41%
bolsa de palabras	12,230	531	0.97	0.97	97.72%
bolsa de palabras + palabras funcionales	12,427	547	0.98	0.97	97.95%
uni-gramas + bi-gramas	65,941	1,576	0.96	0.95	96.13%
uni-gramas + bi-gramas+ tri-gramas	158,243	2,347	0.94	0.94	94.76%
método básico	2,673	635	0.96	0.95	96.81%
método iterativo	27,016	2,455	0.95	0.95	96.00%