UNIVERSIDAD AUTÓNOMA DEL ESTADO DE HIDALGO

Identificación del perfil de autores en redes sociales usando nuevos esquemas de pesado que enfatizan información de tipo personal

TESIS
QUE PARA OBTENER EL TÍTULO DE:
Doctora en Ciencias Computacionales

PRESENTA:
Rosa María Ortega Mendoza

DIRECTORES DE TESIS:

Dra. Anilú Franco Árcega
Universidad Autónoma del Estado de Hidalgo (UAEH)

Dr. Manuel Montes y Gómez
Instituto Nacional de Ástrofísica, Óptica y Electrónica (INAOE)

Mineral de la Reforma, Hidalgo, Agosto, 2017

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE HIDALGO
Instituto de Ciencias Básicas e Ingeniería
Institute of Basic Sciences and Engineering
Área Académica de Computación y Electrónica
Computer Science and Electronics Department

ICBI-AACyE/1567/2017

Rosa María Ortega Mendoza
Presente.

Por este conducto le comunico que el jurado asignado para la revisión de su trabajo de tesis titulado **"Identificación del perfil de autores en redes sociales usando nuevos esquemas de pesado que enfatizan información de tipo personal"**, que para obtener el grado de Doctorado en Ciencias Computacionales, fue presentado por usted ha tenido a bien, en reunión de sinodales **autorizarlo para impresión.**

PRESIDENTE:	**DRA. MARÍA DE LOS ÁNGELES ALONSO LAVERNIA**
VOCAL:	**DR. RUSLAN FAIZOVICH GABBASOV**
SECRETARIO:	**DR. MANUEL MONTES Y GÓMEZ**
SUPLENTE:	**DRA. ANILU FRANCO ARCEGA**

Agradeciendo su puntual asistencia quedo a sus órdenes.

A T E N T A M E N T E.
"AMOR, ORDEN Y PROGRESO"
Mineral de la Reforma, Hgo., a 23 de agosto de 2017

UNIVERSIDAD AUTÓNOMA DEL ESTADO DE HIDALGO

Dr. Omar López Ortega
Coord. del DCC

Instituto de Ciencias Básicas e Ingeniería
Área Académica de Computación y Electrónica

c.c.p. Archivo OML/apl

Ciudad del Conocimiento
Carretera Pachuca - Tulancingo km. 4.5
Colonia Carboneras
Mineral de la Reforma, Hidalgo, México, C.P. 42184
Tel. +52 771 7172000 exts. 2250 y 2251
aacye_icbi@uaeh.edu.mx

A mis hijas,
el amor que me inspira, alegra y fortalece

Agradecimientos

Agradezco a Dios por permitirme alcanzar esta meta, brindarme fortaleza e iluminar mi camino.

A mis padres por su inagotable apoyo y palabras de aliento. De manera especial, a mi esposo e hijas por acompañarme siempre, por su amor, la paciencia brindada y por ser mi motivo de inspiración.

A mi asesora Dra. Anilú Franco Árcega por su colaboración en esta investigación, su apoyo constante durante estos años y por contribuir solidariamente en mi formación académica-científica.

A mi asesor Dr. Manuel Montes y Gómez, quien siempre estuvo presente compartiendo su entusiasmo por la investigación. Gracias por brindarme valiosas enseñanzas, consejos y por confiar en mí.

A mis sinodales, Dra. María de los Ángeles Alonso Lavernia por sus recomendaciones, la motivación brindada y el seguimiento a mi trabajo.
Dr. Ruslan Gabbasov por todas las asesorías, las ideas compartidas y sobre todo, por el interés mostrado en mi investigación.

A la Universidad Autónoma del Estado de Hidalgo (UAEH) por brindarme la oportunidad de realizar mi formación doctoral.

A CONACyT por la beca recibida a través del proyecto 247870.

Al grupo de Tecnologías del Lenguaje del Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), quienes siempre me hicieron sentir parte importante de su comunidad científica. Especialmente, al Dr. Luis Villaseñor Pineda y a mi compañero Adrián Pastor López Monroy por su colaboración.

Al Instituto Tecnológico Superior del Oriente del Estado de Hidalgo (ITESA) por todas las facilidades otorgadas durante mis estudios de doctorado.

Resumen

Los textos son un vehículo para expresar sentimientos, pensamientos y conocimiento. Por lo tanto, se han consolidado como una fuente valiosa de comunicación donde la información transmitida va más allá del contenido de los mismos. Particularmente, la forma en que se relacionan las palabras, que muchas veces se realiza de manera inconsciente, y la propia elección de su uso, integran el estilo de escritura del autor de un texto, y a su vez, señalan sus intereses temáticos. De ahí que, la expresión escrita es explorada por tareas de investigación como la identificación del perfil de autores cuyo objetivo es encontrar patrones lingüísticos que predigan rasgos socio-demográficos del perfil de los autores como: género, edad y personalidad. Desde esta perspectiva, la presente investigación propone un nuevo enfoque para la tarea usando textos provenientes de redes sociales.

Diferente de los métodos actuales, los cuales usan representaciones basadas en la frecuencia de los términos en los documentos, en este trabajo se propone evaluar la calidad de los fragmentos de texto donde ocurren tales términos.

La hipótesis principal estipula que las personas revelan información valiosa de su perfil cuando escriben acerca de ellas mismas, esto es, cuando usan pronombres personales en primera persona del singular. Por lo tanto, en este trabajo se estudia el rol que juegan las frases que contienen tales pronombres en la tarea.

Asimismo, se diseña una novedosa medida llamada índice de expresión personal, la cual cuantifica el grado de expresión personal de cada término dentro de los textos. La medida diseñada es explotada en el proceso de construcción de la representación de los textos mediante dos esquemas propuestos: a) una novedosa técnica de selección de términos llamada pureza personal discriminativa, la cual está dirigida a seleccionar términos muy asociados al contexto personal de los autores, aun cuando su frecuencia sea baja; b) un nuevo esquema de pesado de términos denominado recompensa exponencial de información personal, el cual considera además de la frecuencia, la calidad de los fragmentos de texto donde los términos ocurren, enriqueciendo así, la frecuencia de los términos con información cualitativa de sus contextos.

Las propuestas son evaluadas mediante una serie de experimentos realizados en distintas colecciones de documentos provenientes de redes sociales. Fructuosamente, los resultados evidencian la usabilidad de los esquemas diseñados y demuestran que las frases con pronombres en primera persona del singular contienen un alto valor para obtener información acerca del perfil de los autores. Por otra parte, también se muestra que enfatizando el valor de los

términos contenidos en este tipo de frases, se puede mejorar ventajosamente el desempeño en la identificación de perfiles de autores. Específicamente, se obtuvieron ganancias promedio correspondientes a 7.34% en la identificación de edad y 5.76% en género, con respecto al enfoque que ha reportado el desempeño más alto en las mismas colecciones. Por consiguiente, los esquemas propuestos permiten enriquecer enfoques de análisis de autoría.

Abstract

Documents are the main mean by which people express and share their feelings, thoughts and knowledge. Hence, they have consolidated as a valuable source of communication where the transmitted information goes beyond the content. In particular, the way in which people choose and connect the words indicates their writing style as well as their thematic interests. Therefore, the writing expression is explored in research fields such as author profiling, where the aim is to analyze the texts to find linguistic patterns that predict socio-demographic traits of authors such as: gender, age and personality. In this context, the present research work proposes a new approach for author profiling using text from social media.

Unlike current methods, which mainly use text representations based on term frequency, this work evaluates the quality of text fragments where the terms occur. The hypothesis is that words reveal valuable information for profiling when people write about themselves, that is, when first-person pronouns are

used. Therefore, this work studies the role of phrases containing such pronouns in the task of author profiling.

Furthermore, a novel measure called personal expression index is designed in order to estimate the level of personal information captured by each term of a given text. The new measure is used for constructing the text representation by means of the following schemes: i) a novel feature selection technique, discriminative personal purity, which selects the terms tightly associated to personal contexts even if their frequency is low; and ii) a novel term weighting scheme, Exponential Rewarding of Personal Information, which considers besides the frequency the quality of text fragments where the terms occur, in this way the term frequency is enriched with qualitative information from their contexts.

The proposed approach is evaluated by means of a series of experiments using text collections from social media. The results are very encouraging and show strong evidence about the effectiveness of the proposed schemes. In this way, it is demonstrated that phrases with first-person pronouns convey huge information about authors' profiles. On the other hand, it is shown that emphasizing the value of the terms from this kind of phrases, the performance of profile classification can be improved advantageously as compared to current methods. Specifically, average improvements of 7.34% and 5.76% for age and gender classification were obtained when comparing to the best result from state-of-the-art in the same collections. In conclusion, the proposed schemes are able to enrich the existing approaches for authorship analysis.

Contenido

Lista de figuras

Lista de tablas

Lista de siglas

Abrev.	Significado
AA	*Authorship Attribution*, (Análisis de autoría)
AP	*Author Profiling*, (Identificación del perfil de autores)
BOW	*Bag of words*, (Bolsa de palabras)
DPP	*Discriminative Personal Purity* (Pureza personal discriminativa)
FP	*Personal Phrases*, (Frases personales)
FNP	*Non Personal Phrases*, (Frases no personales)
IG	*Information Gain*, (Ganancia de información)
IM	*Impostors Method*, (Métodos de los impostores)
KNN	*K-Nearest Neighbors*, (K vecinos más cercanos)
MNB	*Multinomial Naïve Bayes*, (Naïve Bayes Multinomial)
NB	Naïve Bayes
NEI	*Non-Personal Expression Index*, (Índice de expresión no personal)
PEI	*Personal Expression Index*, (Índice de expresión personal)
POS	*Part Of Speech*, (Etiquetas de partes de la oración)
PP	*Personal Pronoun*, (Pronombre personal)
SVM	*Support Vector Machines*, (Máquinas de vectores de soporte)
TC	*Text Categorization*, (Categorización de textos)
TF	*Term Frequency*, (Frecuencia del término)
TSS	*Feature Selection Schemes*, (Esquemas de selección de términos)
TWS	*Term Weighting Schemes*, (Esquemas de pesado de términos)
$10FCV$	*10-fold cross validation*, (Validación cruzada de 10 capas)

Capítulo 1

Introducción

Cada vez que alguna persona redacta un texto en lenguaje natural, transmite en él sentimientos, intereses, pensamientos, conocimiento, etc. Por lo tanto, la información expuesta en un texto, las estructuras gramaticales utilizadas en la redacción, las palabras, la forma en la cual son relacionadas, así como la propia elección de su uso, por un lado, integran el estilo de escritura del autor del texto y por el otro, reflejan sus intereses temáticos. En consecuencia, el estilo de redacción y el contenido expresado en un texto conlleva información que refleja rasgos del perfil del autor. Por lo tanto, mediante un análisis del texto se podrían predecir dichos rasgos e incluso se puede revelar la autoría del texto cuando ella se desconoce.

Si bien cada autor transmite parte de su perfil en cada texto que redacta, un grupo de personas que comparten características demográficas como la edad, género, lengua materna entre otros rasgos, comparten también similaridades en el estilo de redacción e incluso intereses temáticos. En otras palabras, los intereses y el estilo de redacción de un grupo de personas que han compartido un contexto social similar puede ser generalizado. En consecuencia, algunos rasgos del perfil de los autores pueden predecirse a partir del análisis de los textos mediante técnicas de *Author Profiling* (AP, por sus siglas en inglés), tarea que puede ser traducida como identificación del perfil de autores. A la fecha, se han propuesto algunos métodos para abordar dicha tarea. Por su parte, esta tesis propone un enfoque novedoso cuya motivación se describe en la siguiente sección.

1.1 Motivación

Específicamente, la tarea AP consiste en analizar textos para predecir atributos demográficos o generales del perfil del autor como: edad, género, personalidad, lengua nativa, orientación política, entre otras. Recientemente, debido a una gran variedad de aplicaciones, AP ha ganado un interés especial. Por ejemplo, en mercadotecnia, conocer el perfil de las personas que gustan o disgustan de un producto puede ayudar a las empresas a crear publicidad dirigida. En lingüística forense, el perfil lingüístico puede generar evidencia adicional para conocer rasgos socio-demográficos del autor de un mensaje de acoso, en tal caso, tal vez no indique quién es el verdadero responsable pero sí orienta la

búsqueda a cierto perfil reduciendo el número de candidatos; además puede apoyar a determinar la veracidad de una carta suicida.

El área de AP es un campo de investigación comúnmente abordado por la comunidad científica como un problema de clasificación de textos, es decir, a partir de un texto se genera una representación manipulable por la computadora. Posteriormente, se usan algoritmos de aprendizaje automático para encontrar similaridades del estilo de escritura y contenido de los textos de un grupo de personas que conforman una muestra de textos escritos por autores etiquetados de acuerdo con los valores de la característica del perfil a predecir (e.g. en la predicción de género: hombres y mujeres). Las similaridades o patrones encontrados conforman el modelo de predicción, el cual posteriormente es aplicado para predecir los rasgos del perfil de nuevos textos.

Ahora bien, la tendencia y contribución de los trabajos de investigación en AP está dirigida a estudiar y/o definir atributos útiles para lograr la representación del texto[1]. Este trabajo, más que definir un conjunto de atributos adecuados para AP, se enfoca en estudiar la relevancia de sentencias que contienen pronombres en primera persona del singular, las cuales serán referidas como frases personales. El interés está motivado por recientes trabajos en psicología, los cuales han demostrado que los pronombres y preposiciones revelan información importante acerca del perfil lingüístico de un autor (Pennebaker, 2011), incluso que la gente tiende a ser más honesta cuando hablan de ellos

[1]De acuerdo con la literatura, dos tipos de atributos son las más relevantes: i) temáticos, principalmente capturadas por nombres, verbos y adjetivos ii) estilísticos, por ejemplo, palabras de función, signos de puntuación y etiquetas POS.

mismos (Newman et al., 2003). Basado en estas contribuciones, la idea base del presente trabajo de investigación indica que las palabras o términos alrededor de los pronombres personales en primera persona expresan mejor los intereses temáticos y estilo de escritura de los autores, por lo tanto, ellas pueden revelar información valiosa para la predección de su perfil del autor.

Tomando en consideración lo anterior y aprovechando la importancia de los términos contenidos en frases personales, en esta tesis se plantea un enfoque novedoso para AP cuyo objetivo es enfatizar el valor de estos términos y en consecuencia obtener mejoras en el desempeño. A través de los resultados de la experimentación, se confirma que las palabras alrededor de un pronombre personal tienen información valiosa, es decir, altamente discriminativa para AP.

1.2 Descripción del problema

En el presente trabajo de investigación se considera la identificación del perfil de autores en redes sociales como una tarea de categorización automática, a través de la caracterización del estilo de escritura y contenido de los textos. La finalidad es inferir patrones del uso del lenguaje que distingan categorías de autores. Sin embargo, la dificultad de generalizar patrones socio-lingüísticos no de una, sino de un grupo de personas (e.g. hombres y mujeres) para predecir rasgos demográficos de los autores, ha propiciado enfoques con resultados aún mejorables. A pesar de esfuerzos numerosos en la tarea y logros conseguidos, aún queda mucho espacio para proponer técnicas y métodos que aumenten los resultados de la clasificación de perfiles. Como un ejemplo representativo, en

la última competencia de AP del Foro internacional PAN 2016[2], se reportó una exactitud [3] promedio global en el idioma inglés de 0.7564 para género y 0.5897 para edad (Rangel et al., 2016) [4]. En AP son muchos los retos a los cuales se enfrenta la comunidad científica, entre ellos destacan:

1. Información no valiosa presente en los textos provenientes de medios sociales (ruido en los datos). En este tipo de medios, los textos pueden contener información que no es propia del autor. Por ejemplo, cuando se repite lo que dicen otros autores (incluso los *retweets*) y la transcripción de información de otras fuentes. Lo anterior indica, que el contenido de otros autores o fuentes puede reflejar los intereses del autor, pero no así, su estilo de redacción. En este contexto, los métodos existentes analizan la información contenida en los textos tratándola con el mismo nivel de importancia, sin considerar que existen fragmentos donde las personas se expresan honestamente y es, en esas secciones donde el estilo de escritura e intereses de las personas podrían estar enfatizados.

2. Dificultad para encontrar patrones temático-estilísticos. La variabilidad de contenidos es una característica presente en textos de medios sociales. Es decir, los usuarios se expresan libremente sobre cualquier tópico, propiciando representaciones del texto con alta dimensionalidad y muy

[2]http://pan.webis.de, AP Task.

[3]Aunque el desempeño depende en gran medida de las condiciones de evaluación como las características de las colecciones de textos, la exactitud mencionada es un punto de referencia para determinar la dificultad y relevancia de la tarea. Existen trabajos donde se reportan valores de exactitud muy altos, pero las colecciones son diferentes.

[4]Los equipos participantes obtuvieron exactitudes en Twitter desde 0.0201 hasta 0.5230 (el mejor resultado) en género y para edad desde 0.0086 hasta 0.3879. En blogs desde 0.4615 hasta 0.6923 para género y 0.3205 hasta 0.5513 para edad.

dispersas[5] que aumentan la dificultad para encontrar patrones temáticos que describan y diferencien perfiles de autores. Adicionalmente, la informalidad de estos medios genera una especie de homogeneización del estilo produciendo representaciones con un vocabulario común de tamaño insuficiente para generar patrones de índole estilística. Por ejemplo, el uso de emoticones se está volviendo muy popular reduciendo su valor discriminativo. Otro ejemplo es el desuso común de acentos. En este contexto, poco se ha explorado sobre las secciones de un texto donde las diferencias entre perfiles de personas son más evidentes. Más aún, hasta el momento, no se ha explorado la existencia de algunas frases en los documentos más discriminativas que otras, lo cual facilitaría el descubrimiento de patrones.

3. La carencia de información cualitativa en los esquemas de selección y pesado de términos. Las estrategias de selección y pesado usadas en AP provienen de recuperación de información (Baeza-Yates y Ribeiro-Neto, 1999; Salton y Buckley, 1988) y en su mayoría usan inferencias estadísticas acerca de las ocurrencias en los documentos sin considerar características cualitativas de esas ocurrencias. Sin embargo, en la identificación del autor dichos esquemas pueden no ser suficientes debido a la naturaleza de los textos destacando: un gran vocabulario, longitudes generalmente cortas, la mayoría aborda los mismos sucesos sociales. Los enfoques actuales (e.g. ganancia de informacion, TF-IDF, etc.) evalúan la

[5]Usualmente, los documentos son representados como vectores de términos cuya dimensionalidad es igual al tamaño del vocabulario (términos distintos) de la colección de documentos. La dispersidad refleja que cada documento contiene un pequeño subconjunto de términos del vocabulario (Joachims, 2001).

frecuencia sin considerar el tipo de contexto de los atributos o el tipo de información que representan.

Para solventar los problemas descritos anteriormente, la presente investigación aborda AP con un enfoque de clasificación supervisada. Específicamente, para enfrentar la comunicación informal y dificultad de encontrar patrones, se considera que la información contenida en los textos no es igualmente relevante para la tarea, destacando una riqueza especial en los términos de las frases personales (que tienen un pronombre personal en primera persona del singular), pues desde una perspectiva psicológica, es ahí donde los autores hablan de ellos mismos expresando honestamente información personal como intereses, preferencias y hábitos, incluso reflejan mejor el estilo de escritura (Pennebaker, 2011). Dicha información es esencial para encontrar diferencias entre categorías de autores.

Asimismo se proponen nuevos esquemas de selección y pesado de términos que enriquecen la frecuencia de los términos al considerar características cualitativas correspondientes al contexto de su ocurrencia. Básicamente, se propone ponderar, como característica cualitativa, la relevancia de los fragmentos donde aparece un término, tomando ventaja de su ocurrencia en frases personales. De esta manera, el poder descriptivo-discriminativo de los términos está calificado por la posibilidad de reflejar información personal del autor.

1.3 Hipótesis

- Las palabras cercanas a un pronombre personal contienen información más valiosa que otros fragmentos del texto para la identificación del perfil de autores.
- La selección y pesado de términos que enfatizan el valor de la información de tipo personal del autor, mejora los resultados de clasificación con respecto a dichas tareas cuando se realizan usando exclusivamente información de la frecuencia o distribución de los términos.

1.4 Objetivos

1.4.1 General

Proponer un nuevo enfoque para identificar automáticamente el perfil de autores en redes sociales basado en esquemas de pesado de términos que enfaticen la información de tipo personal contenida en los documentos.

1.4.2 Específicos

- Analizar el valor de las frases personales para la identificación del perfil de autores en redes sociales contrastando con el resto de las frases.
- Proponer un criterio para facilitar la construcción de colecciones de documentos etiquetadas para identificar el perfil de los autores a través de la evaluación del contenido de los documentos.

- Definir una medida de relevancia para los términos que determine su grado de asociación con el contexto personal del autor.
- Generar un método de selección de términos que enfatice el valor de aquellos fuertemente asociados al contexto personal de los autores.
- Diseñar un esquema de pesado de términos que pondere el tipo de contexto donde aparece el término, enfatizando la información de tipo personal contenida en los textos.
- Evaluar la robustez de los esquemas propuestos de selección y pesado de términos para clasificar perfiles de autores mediante su aplicación en distintas redes sociales y contrastándolos con enfoques actuales del estado del arte.

1.5 Organización de la tesis

El contenido de esta tesis está estructurado de la siguiente forma:

En el Capítulo 2 se presentan los conceptos básicos que introducen al lector dentro del contexto del presente trabajo de investigación. Específicamente, se estudia la tarea conocida como categorización de textos y su aplicación como perspectiva para abordar tareas del análisis de autoría, donde se ubica AP. Primero, se presenta una introducción al concepto formal de categorización de textos. Posteriormente, se describen las tres etapas esenciales en la construcción de un clasificador de textos: la representación del texto, el proceso

de aprendizaje y la evaluación. Finalmente, se introduce al lector en el marco de referencia del análisis de autoría.

En el Capítulo 3 se presenta una revisión general de los trabajos enfocados en la identificación del perfil de autores. Asimismo, se analiza el uso de información personal en AP. Enseguida, se muestran los principales enfoques utilizados en AP para realizar selección de términos y pesado de términos. Posteriormente, se describen las colecciones de documentos utilizadas en las secciones experimentales de los siguientes capítulos.

En el Capítulo 4 se presenta la primer contribución del presente trabajo de investigación. Específicamente, se describe un estudio formal para conocer el rol que tienen las frases con pronombres personales en la identificación del perfil del autor.

En el Capítulo 5 se describe el concepto *indice de expresión personal*, una novedosa contribución de esta investigación. Dicho concepto está dirigido a cuantificar la cantidad de información personal revelada por un término. En este capítulo se expone la base de su creación, así como, sus componentes.

En el Capítulo 6 se ubica otra contribución de esta tesis. Ahí se desarrolla la técnica propuesta de selección de términos, la cual es llamada *pureza personal discriminativa*.

En el Capítulo 7 se expone el esquema propuesto de pesado de términos, el cual es denominado EXPEI, una propuesta novedosa que asigna un valor a cada término de acuerdo con el valor del índice de expresión personal.

En el Capítulo 8 se presentan los resultados del enfoque propuesto, el cual es generado por seleccionar términos usando la técnica propuesta en el Capítulo 6 y pesarlos mediante del esquema propuesto en el Capítulo 7.

Finalmente, se exponen las conclusiones, un resumen de las contribuciones y las direcciones futuras de esta investigación.

Capítulo 2

Análisis de autoría basado en categorización de textos

El Análisis de Autoría (AA) es un área dentro del procesamiento del lenguaje natural orientada a analizar textos para inferir información de su autor e incluso la autoría cuando ella se desconoce. De forma general, las tareas referentes a AA pueden ser ubicadas en tres tipos (Ding et al., 2016): atribución de autoría, verificación de autoría y la identificación del perfil del autor, siendo esta última, la tarea abordada en la presente investigación.

Típicamente, las tareas de AA se han estudiado como un problema de categorización supervisada de textos (*Text Categorization*, TC, por sus siglás

en inglés) mediante técnicas de aprendizaje automático. De ahí que en este capítulo, por un lado, se presenta el marco conceptual de la tradicional TC describiendo las etapas esenciales en la construcción de un clasificador de textos. Por el otro lado, se discute la relación y diferencias de TC con AA. Asimismo, se presenta una introducción a las tareas de AA colocando especial interés en AP, pues es donde se ubican las contribuciones de esta tesis.

2.1 Categorización de textos

La categorización tradicional de textos es también conocida como clasificación de textos. Es importante señalar, que en esta tesis, TC hace referencia a la tarea de clasificación bajo el enfoque de aprendizaje automático mediante la perspectiva supervisada (Mitchell, 1997), la cual es diferente del agrupamiento referido como clasificación no supervisada.

Este apartado está organizado como sigue: primero se presentan los fundamentos del problema de TC (Sección 2.1.1). Posteriormente, se describen las tres fases esenciales que integran el marco de solución común de los sistemas de TC: la representación de los documentos (Sección 2.1.2), el proceso de clasificación (Sección 2.1.3) y la evaluación (Sección 2.1.4).

2.1.1 El problema de categorización de textos

TC es la tarea de etiquetar (categorizar) textos escritos en lenguaje natural con categorías temáticas (temas) (Sebastiani, 2002), donde el conjunto de categorías es definido previamente.

Formalmente, el tradicional problema de TC puede ser definido considerando un dominio de documentos $\mathcal{D} = \{d_1, ..., d_{|\mathcal{D}|}\}$ y de categorías predefinidas $\mathcal{C} = \{c_1, ..., c_{|\mathcal{C}|}\}$ como la tarea de aproximar la función objetivo desconocida $F : \mathcal{D} \times \mathcal{C} \rightarrow \{0, 1\}$ (que indica cómo los documentos deberían ser clasificados) por medio de una función $M : \mathcal{D} \times \mathcal{C} \rightarrow \{0, 1\}$ llamada clasificador o modelo. En tal caso, el valor de la función $F(d_j, c_k)$ es 1 si el documento d_j pertenece a la categoría c_k y 0 en caso contrario. Entonces, la tarea es construir un modelo M que produzca resultados lo más cercanos posible a la función F (Sebastiani, 2002).

Dependiendo del número de categorías en el conjunto $\mathcal{C}$, es decir, $|\mathcal{C}|$, así como del número de categorías que pueden asignarse a cada documento $d_j \in \mathcal{D}$, los problemas de TC pueden ser de diferentes tipos:

Etiqueta única. Un problema de TC es considerado de única etiqueta si a cada documento $d_j \in \mathcal{D}$ se le debe asignar exactamente una categoría c_k. En este caso, si el conjunto de categorías predefinidas contiene dos elementos, es decir, $|\mathcal{C}| = 2$ (e.g., c_k o $\overline{c_k}$; si o no; verdadero o falso) se trata de un **problema binario** (Sebastiani, 2002). Como ejemplo se puede mencionar la clasificación de correos electrónicos para definir si es o no correo basura (*spam*). La categorización binaria es la más simple y a su vez, es la más usada en la literatura para las demostraciones sobre TC. En cambio, cuando $|\mathcal{C}| > 2$, el problema es considerado **multicategoría** (Tsoumakas y Katakis, 2007; Joachims, 2002).

Multi-etiqueta. En un problema multi-etiqueta cada documento $d_j \in \mathcal{D}$ puede recibir cualquier número categorías, desde 0 hasta $|\mathcal{C}|$ (Sebastiani, 2002). Por ejemplo, un historial clínico puede ser etiquetado con varias enfermedades simultáneamente. Según Joachims (2002) a diferencia del caso multicategoría, aquí no hay correspondencia uno a uno entre las categorías y los documentos. En su lugar, cada documento puede tener cero, una o incluso varias categorías. A saber, el caso multi-etiqueta puede ser resuelto por $|\mathcal{C}|$ clasificadores binarios independientes, uno para cada categoría $c_k \in \mathcal{C}$; siempre y cuando, las decisiones para asignar un documento a diferentes categorías sean independientes (Feldman y Sanger, 2006).

Naturalmente, cualquier tipo de TC requiere transformar los textos a una representación que los prepare para el proceso de clasificación. Los detalles se muestran en la siguiente sección.

2.1.2 Representación de los documentos

Aún cuando actualmente los textos se almacenan en formatos digitales (HTML, PDF, PS), los clasificadores no pueden procesar el texto en su forma original (cadenas de caracteres), principalmente porque son redactados en lenguaje natural. Por ello, es necesario convertirlos en una representación formal que preserve sus características relevantes y que a su vez, sea manejable por los algoritmos de aprendizaje. La representación de documentos es un proceso importante que influye significativamente en los resultados de los sistemas TC (Lewis, 1990; Song et al., 2005).

De forma general, la representación de los textos se basa en el modelo de espacio de vectores (Baeza-Yates y Ribeiro-Neto, 1999) cuyo objetivo es representar cada documento como un vector de pesos $\vec{d_j} = \langle w_{1,j}, ..., w_{|T|,j} \rangle$ donde T denota el vocabulario de la colección de documentos representada por D. Prácticamente, el vocabulario es el nombre que se asigna al conjunto de términos, también llamados características que ocurren al menos una vez en por lo menos un documento de la colección D. A su vez, $w_{i,j}$ representa el peso del término t_i en el documento d_j cuantificando cómo el término está contribuyendo en la semántica del documento d_j. Finalmente, la colección de documentos (vectores) puede ser mapeada a una matriz documento-atributo (documento-términos) de dimensión $|D| \times |T|$. Para mayor claridad, esta matriz es representada en la siguiente tabla:

Tabla 2.1: Representación documento-atributo correspondiente a una colección de documentos $D = \{d_1...d_n\}$ y vocabulario de términos $T = \{t_1...t_m\}$

		Vocabulario			
		t_1	t_2	...	t_m
Documentos	$\vec{d_1}$	$w_{1,1}$	$w_{1,2}$	...	$w_{1,m}$
	$\vec{d_2}$	$w_{2,1}$	$w_{2,2}$	...	$w_{2,m}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$\vec{d_n}$	$w_{n,1}$	$w_{n,2}$	...	$w_{n,m}$

En la Tabla 2.1 se observa que la dimensionalidad de la representación (espacio de características) está determinada por el tamaño del vocabulario, es decir, $|T|$. Además, es posible notar que los componentes unidad en la representación de documentos son los términos del conjunto T y su extracción inicia con el preprocesamiento del texto, el cual se describe a continuación.

2.1.2.1 Preprocesamiento

El preprocesamiento es un proceso que consiste en eliminar los elementos textuales que no son informativos y preparar los textos para crear su representación formal. Generalmente, este proceso incluye algunas de las siguientes acciones:

- Remover etiquetas. Comúnmente, las etiquetas XML y/o HTML son eliminadas. De esta manera, sólo se mantiene la información textual del documento.
- Convertir a minúsculas. Usualmente, todos los caracteres escritos en mayúsculas son convertidos a minúsculas antes de la clasificación, ya que se asume que no hay diferencia entre ambas formas (Uysal y Gunal, 2014).
- Remover palabras vacías. Comúnmente, en los problemas de TC, las palabras vacías (palabras muy frecuentes en los textos, pero que no contribuyen en la semántica del documento) se eliminan porque son consideradas partículas que no transmiten información. La lista de palabras vacías generalmente incluye palabras de función: artículos, pronombres, preposiciones, conjunciones, etc.
- Lematizar el texto. Este proceso se aplica para obtener las raíces morfológicas de las palabras, pues se asume que las palabras con la misma raíz describen conceptos muy relacionados en el texto. Por ejemplo, palabras semánticamente similares como caminar, camino, caminaré y caminé serán llevadas a una misma raíz léxica.

Aunque el preprocesamiento podría parecer un proceso sencillo, existen estudios que lo consolidan como una etapa crucial que impacta en el desempeño de la tarea de clasificación (Uysal y Gunal, 2014; Haddi et al., 2013).

Específicamente en AA, donde no sólo el contenido de los textos es considerado, sino también el estilo de redacción de los autores, el preprocesamiento no elimina las palabras de función, pues han destacado como los mejores marcadores de estilo de los autores (Chung y Pennebaker, 2007). Incluso, hay propuestas que exploran el uso de etiquetas HTML como elementos del estilo (Weren et al., 2014b). Particularmente, el enfoque propuesto preprocesa el texto a través de un esquema base común para AP: eliminando etiquetas HTML, convirtiendo a minúsculas todo el texto y manteniendo palabras de función.

Cabe señalar que el preprocesamiento también considera la separación de las unidades con significado que definen a los términos de la representación (e.g., caracteres, palabras y frases). La siguiente sección describe los términos más comunes.

2.1.2.2 Definición de términos

De acuerdo con Feldman y Sanger (2006) una característica o término es simplemente una entidad sin estructura interna, la cual corresponde a una dimensión en el espacio de términos del modelo de espacio vectorial. En este

contexto, la siguiente interrogante se hace presente: ¿qué elementos pueden definirse como términos? La respuesta ha generado diversas representaciones.

Según Joachims (2002), los términos se definen en diferentes niveles, tales como: nivel de palabras (palabras e información léxica, e.g., bolsa de palabras), nivel de sub-palabras (descomposición de las palabras, e.g., n-gramas de caracteres[1]), nivel de multi-palabras (frases e información sintáctica, e.g., frases y n-gramas de palabras[2]), nivel semántico y pragmático (significado del texto con respecto al contexto, e.g., sinónimos, categorías de palabras). A saber, la representación más utilizada en TC es la bolsa de palabras, (*Bag of Words*, BOW, por sus siglas en inglés).

Bolsa de palabras. Específicamente, en BOW los términos (elementos del conjunto T) corresponden a las palabras o sus raíces. A su vez, el peso de cada término es determinado por su frecuencia en el documento. Por lo tanto, cada documento d_j es representado por el conjunto de palabras diferentes en toda la colección de documentos $\mathcal{D}$.

Naturalmente, se han explorado representaciones alternativas más complicadas que BOW, tales como: frases (Li et al., 2009), grupos de palabras (Baker y McCallum, 1998; Bekkerman et al., 2003), n-gramas de caracteres (Rahmoun y Elberrichi, 2007; Peng et al., 2003), etc. Sin embargo, también existen estudios que han concluido que representaciones complejas como frases referentes a nombres propios y multipalabras expresando conceptos

[1] El término n-gramas se refiere a secuencias de n caracteres del texto.
[2] Secuencias de n palabras del texto.

de dominio, no mejoran la exactitud de los enfoques (Moschitti y Basili, 2004). Lo anterior sugiere que las representaciones basadas en palabras son muy eficientes.

Es importante señalar que BOW es una representación simple y produce buenos resultados pero tiene algunos inconvenientes: no considera el orden de las palabras[3] ni las relaciones entre ellas; tampoco trata de entender el contenido de los textos.

Existen nuevas representaciones tratando de manejar esos inconvenientes. Algunos ejemplos son: BOW localmente pesadas (Lebanon et al., 2007), bolsa de conceptos (Sahlgren y Cöster, 2004), análisis semántico conciso (CSA) (Li et al., 2011) y modelado a través de tópicos, entre otros. En este último, se ubican representaciones basadas en análisis semántico latente (LSA) (Landauer et al., 1998; Kuralenok y Nekrest'yanov, 2000), *Latent Dirichlet Allocation* (LDA) (Blei et al., 2003; Phan et al., 2008) o categorías de palabras psicolingüísticas usando *Linguistic Inquiry and Word Count* (LIWC)[4] (Pennebaker et al., 2007; Tausczik y Pennebaker, 2010). No obstante, BOW sigue desempeñándose bien en la práctica. Aunque todavía no se puede concluir que BOW es mejor que estas representaciones sofisticadas, ésta sigue siendo ampliamente utilizada en clasificación de textos, así como en análisis de autoría.

[3]Hay varios intentos por integrar el orden de las palabras como los n-gramas de palabras y las secuencias frecuentes maximales.

[4]Este software estudia componentes emocionales, cognitivos y estructurales contenidos en un texto.

Por otra parte, otro aspecto a considerar en la construcción de la representación es la estimación de los pesos de cada término en los documentos.

2.1.2.3 Pesado de términos

La estimación del peso para cada término, es un proceso que impacta altamente en el desempeño de la clasificación, principalmente, porque los términos contribuyen de diferente forma en la descripción de un documento. Los esquemas en TC para asignar un peso $w_{i,j}$ al término t_i en el documento d_j, se denominan esquemas de pesado de términos (*Term Weighting Schemes*, TWS, por sus siglas en inglés) y, en su mayoría, provienen del área de recuperación de información (Salton y Buckley, 1988). Actualmente, existen varios TWS, los siguientes son los más comunes y a su vez, servirán como esquemas de referencia para comparar el esquema propuesto en esta tesis:

- **Ponderado Booleano.** También conocido como indicador de presencia, es un esquema básico, pero común en TC. Como se muestra en la ecuación 2.1, si el término t_i ocurre en el documento d_j, el esquema asigna el valor 1 y el valor 0 en caso contrario:

$$w_{i,j} = \begin{cases} 1, \ si\ t_i\ aparece\ en\ d_j \\ 0, \ en\ otro\ caso. \end{cases} \tag{2.1}$$

- **Frecuencia del término (TF).** Es un esquema altamente utilizado en TC, indica la frecuencia $TF_{i,j}$ (número de ocurrencias) del término t_i en el

documento d_j:

$$w_{i,j} = TF_{i,j}. \tag{2.2}$$

TF tiene algunas variantes como $\log(TF_{i,j})$, $\log(1 + TF_{i,j})$ donde la idea esencial que relaciona directamente la frecuencia del término con su importancia es mantenida. Es importante mencionar, que TF comúnmente es normalizada por la longitud del documento (número de términos) para evitar un sesgo en los resultados. De manera particular, en la descripción de los experimentos realizados en esta investigación, TF siempre hace referencia a la frecuencia normalizada.

- **Frecuencia del Término - Frecuencia inversa del documento (TF-IDF).** El objetivo de este esquema es asignar pesos con base en la frecuencia de cada término en el documento, así como también, considera el número de documentos en el corpus que contienen el término. Como su nombre lo indica, es una combinación de dos factores a través de una multiplicación: la frecuencia del término en el documento (TF) y la frecuencia inversa de documento (IDF):

$$w_{i,j} = TF_{i,j}\, IDF_j, \tag{2.3}$$

donde el componente $TF_{i,j}$ está descrito en los párrafos anteriores y el componente IDF_j es típicamente calculado según la siguiente expresión:

$$IDF_j = \log\left(\frac{N}{n_j}\right), \tag{2.4}$$

donde N representa el número de documentos en la colección y n_j el número de documentos donde el término j aparece. El componente IDF también tiene otras variantes como $\log(N/n_j) + 1$.

Concretamente, TF-IDF señala que los términos que son muy frecuentes en un documento y también en la colección de documentos, no son términos muy discriminativos. En contraste, los términos con alta frecuencia en un documento y poco ocurrentes en la colección son términos de gran valor para distinguir categorías.

Cabe señalar que han emergido novedosas propuestas de TWS, un estudio comparativo de ellas en el área de TC se encuentra en (Lan et al., 2005). Sin embargo, también es importante mencionar que las TWS mencionadas anteriormente funcionan como estándares e incluso su base (la frecuencia) se ha conservado en algunas nuevas propuestas. Como ejemplos se pueden mencionar los esquemas que reemplazan el componente IDF por técnicas comunes de selección de términos como ganancia de información (IG) y χ^2, creando los conocidos esquemas: TF-IG y TF-χ^2, respectivamente. En este contexto, la presente investigación propone un novedoso TWS para AP, el cual es descrito en el Capítulo 7.

Ahora bien, algunas veces se generan representaciones con un conjunto muy grande de términos, lo cual se traduce en un problema menos tratable para los métodos de aprendizaje. También genera un alto costo computacional que, comúnmente, no mejora significativamente los resultados de la clasificación. En cambio, una buena elección de características (reducción del conjunto de

términos) puede mejorar el desempeño de la clasificación. En la siguiente sección se muestran algunas técnicas que permiten elegir los términos más relevantes y a su vez, reducir la dimensionalidad del espacio de términos.

2.1.2.4 Selección de términos

Uno de los grandes problemas de la categorización de textos es la alta dimensionalidad del espacio de términos, es decir, el conjunto de términos que ocurren en la colección de documentos al menos una vez es muy grande, pese a que no todos ellos son relevantes para la tarea. En este contexto, se ha demostrado que esos términos irrelevantes pueden ser eliminados del espacio de características sin afectar el desempeño del clasificador, incluso, en la mayoría de los casos, se reflejan mejoras importantes.

A menudo el problema de alta dimensionalidad es mitigado seleccionando un subconjunto de T, el cual es integrado únicamente por los términos más relevantes para la tarea. Este proceso se conoce como reducción de dimensionalidad y es realizado por esquemas de selección de términos (*Term Selection Schemes*, TSS, por sus siglas en inglés). Formalmente, el objetivo es reducir el tamaño del vector de características de $|T|$ a $|T'| << |T|$. Por lo tanto, el conjunto T' es conocido como el conjunto de términos reducido.

Algunas veces, sobre todo con las colecciones de documentos son de gran tamaño, se define un umbral de frecuencia para determinar los términos más relevantes que conformarán el vocabulario de la colección, pues se asume que aquellos términos “raros” o poco frecuentes en una colección de

documentos no proporcionan información para predecir las categorías. De esta manera, todos los términos cuya frecuencia se encuentre por debajo de umbral, serán excluidos de la representación y aquellos que lo superen conformarán el vocabulario de la colección. Lo anterior, propicia un problema más tratable computacionalmente sin afectar el desempeño. Aunque su simplicidad es evidente, el uso de un umbral es muy recurrido, pues algunos métodos de selección pueden eliminar términos frecuentes por no considerarlos importantes. En los experimentos de esta investigación, se empleo un umbral de frecuencia para definir el vocabulario de las colecciones y posteriormente se aplican técnicas de reducción de dimensionalidad.

En TC, se han propuesto novedosas TSS, las cuales, generalmente, son enmarcadas en dos grupos según el uso de información proveniente de las etiquetas de las categorías:

- Supervisadas, las cuales aprovechan información estadística de las etiquetas de las categorías del conjunto de entrenamiento para seleccionar términos discriminativos que distingan documentos de diferentes categorías.
- No supervisadas, las cuales prescinden de la información de las etiquetas de las categorías.

En particular, en esta tesis se propone un nuevo TSS de tipo supervisado, el cual es comparado con ganancia de información, un esquema también supervisado comúnmente usado en tareas de categorización, su descripción se presenta enseguida.

Ganancia de información

Es una de las técnicas tradicionales para encontrar los términos más valiosos dentro de una colección de documentos (Mitchell, 1997; Sebastiani, 2002). Ganancia de información (*Information Gain*, IG, por sus siglas en inglés) cuantifica la calidad de los atributos para separar el conjunto de entrenamiento de acuerdo con las categorías. Formalmente, la ganancia de información para un término t_i y un conjunto de $|C|$ categorías, $IG(t_i)$, se calcula a través de la siguiente ecuación:

$$IG\left(t_i\right) = -\sum_{k=1}^{|C|} P\left(c_k\right) \log\left(P\left(c_k\right)\right) + P(t_i) \sum_{k=1}^{|C|} P\left(c_k|t_i\right) \log\left(P\left(c_k|t_i\right)\right) + P(\overline{t_i}) \sum_{k=1}^{|C|} P\left(c_k|\overline{t_i}\right) \log\left(P\left(c_k|\overline{t_i}\right)\right), \tag{2.5}$$

donde $P\left(c_k\right)$ es la probabilidad de la categoría c_k, $P(t_i)$ es la fracción de documentos que contiene el término t_i, mientras $P(\overline{t_i})$ denota la fracción de documentos que no lo contienen. $P\left(c_k|t_i\right)$ es la probabilidad de la categoría c_k dado que el documento contiene el término t_i. Por último, $P\left(c_k|\overline{t_i}\right)$ es la probabilidad condicional de que un documento pertenezca a la categoría c_k dado que el documento no contiene el término t_i.

Ganancia de información cuantifica el poder discriminativo de un atributo por conocer la presencia o ausencia del término en los documentos. En otras palabras, estima qué tan común es un término en una categoría particular comparado con qué tan común es en las otras categorías. Como puede observarse, un valor de IG es calculado para cada término del vocabulario.

De esta manera, es posible reducir del vocabulario integrando únicamente los términos con ganancia de información mayor a cero o incluso, los n términos con mayor IG, donde n es definido acorde a las necesidades del problema.

Como se ha observado, la construcción de una representación formal de los documentos involucra, esencialmente, un proceso de selección y pesado de términos. Posteriormente, la representación resultante, alimenta un proceso inductivo que construye un modelo o clasificador, el cual se detalla a continuación.

2.1.3 Proceso de clasificación

Desde el enfoque de aprendizaje automático (Mitchell, 1997), un proceso conocido como aprendizaje observa las características de un conjunto de documentos que han sido previamente clasificados manualmente por un experto en el dominio, bajo c_k ó $\overline{c_k}$. El proceso de aprendizaje deduce las características que un nuevo documento debería tener para ser clasificado como c_k o $\overline{c_k}$. Por lo tanto, se requiere la existencia de una colección inicial de documentos previamente clasificados $\Omega = \{d_1, ..., d_{|\Omega|}\}$ también conocida como conjunto de entrenamiento, tal que $\Omega \subset \mathcal{D}$. Entonces los valores de la función $F : \mathcal{D} \times \mathcal{C} \rightarrow \{0, 1\}$ son conocidos para cada par $\langle d_j, c_k \rangle \in \Omega \times \mathcal{C}$. Un documento d_j es llamado ejemplo positivo de c_k si $F(d_j, c_k) = 1$ y un ejemplo negativo de c_k si $F(d_j, c_k) = 0$. Precisamente, debido a la dependencia del conjunto de entrenamiento, el proceso se denomina aprendizaje supervisado.

Los modelos construidos son usados para asignar automáticamente una categoría a un nuevo documento de varías categorías posibles previamente definidas. El proceso inductivo de aprendizaje se realiza mediante algoritmos de aprendizaje automático supervisado, coloquialmente llamados clasificadores.

A continuación se describen los clasificadores tradicionalmente usados en TC, así como en AA. Los algoritmos mencionados han mostrado un buen desempeño y serán objeto de comparación en la presente tesis.

2.1.3.1 Clasificadores basados en ejemplos

Este tipo de métodos realizan la clasificación con base en la similaridad entre el conjunto de documentos de entrenamiento Ω y el documento a ser categorizado (Mitchell, 1997). Debido a que postergan su decisión hasta la llegada de un nuevo documento se conocen como métodos flojos (*lazy*). El ejemplo más notorio de este tipo de clasificadores es KNN (K vecinos más cercanos, donde $K \in \mathbb{N}$) (Cover y Hart, 2006; Mitchell, 1997).

El algoritmo KNN asume que todos los documentos corresponden a un punto en un espacio n-dimensional (n corresponde al número de atributos en la representación de los textos). Básicamente, a cada documento sin clasificar se le asigna la categoría de sus K vecinos más cercanos. Para decidir si un documento d_j pertenece a la categoría c_k, KNN revisa si los K documentos de entrenamiento más similares a d_j pertenecen a c_k, si la respuesta es positiva

(mediante el voto mayoritario de los K vecinos más cercanos) se asigna una categorización positiva; de lo contrario una decisión negativa es otorgada.

Particularmente, en el caso de 1NN, cada documento cuya categoría se desconoce, será etiquetado con la categoría de su vecino más cercano. Para ilustrar esto, en la Figura 2.1 se muestra una representación de 1NN con documentos cuya categoría es positiva o negativa. En este caso, d_q denota al documento desconocido a clasificar y su vecino más cercano corresponde al documento de color verde con categoría negativa. En consecuencia, d_q será etiquetado con la misma categoría (negativa).

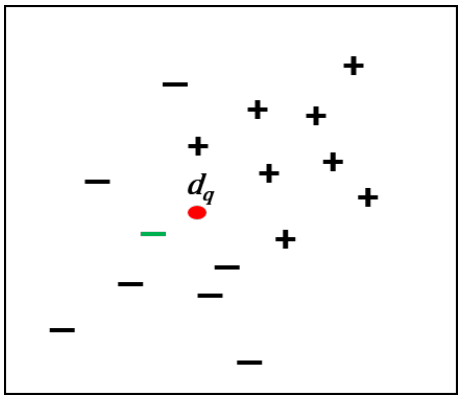


Figure 2.1: Representación de clasificación con 1NN

La cercanía o similaridad entre los documentos es calculada a través de medidas de distancia como son: distancia Euclideana (ampliamente usada), distancia Minkowsky, producto punto y similaridad coseno, entre otras. Un estudio comparativo entre varias medidas puede ser encontrado en el trabajo realizado por Chomboon et al. (2015). Muy recientemente, Kocher y Savoy (2017) publicaron otro estudio donde evaluaron 24 medidas de distancia específicamente en la tarea AP. Entre ellas se encuentran las tradicionales

mencionadas anteriormente, así como la distancia Matusita y la medida Clark (la cual reportó mejores resultados que las demás). En ese estudio, también se señalan las propiedades teóricas que deben presentar las medidas de distancia, encontrándose que la distancia Tanimoto y Matusita cumplen todas ellas.

Debido a que la medida de distancia usada, así como el valor de K impactan en los resultados de la clasificación, es de esperarse la discusión en la literatura sobre el valor de ambos parámetros (Enas y Choi, 1986; Hu et al., 2016; Chomboon et al., 2015). Existen estudios que concluyen que la distancia euclidiana se desempeña razonablemente bien y su elección depende de las características de las colecciones de entrenamiento (Hu et al., 2016; Chomboon et al., 2015). Por otro lado, se ha determinado que los valores altos de K no siempre producen un mejor de desempeño que los valores pequeños (Hand y Vinciotti, 2003). Más aún, se ha mencionado que $K = 1$ genera resultados razonables que pueden ser usados como punto de referencia para evaluar otros clasificadores (Hu et al., 2016).

Aunque KNN, es un método robusto, en el sentido que no requiere que las categorías estén linealmente separadas, tiene la desventaja de un alto costo computacional de clasificación, pues, para cada documento en el conjunto de entrenamiento se calcula la similaridad con todos los otros documentos (Feldman y Sanger, 2006). Lo anterior consume bastante tiempo de procesamiento, sobre todo con representaciones de alta dimensionalidad. En este contexto, Bhatia y Vandana (2010) realizaron una comparación de

diferentes variantes de KNN propuestas para ganar velocidad o eficiencia, concluyendo que cada variante tiene fortalezas en un campo en particular bajo circunstancias particulares.

2.1.3.2 Clasificadores probabilísticos

Los clasificadores probabilísticos construyen un modelo que cuantifica la relación entre los atributos y la categoría como una probabilidad (Aggarwal, 2015). Ejemplos de este tipo de clasificadores son: Naïve Bayes y regresión logística. Estos clasificadores se basan en la aplicación del teorema de Bayes para estimar la probabilidad condicional de una categoría c_k dado un documento d_j, $P(c_k|d_j)$, según la siguiente ecuación:

$$P(c_k|d_j) = \frac{P(d_j|c_k)\, P(c_k)}{P(d_j)}. \tag{2.6}$$

Entre los algoritmos probabilistas frecuentemente usados en los problemas de TC se encuentra Naïve Bayes (NB) (Lewis, 1998; Zhang y Gao, 2011). Este método explota la probabilidad conjunta de los términos y las categorías para estimar la probabilidad $P(c_k|d_j)$ de cada categoría $c_k \in C$, dado un documento d_j. Es importante mencionar que NB asume independencia de los términos que componen un documento $d_j = \{t_1, t_2 ... t_{|T|}\}$.

De forma general, hay dos modelos del clasificador NB. Por un lado, el clasificador Naïve Bayes Simple, expresado en la ecuación 2.7, considera la probabilidad de aparición de cada término dada la categoría de forma binaria, es decir, el término aparece o no y entonces su probabilidad condicional dada

la categoría es o no considerada (Zhang y Gao, 2011):

$$c(d_j) = \arg\max_{c_k \in C} P(c_k) \prod_{i=1}^{|T|} P(t_i|c_k). \tag{2.7}$$

Por el otro lado, el clasificador Naïve Bayes Multinomial (MNB), el cual es expresado en la ecuación 2.8, considera el número de apariciones del término para evaluar la contribución de la probabilidad condicional dada la categoría, con lo que el modelado de cada documento se ajusta mejor a la categoría a la que pertenece:

$$c(d_j) = \arg\max_{c_k \in C} P(c_k) \prod_{i=1}^{|T|} P(t_i|c_k)^{TF_{i,j}}. \tag{2.8}$$

Específicamente, para TC, se ha encontrado que el desempeño del modelo multinomial es superior al modelo simple; pues mientras Naïve Bayes basa sus cálculos en la ausencia o presencia de una palabra, MNB captura información de la frecuencia de las palabras en el documento. Una descripción detallada de las diferencias entre ambos se encuentra en (McCallum y Nigam, 1998).

2.1.3.3 Máquinas de vectores de soporte

Las máquinas vectores de soporte (*Support Vector Machines*, SVM, por sus siglas en inglés) sin duda, han mostrado gran habilidad para trabajar con un espacio altamente dimensional y con datos dispersos, características presentes en problemas de TC. En consecuencia, las SVM son clasificadores con alto desempeño en TC. Además, son menos sensitivas a características irrelevantes que otros clasificadores. Incluso se ha mencionado que la selección de atributos tiene poco impacto usando este algoritmo (Joachims,

1998; Leopold y Kindermann, 2002) principalmente, porque las máquinas están preparadas para trabajar con representaciones de alta dimensionalidad. En términos geométricos, SVM tienen por objetivo básico encontrar hiperplanos que separen los ejemplos (documentos) de las diferentes categorías con el más grande margen, tal como se muestra en la Figura 2.2. Precisamente, los elementos situados a ambos lados del hiperplano óptimo que definen el margen son conocidos como vectores de soporte.

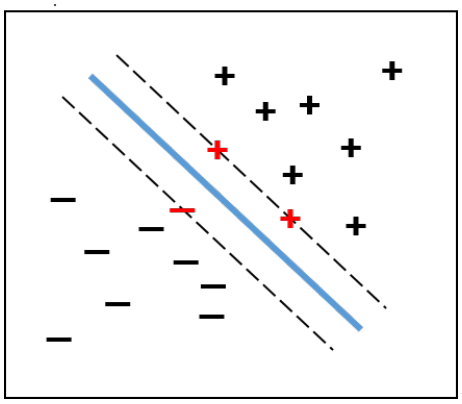


Figure 2.2: Representación de SVM. Los elementos rojos representan los vectores de soporte de las categorías positiva y negativa (+,-). La línea azul representa el hiperplano de separación.

Las SVM pueden ser lineales cuyo objetivo es separar los documentos del conjunto de entrenamiento Ω de forma lineal según las diferentes categorías. En otro caso, las SVMs no lineales pueden ser formuladas a través de funciones conocidas como *kernels*, donde el objetivo es transformar el espacio de atributos iniciales en otro que permita una separación lineal de documentos. Ejemplos de estos *kernels* son las funciones polinomiales y de base radial.

En TC se ha encontrado que las SVM lineales obtienen mejores resultados que las otras SVM con funciones *kernel* más complicadas (Yang y Liu, 1999; Lan et al., 2005). Inclusive se ha señalado que las SVM no requieren

una configuración de parámetros, porque ellas aprenden la configuración automáticamente (Joachims, 1998). Sin embargo, también se ha concluido que las características de la representación del texto dominan el desempeño de TC más que las funciones kernel (Leopold y Kindermann, 2002).

Existen varias implementaciones de software acerca de SVM. Una lista comparativa puede ser consultada en el trabajo realizado por Danenas y Garsva (2011). Algunos ejemplos de las implementaciones son: SVM-Light (Joachims, 1999), LIBSVM (Chang y Lin, 2011) y LIBLINEAR (Fan et al., 2008). Esta última es usada en el presente trabajo de investigación.

Es necesario mencionar que la elección de un clasificador está estrechamente relacionada con la evaluación de su desempeño.

2.1.4 Evaluación en clasificación de textos

Varias medidas y estrategias han sido utilizadas para evaluar el desempeño de un clasificador. El objetivo es calificar su habilidad para tomar las decisiones correctas de categorización. A continuación se describen las medidas y estrategias de evaluación más utilizadas en sistemas de TC.

2.1.4.1 Medidas de desempeño

El desempeño de un clasificador es evaluado analizando sus decisiones. Para un conjunto de documentos a categorizar, las decisiones del clasificador para cada categoría son ubicadas en cuatro grupos, los cuales están representadas en la conocida matriz de confusión representada a través de la Tabla 2.2:

- Verdaderos positivos (tp): número de documentos correctamente asignados a la categoría.
- Verdaderos negativos (tn): número de documentos correctamente reconocidos como no pertenecientes a la categoría.
- Falsos positivos (fp): número de documentos incorrectamente asignados a la categoría.
- Falsos negativos (fn): número de documentos no reconocidos como pertenecientes a la categoría.

Tabla 2.2: Matriz de confusión para una tarea de clasificación binaria

Categoría	Clasificado como *positiva*	Clasificado como *negativa*
positiva	tp	fp
negativa	fn	tn

A través de los valores anteriores, se definen medidas comúnmente empleadas para estimar el desempeño de un clasificador binario:

- Precisión. Fracción de predicciones positivas correctas. Esta medida corresponde a la probabilidad de que un documento d_j sea etiquetado con la categoría *positiva*, y realmente pertenezca a esta categoría:

$$precisión = \frac{tp}{tp + fp}. \tag{2.9}$$

- Cobertura[5]. Fracción de predicciones verdaderas positivas. Esta medida corresponde a la probabilidad de que si un documento d_j debería ser asignado a una categoría *positiva*, esta decisión sea tomada:

$$cobertura = \frac{tp}{tp + fn}. \tag{2.10}$$

- Exactitud. Porcentaje de predicciones correctas:

$$exactitud = \frac{tp + tn}{tp + fp + tn + fn}. \tag{2.11}$$

Para evaluar el promedio de desempeño entre categorías (evaluación global), generalmente se estima el macro-promedio, el cual consiste en calcular primero las medidas de rendimiento por categoría (precisión, cobertura, exactitud) y posteriormente, promediar esos valores. Habitualmente, el promedio de la exactitud es la medida elegida en la literatura para reportar el desempeño de un clasificador.

2.1.4.2 Estrategias de evaluación

Aplicando las medidas de desempeño, se puede estimar la calidad de un clasificador para categorizar nuevos documentos. Normalmente, los documentos de la colección son separados en dos grupos: entrenamiento y prueba. Para evitar selecciones de datos sesgadas, comúnmente se repite el proceso varias veces con diferentes muestras aleatorias. Generalmente, se

[5]Este término proviene de la palabra *recall* en el idioma inglés. En español existen varias acepciones para referirse a este término. Entre ellas se tienen: recuerdo, alcance, cobertura, evocación y recubrimiento. En particular, en esta tesis se utilizará la acepción cobertura.

utiliza una técnica conocida como validación cruzada, la cual consiste en dividir los datos en q particiones disjuntas del mismo tamaño. Una de las particiones se usa para probar y las particiones restantes $(q-1)$ se emplean para entrenar. El proceso es repetido q veces con cada una de las posibles particiones de prueba y se reporta la media de los resultados de cada iteración. Usualmente, al igual que en la presente investigación, esta técnica se realiza con $q = 10$, mejor conocida como validación cruzada de 10 capas (*10-fold cross validation*, 10FCV).

Otra estrategia de evaluación de un sistema de aprendizaje es el método conocido como *leave-one-out cross validation*, el cual consiste en realizar validación cruzada con cada uno de los datos. Se aisla un dato del conjunto, el resto se usa para entrenar y el modelo se prueba con el dato aislado; el proceso se repite para cada dato del conjunto. Sin embargo, es un proceso costoso y usado en menor medida.

2.1.4.3 Comparación de clasificadores: pruebas de significancia estadística

Recientemente, las pruebas de significancia se han utilizado para comparar estadísticamente el desempeño de un conjunto de clasificadores. La idea es determinar si la diferencia en los resultados fue significativa, indicando así, si un algoritmo se desempeñó mejor que otro(s). De esta manera, se indica que el resultado es superior al obtenido de manera azarosa. La aplicación de tales pruebas depende del número de clasificadores a comparar. Enseguida,

se describen brevemente las pruebas de significancia estadística usadas en la experimentación de esta investigación.

Comparación de dos clasificadores. En este contexto, las pruebas de significancia estadística toman como entrada dos clasificaciones, generalmente producidas por dos clasificadores independientes. Las siguientes son dos pruebas pertenecientes a este caso:

- **La prueba pareada t (*paired t-test*)**. Esta prueba estima si la diferencia del promedio del desempeño sobre los conjuntos de datos es significativamente diferente de cero (Dietterich, 1998). En este contexto, Demšar (2006) explica que si c_i^1 y c_i^2 representan estimaciones del desempeño de dos clasificadores sobre el i-ésimo conjunto de N conjuntos de datos y d_i es la diferencia $c_i^1 - c_i^2$, entonces, el estadístico de prueba t es calculado como $\overline{d}/\sigma_{\overline{d}}$, donde el denominador corresponde al error estándar de las diferencias de la media. El estadístico t es comparado con la distribución t de Student con $n-1$ grados de libertad. La prueba es exacta si las poblaciones son normalmente distribuidas, en caso de no seguir la distribución normal, la prueba puede ser considerada aproximada (Kanji, 2006) produciendo resultados útiles para comparación. Similar a enfoques propuestos en AP (Weren et al., 2014b; Rangel et al., 2015), en esta investigación esta prueba fue utilizada en la primera parte experimental (Sección 2.1.4.3).

- **La prueba del rango con signo de Wilcoxon**. Es una prueba no paramétrica que compara y ordena las diferencias en los desempeños de

dos clasificadores para cada conjunto de datos, ignorando los signos. La prueba trabaja con las posiciones ordenadas que ocupan las diferencias d_i de cada i-ésimo conjunto de N conjuntos de datos (Demšar, 2006). Más específicamente, considere R^+ y R^- la suma de las posiciones con diferencias positivas (el primer algoritmo supero al segundo) y negativas (el caso contrario), respectivamente. Entonces, $T = \min(R^+, R^-)$ cuyos valores críticos pueden ser encontrados en las tablas para la prueba Wilcoxon. Esta prueba es una de las preferidas en el estado del arte de AP.

Comparación de múltiples clasificadores. Las pruebas de múltiples clasificadores permiten evaluar las diferencias entre más tres o más algoritmos usando el mismo conjunto de datos. De esta manera, se puede deducir si la diferencia del desempeño de los algoritmos es significativa. Las pruebas más conocidas son Anova para pruebas paramétricas y de Friedman para pruebas no paramétricas. Si la hipótesis nula[6] es rechazada, comúnmente se aplica un procedimiento *post-hoc* para encontrar cuáles clasificadores tienen un desempeño significativamente diferente. Ejemplo de estos procedimientos son: Bonferroni, Holm y Hochberg.

Cuando el número de algoritmos es pequeño, se recomienda usar una variante de la prueba de Friedman: *Friedman's Aligned Rank* (García et al., 2010). En esta investigación esta prueba es utilizada acompañada del procedimiento *post-hoc* Holm. El objetivo es comparar el desempeño del

[6] En la comparación, la hipótesis nula establece que las medias (promedios) de los resultados de dos o más algoritmos son las mismas.

enfoque propuesto mediante varios algoritmos de clasificación a través de las diferentes colecciones de datos (Sección 8.3.2) . Cabe señalar que hasta el momento, no se tiene evidencia del uso de esta prueba en AP. Sin embargo, su uso es altamente recomendable, pues de acuerdo con García et al. (2010), esta prueba ofrece mejores resultados que la prueba de Friedman cuando la comparación involucra un número de algoritmos no mayor a cinco.

Una vez establecido el marco común de TC que sustenta los enfoques de solución para las tareas de AA, se procede a describir el contexto del análisis de autoría.

2.2 Análisis de autoría

En sus inicios AA se enfocó a descubrir el autor de un texto dentro un conjunto de autores candidatos. Posteriormente, surgió el interés de conocer o verificar no sólo la autoría de un texto, si no también, las características del autor (AP).

El origen de la investigación sobre AA asistida por computadora se concentra en los intentos de cuantificar automáticamente el estilo de escritura, a través del estudio realizado por Mosteller y Wallace (1984) sobre la disputa de autoría de 12 artículos federalistas (*federalist papers*) (Coulthard y Johnson, 2007). En este problema, Alexander Hamilton y James Madison reclamaban la autoría de 12 ensayos anónimos publicados para convencer a ciudadanos de aceptar la Constitución de EUA. El trabajo de Mosteller y Wallace (1984) se basó en un sencillo análisis estadístico de la frecuencia de palabras simples como *and* y

to. Los resultados obtenidos fueron significativamente discriminantes entre los escritores, permitiendo atribuir la autoría a Madison.

A partir de ese acontecimiento, surgieron varios trabajos (Holmes, 1998; Rudman, 1998) tratando de definir características para cuantificar el estilo de escritura. Sin embargo, los métodos propuestos en esa época, estaban enfocados únicamente a la identificación del autor y en su mayoría, necesitaban grandes colecciones de datos (libros enteros) y un número de autores demasiado pequeño; además, algunas veces los resultados eran controversiales (Stamatatos, 2009). Concretamente, esos trabajos son sustentados en métodos estadísticos y por ello, son conocidos como asistidos por computadora.

Hoy en día, los avances de las técnicas de aprendizaje automático, lingüística computacional y procesamiento del lenguaje natural, han permitido el desarrollo de sistemas basados en computadora que exploran incluso la predicción del perfil del autor. Generalmente estos sistemas, comparten un marco similar con TC al resolver tareas de AA.

2.2.1 Relación con clasificación de textos

Actualmente, las tareas automáticas relacionadas con AA se tratan como problemas supervisados de TC, de tipo multicategoría y de etiqueta única[7]

[7]Contextualmente, pocas veces se ha tratado como un problema multi-etiqueta. Por ejemplo, se expresa así en el trabajo de Posadas-Durán et al. (2015), pero técnicamente entrenan un clasificador para cada etiqueta y la predicción para una instancia es la unión de la salida de cada clasificador.

(Stamatatos, 2009). Las principales diferencias con la TC tradicional se observan en los siguientes aspectos:

- Tipo de etiquetas a asignar. Mientras en TC se asigna a cada documento de texto etiquetas concernientes a tópicos, en AA las etiquetas corresponden a autores o a rasgos del perfil del autor (e.g., hombre y mujer).
- El enfoque considerado. Mientras TC es considerado un problema de clasificación temática, el enfoque aplicado en AA corresponde a una clasificación basada en estilo. En consecuencia, tanto el estilo como el contenido son usados en AA.
- Tipo de atributos. En AA se han considerado atributos de distinta naturaleza con la finalidad de capturar el estilo de redacción así como los intereses temáticos de los autores, distinguiéndose el uso de las palabras de función. Mientras en TC, estas palabras son eliminadas, en los problemas de AA son incluidas, pues se relacionan con el estilo de redacción de los autores (Argamon y Levitan, 2005; Zhao y Zobel, 2005; Segarra et al., 2013).
- Dimensionalidad y longitud de los textos. Usualmente, la combinación de atributos de distinta naturaleza en AA ocasiona representaciones dispersas de alta dimensionalidad. En adición, en AA los textos son comúnmente cortos, lo cual se ha traducido en un reto para la comunidad científica. En contraste, estas características no son comúnmente observadas en los enfoques de TC.

Concretamente, en la literatura relacionada a AA, varias características han sido combinadas para representar los textos. Sin embargo, se han distinguido dos tipos: características de contenido y de estilo. Las primeras están relacionadas con los intereses temáticos de los autores y corresponden a nombres, sustantivos, adjetivos, etc. Por su parte, las segundas, corresponden a elementos que representan el estilo de redacción de un autor, como: pronombres, palabras de función, etc. Una lista completa de características se encuentra en el trabajo realizado por Stamatatos (2009).

Con respecto al uso de algoritmos de clasificación, no existe una diferencia sustancial entre TC y AA. Sin embargo, dado que en AA se trabajan con representaciones de alta dimensionalidad e inclusive dispersas, las SVM han destacado ampliamente por su habilidad para abordar representaciones con tales características (Houvardas y Stamatatos, 2006; Escalante et al., 2011; Mukherjee y Liu, 2010).

2.2.2 Tareas del análisis de autoría

De manera general, las tareas del análisis de autoría pueden ser categorizadas en tres tipos:[8] (Ding et al., 2016) atribución de autoría, verificación de autoría e identificación del perfil de autores. Cada una de estas tareas es descrita en las siguientes secciones, mostrando un interés especial en AP, pues es la tarea donde se ubican las contribuciones de la presente investigación.

[8]Recientemente ha surgido una tarea relacionada con el análisis de autoría conocida como *Author obfuscation* (ofuscación del autor) Rosso et al. (2016), la cual está dirigida a parafrasear automáticamente un texto determinado para ocultar el estilo de redacción de los autores.

2.2.2.1 Atribución de autoría

La tarea correspondiente a atribucion de autoria consiste en determinar, de un conjunto de candidatos, el autor de un documento dado. De acuerdo con Stamatatos (2009), en un problema de atribución de autoría, existe conjunto de autores candidatos, un conjunto de textos "muestra" de autoría conocida (corpus de entrenamiento) y un conjunto de textos de autoría desconocida (corpus de prueba), donde cada documento deberá ser atribuido a un autor candidato. Cómunmente, este problema es tratado como un problema de clasificación multicategoría, donde cada categoría es un autor candidato.

Por ejemplo, Halteren (2004) extrajo características léxicas y sintácticas de un conjunto de 42 ensayos escritos en alemán correspondientes a 8 estudiantes. En su propuesta desarrolló un método que permite: 1) atribuir ensayos al estudiante cuando olvidó colocar su nombre y 2) detectar estudiantes fraudulentos que copian el ensayo de otro estudiante. Por otra parte, Escalante et al. (2011) usaron histogramas locales sobre trigramas a nivel de caracteres para capturar información secuencial (posiciones de los términos en documentos) del estilo de escritura de un autor. Los autores emplearon SVM para atribuir autoría a un subconjunto de noticias de *Reuters* correspondiente a 10 autores.

Varios métodos novedosos han sido propuestos sobre atribución de autoría. Un estudio detallado de éstos puede ser encontrado en los trabajos de Stamatatos (2009), así como de Bouanani y Kassou (2014). Se debe agregar que nuevas subtareas han emergido, como el agrupamiento automático de documentos

de un mismo autor y la subtarea conocida como *Diarization* cuyo objetivo es identificar diferentes autores dentro de un sólo documento (Rosso et al., 2016).

2.2.2.2 Verificación de autoría

Comúnmente, los trabajos enfocados a verificar autoría (Halteren, 2004) tratan la tarea como un problema de clasificación dual. Si se desea conocer si un texto fue escrito por el autor A dentro de un conjunto cerrado de autores, este enfoque corresponde a preparar una mezcla compuesta de trabajos escritos (ejemplos de entrenamiento) por A y por los otros autores. De esta manera, se crea un modelo para discriminar entre el A y el resto del conjunto de autores. Sin embargo, la clasificación dual puede ser una solución un tanto irreal, porque generalmente no se tiene un conjunto cerrado de autores. Por ello, también se ha usado la técnica conocida como *one-class classification* (clasificación con una única categoría) aplicada por Koppel y Schler (2004). Por ejemplo, el trabajo de Stein et al. (2011) usa esta técnica para detectar plagio intrínseco en documentos.

Recientemente, el método de los impostores (*Impostor Method*, IM, por sus siglas en inglés) introducido por Koppel y Winter (2014) ha logrado un auge importante. En la competencias del PAN 2013 y 2014, los enfoques ganadores en verificación de autoría usaron modificaciones de este método. Fundamentalmente, IM transforma el problema de la forma *one-class* a un problema de clasificación multicategoría (multiclase) mediante la recolección de

textos impostores (e.g., usando un motor de búsqueda) y estimando similiaridad entre parejas de documentos[9].

2.2.2.3 Identificación del perfil de autores

La tarea conocida como AP puede ser traducida al español como: identificación del perfil de autores, caracterización del perfil de autores, perfilado de autores e incluso clasificación de perfiles. AP está dirigido a inferir tanta información como sea posible de los autores (rasgos que integran su perfil) tan sólo por analizar el texto que escribe. Entre los rasgos más comunes que pueden predecirse del perfil de un autor se encuentran: edad, educación, género, personalidad, nacionalidad, lenguaje de origen, etc.

Es importante notar que en contraste con atribución de autoría y verificación de autoría, en AP no se tienen textos de autores conocidos, sino de grupos que comparten rasgos de su perfil (género, edad, personalidad, etc.). La idea es explotar la observación sociolingüística en la escritura de diferentes grupos de personas, lo cual propicia mayor dificultad en la tarea. No obstante, se han generado conclusiones importantes a partir de algunos resultados, por ejemplo, se ha declarado que las mujeres tienden a usar pronombres frecuentemente y los hombres utilizan más representaciones numéricas (Sumit et al., 2009).

Una de las características distintivas de los enfoques de AP reside en la combinación propuesta de los términos que conforman la representación de los

[9]Si frecuentemente, los documentos cuestionados tienen la más alta similaridad comparado con los textos impostores, entonces, se establece que los documentos son escritos por el mismo autor.

textos. La Tabla 2.3 muestra una clasificación de tales términos de acuerdo con el tipo de información que representan. Se ha explorado una gran diversidad de términos de distinta naturaleza, pues el objetivo es capturar el estilo de redacción e intereses de grupos de personas que comparten rasgos de su perfil. Algunos tipos de términos corresponden directamente al contenido de los textos, tal es el caso de los n-gramas léxicos. En cambio, otros tipos se derivan de un estudio más sofisticado del contenido de los textos, por ejemplo, los n-gramas sintácticos o las medidas de recuperación de información.

Precisamente, esa búsqueda de la combinación adecuada de atributos ha marcado la tendencia de la investigación en esta disciplina. En este contexto, los atributos de segundo orden han surgido como una nueva propuesta que ha impactado favorablemente en los resultados de la tarea (López-Monroy et al., 2013, 2015; Rangel et al., 2016).

Sobre los algoritmos de clasificación, las SVM han destacado (Houvardas y Stamatatos, 2006; Escalante et al., 2011; Mukherjee y Liu, 2010). Aunque existen otros algoritmos, no muy comunes, que también ofrecen resultados satisfactorios en AP como: Regresión Bayesiana Multinomial (Argamon et al., 2009) y *Multi-Class Real Winnow* (Schler et al., 2006). Naïve Bayes también ha mostrado un desempeño competitivo en tareas de AP Goswami et al. (2009). Recientemente el modelo multinomial (MNB) ha ganado importancia en las tareas de AP, aunque con un desempeño a veces inferior a SVM. Por ejemplo, en el trabajo de Villena Román y González Cristóbal (2014) se describe un software para identificar edad y género, ahí los autores reportan

Tabla 2.3: Características estilométricas comunes en AP. Adaptación de (Reddy et al., 2016)

Tipo	Ejemplos
Basadas en carácter	Número total de caracteres, letras en mayúscula, n-gramas de caracteres, promedio de números, frecuencia de caracteres especiales, proporción de datos numéricos, etc.
Léxicas	Número de palabras, palabras emocionalmente positivas, emocionalmente negativas, lista de palabras foráneas n-gramas de palabras, acrónimos, promedio de la longitud de las palabras, proporción de palabras distintas en el texto, número de palabras que ocurren una sola vez (hapax legomena), etc.
Sintácticas	N-gramas de etiquetas de partes de la oración (*Part Of Speech*, POS, por sus siglas en inglés), n-gramas sintácticos, frecuencia de palabras de función y signos de puntuación, palabras vacías, verbos en pasado, errores gramaticales, etc.
Estructurales	Número de: párrafos, sentencias, caracteres o palabras por párrafo, etiquetas (*hashtags*), retweets, URL's usados, emoticones. Así como jerga y etiquetas HTML.
Contenido específico	Frecuencia de palabras clave específicas de contenido, palabras relacionadas con sentimientos, palabras señalando tópicos específicos, palabras provenientes de LIWC.
Legibilidad	Medidas como: facilidad de lectura, LIX, RIX, SMOG, etc.
Recuperación de información	Coseno y Okapi BM2.5

que MNB obtuvo mejor desempeño que otros algoritmos que probaron (aunque no detallan cuales son).

Adicionalmente, algunos autores sugieren que el clasificador KNN puede influir positivamente en el desempeño de AP. Por ejemplo, Duong et al. (2016) predijeron edad, género, origen geográfico y ocupación de autores de blogs[10]. En su trabajo, contrastan el desempeño de 10 algoritmos[11] del kit de weka encontrando que IBk obtuvo los mejores resultados en la predicción de la mayoría de los rasgos. Por su parte, Ramnial et al. (2016) exploraron género en tesis de posgrado encontrando un desempeño similar entre KNN y SVM.

Específicamente, el presente trabajo exploró el desempeño de tres algoritmos: MNB, KNN (K=1) y las SVM.

2.3 Resumen

La tarea AP – abordada en la presente investigación – se ubica dentro del análisis de autoría, campo que ha sido enfocado como un problema de categorización de textos. Por lo tanto, en este capítulo, primero, se presentó y discutió el marco fundamental de la tradicional categorización de textos. En específico, se describieron las tres etapas presentes en la construcción de un clasificador de textos y que también corresponden a un clasificador de perfiles de autores: la construcción de la representación, el proceso de clasificación y

[10] Específicamente, se trata de blogs escritos en idioma Vietnamita.

[11] Los algoritmos considerados son: ZeroR, árboles de decisión (J4.8), bosques aleatorios, bagging, IBk (IB1), SVM (SMO), NB, redes bayesianas, redes neuronales (perceptrón multicapa) y RandomTree.

la evaluación. Posteriormente, se introducen las tareas del análisis de autoría, donde se coloca especial interés en AP. Asimismo, se ha discutido una serie de métodos que la comunidad científica ha propuesto para abordar tales tareas.

Durante el desarrollo del capítulo, se discutió el contexto actual de la categorización de documentos marcando las similaridades y sobre todo, las diferencias más relevantes que existen entre ambas disciplinas (TC y AA con especial énfasis en AP). La diferencia principal se ve señalada en el tipo de términos considerados para crear la representación de los textos. Mientras en TC las palabras de función no aportan información, en AA son básicas para caracterizar el estilo de redacción de los autores. Cabe señalar que en TC, las palabras de contenido han destacado por ser descriptivas de la semántica de los documentos. En cambio, en las tareas de AA, dos tipos de atributos (términos) han mostrado ser de utilidad: temáticos (contenido) y estilísticos.

De forma específica, en la tarea AP aplicada a redes sociales se han usado combinaciones de ambos tipos de atributos (Rangel et al., 2014, 2015; Rangel et al., 2016). Recientemente, en AP se ha sugerido una relevancia mayor de los atributos de contenido indicando que los perfiles de grupos de personas pueden ser discriminados por sus intereses temáticos más que por los estilos de escritura (Rangel et al., 2014; López-Monroy et al., 2015; Fatima et al., 2017). Con respecto a las técnicas de selección y pesado de términos dirigidas a construir la representación vectorial de los documentos, se han usado técnicas tradicionales provenientes de TC.

Todas las tareas mencionadas comparten un marco común del proceso de clasificación así como evaluación. Se ha destacado el uso SVM por su habilidad para trabajar con representaciones de alta dimensionalidad. Por otra parte, comúnmente, los resultados se reportan mediante la exactitud y se realizan pruebas de significancia estadística que permiten comparar el desempeño de dos o más clasificadores.

La tarea correspondiente a AP será descrita con mayor detalle en el siguiente capítulo, por ser el área donde se ubican la contribuciones de la presente tesis.

Capítulo 3

Enfoques de identificación del perfil de autores

En este capítulo se presenta el estado del arte relacionado con las contribuciones de esta tesis. Primero, se describen algunos métodos enfocados a identificar rasgos del perfil de autores mediante análisis de textos provenientes de redes sociales (Sección 3.1). Se ha colocado especial interés en la identificación de edad y género, por ser los rasgos en los cuales que se valida el enfoque propuesto (Sección 3.1.2). Posteriormente, para sustentar la idea principal del enfoque, la cual señala que la información personal contenida en los textos revela el perfil de autores, se presentan trabajos, desde una perspectiva psicológica, que conectan el uso de pronombres personales

con el perfil de los autores (3.2). Después, considerando que esta tesis propone novedosos esquemas de selección y pesado de términos, se presenta el panorama actual de este tipo de esquemas comúnmente usados en AP (Secciones 3.3 y 3.4). Finalmente, se describen algunas colecciones de documentos de entrenamiento, así como, los resultados reportados hasta el momento mediante su uso.

3.1 Identificación del perfil de autores en redes sociales

En sus inicios, AP fue una tarea enfocada a trabajar con corpus formales (Koppel et al., 2002; Argamon et al., 2003, 2005; Koppel et al., 2005; Bergsma et al., 2012). Por ejemplo, Koppel et al. (2002) así como Argamon et al. (2003), pioneros en AP, usaron el *British National Corpus (BNC)*, documentos formales (artículos y libros) para demostrar que hay diferencias lingüísticas entre el estilo de escritura de hombres y mujeres. Ambos estudios coinciden en que las mujeres recurren más al uso de pronombres que los hombres.

Más tarde, debido al auge de los medios electrónicos de comunicación, nace el interés y reto de procesar datos no formales, los cuales pueden ser llamados textos informales de la web y que son de gran relevancia para atender tareas de la lingüística forense. Entre ellos se encuentran: correos electrónicos (Estival et al., 2007b; Cheng et al., 2011), blogs (Schler et al., 2006; Yan y Yan, 2006; Argamon et al., 2009; Mukherjee y Liu, 2010; López-Monroy et al., 2015), microblogs, publicaciones de redes sociales (Peersman et al., 2011; Schwartz

et al., 2013; Sap et al., 2014), críticas (*reviews*) (López-Monroy et al., 2015), etc.

Prácticamente el reto de trabajar con textos informales de la web radica en la longitud comúnmente corta de los documentos, la existencia de pocos ejemplos de entrenamiento y la presencia de categorías desbalanceadas. Además, la tarea se dificulta debido a la alta variabilidad léxica ocasionada por la existencia recurrente de errores ortográficos, abreviaciones, modismos, emoticones, palabras con repeticiones de caracteres, etc. Sin embargo, estos elementos léxicos informales también han facilitando la tarea, pues se han explorado como marcadores estilísticos para discriminar grupos de personas. Por ejemplo, se ha estudiado el uso de las palabras informales (jerga) como atributos discriminativos (Schler et al., 2006; López-Monroy et al., 2015), así como los errores ortográficos (Marquardt et al., 2014) y los emoticones (Rao et al., 2010; Marquardt et al., 2014; Posadas-Durán et al., 2015; Rangel y Rosso, 2016). Este valor dual de este tipo información informal hace más interesante la tarea.

Ahora bien, dada la relevancia de crear herramientas que se adapten a nuevas formas de comunicación, como las redes sociales, se han propuesto varios enfoques de solución para AP. La base común para afrontar los retos mencionados recae en la búsqueda de atributos adecuados que identifiquen el perfil del autor. En los siguientes párrafos se examinarán brevemente algunos trabajos que se suman al reto de trabajar con textos informales prediciendo algún(os) rasgos socio-demográficos.

3.1.1 Rasgos del perfil de autores comúnmente explorados

Existe una diversidad de rasgos socio-demográficos que conforman el perfil de los autores y que la literatura ha explorado mediante enfoques de AP. Algunos ejemplos son:

Personalidad. Generalmente, los enfoques de AP que reconocen personalidad (Nowson y Oberlander, 2006; Schwartz et al., 2013; Álvarez-Carmona et al., 2015; Kocher, 2015), se basan en el modelo *Big Five*, que puede ser traducido como los *Cinco Grandes*: apertura, responsabilidad, extraversión, amabilidad e inestabilidad emocional (McCrae y John, 1992). Para abordar la tarea, se han empleado representaciones sencillas basadas en el uso de términos más frecuentes (Kocher, 2015). También, se han propuesto representaciones enriquecidas mediante categorías de palabras (características lingüísticas y psicológicas) obtenidas a través de LIWC (Nowson y Oberlander, 2006; Qiu et al., 2012; Schwartz et al., 2013; Golbeck, 2016), e incluso se ha integrado el uso de atributos de segundo orden (Álvarez-Carmona et al., 2015). En relación con los algoritmos de clasificación, la mayoría de los enfoques resalta la conveniencia del uso de las SVM. Una descripción completa de trabajos orientados hacia la detección de personalidad puede ser encontrada en (Vinciarelli y Mohammadi, 2014) y en las notas del PAN 2015 (Rangel et al., 2015).

Origen regional y orientación política. Aunque son rasgos poco explorados, existen algunas propuestas interesantes. Por ejemplo, el trabajo de Rao

et al. (2010) describe un método para encontrar el origen regional y la orientación política (conservador vs demócratas). En ese trabajo, se exploraron características socio-lingüísticas tales como: presencia de emoticones, datos correspondientes al uso de la red (e.g., número de seguidores, amigos) y comportamiento de comunicación (e.g., frecuencia de *"retweet"*). Así como también, un ensamble formado por un modelo socio-lingüístico y un modelo basado en n-gramas. Por su parte, Pennacchiotti y Popescu (2011) construyeron una colección de textos de Twitter para estudiar la predicción de: afinación política clasificando a demócratas y republicanos; origen étnico detectando entre afroamericanos y otras etnias; así como también la afinidad o antipatía para seguir un negocio en particular en Twitter. Los autores probaron un conjunto extenso de características de distinta naturaleza (perfil, comportamiento, contenido lingüístico e información de la red social).

Idioma nativo. Esta tarea consiste en determinar el idioma nativo de los autores analizando textos escritos en un segundo idioma. Mientras la mayoría de los enfoques propuestos usa ensayos, considerados como textos formales (Koppel et al., 2005; Argamon et al., 2009; Bykh y Meurers, 2014), son pocos los trabajos que exploran la tarea en dominios sociales, como el trabajo realizado por Tofighi et al. (2012), quienes concluyeron que los atributos léxicos y sintácticos son buenos discriminadores.

Ocupación y nivel de educación. Aunque la ocupación es uno de los rasgos del perfil del autor poco explorados, existen propuestas como la de Duong et al. (2016), quienes usaron textos de tres foros web de diferente dominio

para inferir tres ocupaciones populares: negocios-ventas-administración, tecnología-educación y cuidado de la salud. La representación incluye una combinación de atributos de contenido y estilo. Por otro lado, Estival et al. (2007) identificaron, a partir de correos electrónicos escritos en árabe, el nivel de educación de los autores: educación superior y básica (no superior). Usan varias características a nivel carácter, morfológico y entidades nombradas.

Otros rasgos ampliamente explorados son edad y género de los autores, contexto en el que se desenvuelve la presente investigación. Por tal motivo, se coloca especial interés en su descripción en la siguiente sección.

3.1.2 Identificación de edad y género

De forma general, la identificación de edad y género se realiza de manera supervisada y como variables categóricas. Si bien existen trabajos donde la predicción de la edad se modela como una variable continua (Nguyen et al., 2011; Sap et al., 2014), la tendencia ha sido marcada por la clasificación categórica de rangos de edad (e.g., 10-20 años, 20-30 años, 30 o más) y los dos géneros (hombre, mujer).

Existe un gran número de trabajos que abordan la predicción de edad y género. Como se observa a continuación, la mayoría están dirigidos a determinar un conjunto adecuado de características de diferente naturaleza para modelar el perfil de escritura de los autores. Como se ha mencionado, comúnmente, se propone una combinación de características temáticas y estilísticas.

Uno de los trabajos pioneros en AP trabajando con corpus no formales, es el trabajo de Schler et al. (2006), quienes usaron blogs de 19,320 autores escritos en inglés y el clasificador Multi-Class Real Winnow para probar una representación basada en una combinación de atributos de contenido y estilo: palabras de contenido, palabras función y etiquetas POS. En particular, usando los 1000 términos con mayor ganancia de información y pesando con frecuencia normalizada, reportaron una exactitud de clasificación de género y edad de 80% y 76.2%, respectivamente. A través de los resultados concluyen que existen diferencias temáticas (atributos de contenido) y también estilísticas entre hombres y mujeres, así como entre los grupos de diferentes edades. Por ejemplo, los hombres escriben más acerca de política y tecnología. En cambio, las mujeres discuten comúnmente su vida personal y reflejan más su estilo de escritura.

Por su parte, Nowson y Oberlander (2006) identificaron el género de 71 autores de blogs usando términos obtenidos a partir de LIWC y MRC (una base de datos psicolingüística), así como n-gramas de palabras. Usando SVM, los autores reportaron una exactitud de 91.5%. Sin embargo, concluyeron que el uso de n-gramas genera mejores resultados (hasta 93%) que los atributos obtenidos de diccionarios (LIWC y MCR).

Por otra parte, Yan y Yan (2006) presentaron un enfoque basado en Naïve Bayes para identificar el género de 3000 autores de blogs. En su trabajo resaltan su contribución al ser los primeros que usaron atributos específicos de los blogs como colores del fondo, emoticones y tipografía. Dentro de su estudio

concluyen que el uso de atributos no tradicionales enriquece el desempeño de BOW. Como una observación adicional, los autores señalan que remover palabras vacías provoca un decremento del desempeño. Hoy en día, es bien conocido que este tipo de palabras está relacionado con el estilo de redacción, de ahí que enriquecen la tarea.

Posteriormente, Estival et al. (2007) predijeron edad y género de autores de correos electrónicos mediante el uso de SVM. Su representación consistió de atributos a nivel de carácter, léxico y estructurales. Por ejemplo, longitud de palabras, etiquetas POS y categoría del correo electrónico. Los resultados indicaron que todas esas características fueron muy importantes para generar exactitudes de 56.46 y 69.26 correspondientes a edad y género, respectivamente.

Argamon et al. (2009) usaron características de estilo y contenido. Los resultados mostraron una mejor exactitud usando ambos tipos de atributos que empleando sólo un grupo. Los valores de exactitud obtenidos corresponden a 76.1 % y 77.7 % para género y edad, respectivamente.

Continuando la búsqueda de atributos de diferente naturaleza, se han propuesto atributos más sofisticados. Por ejemplo, Mukherjee y Liu (2010) propusieron el uso de patrones de secuencias POS de longitud variable extraídos del conjunto de entrenamiento para abordar la clasificación de género en blogs. Al mismo tiempo, proponen un método de selección de términos que prácticamente corresponde a un ensamble de varios enfoques de selección de términos. Su propuesta fue probada con SVM y Naïve Bayes generando una

exactitud máxima de 88.56% con el primer algoritmo. Dados los resultados, se observa que los patrones de secuencias POS capturan regularidades lingüísticas complejas de hombres y mujeres.

Más aún, han sido exploradas características específicas de la forma de expresión en redes sociales. Por ejemplo, Rao et al. (2010) usan la presencia de emoticones, características del uso de la red (e.g., número de seguidores, amigos) y comportamiento de comunicación (e.g., frecuencia de "retweet"). Por su parte, Rangel y Rosso (2013) propusieron un conjunto de características para discriminar género y edad con un clasificador basado en SVM: frecuencias de mayúsculas, longitud de las palabras, número de palabras con repeticiones de caracteres (e.g., *Heeeelloooo*) signos de puntuación, etiquetas POS, emoticones y léxico en español relacionado con las emociones. En su estudio concluyen que dichas características se desempeñan mejor para edad que para género. Por lo tanto, sugieren que el estilo de escritura depende más de la edad del autor que del género.

En esta búsqueda de la combinación adecuada de atributos, incluso han sido exploradas millones de características mediante un enfoque MapReduce[1] (Maharjan et al., 2014).

Particularmente, en 2013 surge el foro internacional de evaluación denominado PAN[2], donde la comunidad científica compite evaluando sus enfoques de solución en la identificación del perfil de autores. El atractivo desafío así como la

[1] Modelo de programación para procesamiento paralelo y distribuido
[2] http://pan.webis.de/

disposición de colecciones de textos etiquetados, han alentado el surgimiento de un gran número de investigaciones prediciendo el perfil del autor con un marco común de comparación. Este foro se ha realizado anualmente desde 2013, con un enfoque en la identificación de edad y género; también se ha sumado la predicción de personalidad en el año 2015 e identificación del idioma en el presente 2017, aunque este último año el caso edad no fue considerado.

En los resúmenes del PAN (Rangel et al., 2013, 2014, 2015; Rangel et al., 2016), también se observa que las contribuciones de los trabajos están orientadas a combinar atributos de distinta naturaleza. También se han propuesto representaciones basadas en atributos más elaborados donde sobresale el uso exitoso de los atributos de segundo orden introducidos por López-Monroy et al. (2013). En las cuatro ediciones del PAN, las representaciones que incluyen el uso de atributos de segundo orden han logrado las primeras posiciones en algunas tareas. También se han propuesto representaciones continuas de palabras (*word embeddings*) (Bayot y Gonçalves, 2016). Por otro lado, en relación con los algoritmos de aprendizaje, las SVM han destacado por su buen desempeño, aunque también se han explorado algoritmos basados en árboles de decisión, MNB y diversos ensambles. En adición, se han utilizado técnicas de aprendizaje profundo como las redes neuronales convolucionales (Surendran et al., 2017).

Recientemente, en la búsqueda de enfoques que modelen la forma en que los autores construyen sus textos, Rangel y Rosso (2016) han estudiado el impacto de las emociones en AP. Su propuesta se basa en el uso del método EmoGraph

para modelar la forma en la cual se usa el lenguaje y las emociones cuando se escribe. Para ello, proponen una representación basada en grafos de etiquetas POS, manteniendo así la secuencia de las palabras. En su representación evitan información de contenido para obtener independencia de tópicos. A través de su estudio, se determinó que la forma en que los usuarios expresan sus emociones depende de su edad y género. Adicionalmente, observaron que las características obtenidas de los grafos son buenos discriminadores de la edad y género.

Diferente de la mayoría de los trabajos en el estado del arte de AP, la presente investigación se enfoca en estudiar y aprovechar la relevancia de las sentencias que contienen pronombres personales en primera persona, más que en definir un conjunto adecuado de características para AP. El interés de trabajar con frases con pronombres personales tiene un sustento psicológico, el cual es presentado en la siguiente sección.

3.2 Información personal en la identificación del perfil de autores

AP es soportado en la idea de que las personas con características comunes de su perfil también comparten similaridades lingüísticas, en parte, debido a su ambiente cultural o social. De hecho, hay estudios interesantes acerca de cómo el lenguaje es compartido por la gente desde una perspectiva psicológica (Pennebaker y Stone, 2003; Pennebaker, 2011).

3.2.1 Perspectiva psicológica

Algunos estudios han establecido una relación entre el uso del lenguaje y los rasgos del perfil de las personas (Pennebaker et al., 2003; Pennebaker, 2011) como personalidad (Fast y Funder, 2008; Yarkoni, 2010) y diferencias de género (Mehl et al., 2007). Más específicamente, se ha establecido un vínculo entre las palabras de función y actividades biológicas, sociales, cognitivas y de personalidad (Chung y Pennebaker, 2007; Pennebaker, 2011).

Particularmente, un tipo de palabras de función que son de primordial interés en esta tesis son los pronombres personales, los cuales actúan como marcadores de estilo aportando información para distinguir diferentes grupos de personas. Por ejemplo, el uso frecuente de pronombres en primera persona del singular está relacionado con:

- Gente joven. Pennebaker y Stone (2003) analizaron muestras de escritura y transcripciones de entrevistas de 3280 participantes. En sus resultados encontraron que el uso de frases con referencias al hablante (uso de pronombres en primera persona del singular) disminuye con la edad.
- Mujeres. Varios estudios han señalado que las mujeres recurren al uso de pronombres en primera persona del singular en tasas más altas que los hombres (Pennebaker y Stone, 2003; Newman et al., 2008; Argamon et al., 2009).
- Bajo estatus social. El uso relativo de la primera persona del singular es un marcador de bajo estatus social. Por el contrario, las personas

consideradas con alto estatus social usan más pronombres en primera persona del plural (Kacewicz et al., 2014).

- Depresión. Se ha señalado que las personas con depresión usan el pronombre *I* de forma más frecuente que aquellas que nunca han sufrido depresión (Rude et al., 2004).
- Honestidad. Se ha determinado que cuando las personas están diciendo la verdad, se expresan de manera más personal y al mismo tiempo, describen su historia en una forma cognitivamente más compleja (Newman et al., 2003).

Puede notarse, que el uso de pronombres personales en primera persona del singular se relaciona directamente con rasgos del perfil de los autores o con características que ayudan a derivar su perfil como el estatus y estado emocional. De hecho, la mayoría de las contribuciones acerca del uso de pronombres se ubican en el campo psicológico y social. En definitiva, tales contribuciones podrían aprovecharse para analizar textos y crear métodos automáticos que permitan predecir rasgos del perfil de los autores como se muestra en la siguiente sección.

3.2.2 Perspectiva computacional

Algunos enfoques computacionales de AP han remarcado en sus conclusiones la relevancia de los pronombres personales a través de su uso como atributos, pero ningún método ha analizado la riqueza de su contexto como la información más discriminativa de perfiles de autores. Por ejemplo, en la predicción de

género, Argamon et al. (2009) encontraron que los pronombres personales son fuertes marcadores del estilo de escritura de las mujeres y que la palabra *im* es un fuerte identificador de la gente joven. Por otro lado, Newman et al. (2003) encontró que las mujeres usan más pronombres en primera persona del singular que los hombres. Por su parte, Pennebaker y Stone (2003) concluyeron que el uso de los pronombres en primera persona del plural aumenta con la edad.

Inclusive, en el último foro del PAN, para la identificación de género, Gencheva et al. (2016) indagaron la presencia de sentencias con pronombres, por ejemplo: *my wife*, *my man*, *my girlfriend*, las cuales indican fácilmente evidencia de la expresión del género del autor. En lo que concierne a la edad usaron un patrón expresado por la palabra *I'm* seguida de una expresión numérica que expresa la edad. Aunque, en ese trabajo únicamente se consideró si se había mencionado algo que hiciera obvia la pertenencia a una categoría, desde la perspectiva de esta tesis, se está reforzando la idea base que indica que las palabras cercanas a un pronombre personal reflejan información discriminativa del perfil de los autores.

Ahora bien, la presente investigación estudia el uso de los pronombres personales como una fuente reveladora de los perfiles de autores. Pero, a diferencia de los trabajos mencionados anteriormente, no se estudian como atributos de la representación. Más bien, esta tesis sugiere que los términos contenidos en las frases con pronombres en primera persona del singular, son las secciones (fragmentos) más importantes que contienen información

valiosa de estilo y contenido para discriminar perfiles de autores. Con base en estas ideas, en esta investigación se proponen nuevos métodos de selección y pesado de términos. Por tal motivo, en las siguientes secciones, se presenta la literatura relacionada con las etapas de selección y pesado de términos en AP.

3.3 Esquemas comunes de selección de términos en AP

Es conocido que en la tradicional tarea de TC, han emergido nuevos métodos para reducir la dimensionalidad o para encontrar un conjunto representativo de características (Singh et al., 2010; Wang et al., 2012; Zong et al., 2015; Uysal, 2016). Sin embargo, en AP se han utilizado comúnmente esquemas tradicionales de TC (Yang y Pedersen, 1997; Sebastiani, 2002). Por ejemplo, varios trabajos en el estado del arte de AP usan un umbral de frecuencia de términos (Bayot y Gonçalves, 2016; Gencheva et al., 2016), evitando que características estilísticas valiosas sean removidas por los métodos de selección de características[3].

Ganancia de información es otro esquema ampliamente utilizado (Agrawal y Gonçalves, 2016; Schler et al., 2006) principalmente para analizar e interpretar las características usadas (Schler et al., 2006) o para extraer términos temáticos en sistemas con múltiples tipos de características (Agrawal y Gonçalves, 2016).

[3]Muchas caracterítiscas estilísticas tienen alta frecuencia en los documentos así como en la colección de docuementos, por tanto no son considerados discriminativos y son eliminados. Por otro lado, muchas características poco frecuentes también son eliminadas por considerar que su frecuencia no refleja una asociación con categoría alguna.

Otros esfuerzos incluyen el uso de la medida χ^2 (Bilan y Zhekova, 2016). Sin embargo, la búsqueda de enfoques más adecuados y específicos para AP en dominios sociales comienza a ser más evidente. En este contexto, Mukherjee y Liu (2010) propusieron una combinación de un *filter* y un *wrapper* para mejorar la selección de características. Recientemente, Markov et al. (2016) evaluaron una técnica conocida como punto de transición (*transition point*) con el objetivo de elegir atributos más apropiados para la representación de un documento.

Otro esquema de selección también empleado en tareas relacionadas con clasificación texto es el índice Gini (Shankar y Karypis, 2000; Shang et al., 2007). En este contexto, Singh et al. (2010) propusieron un nuevo método de selección llamado *within-class-popularity*, donde se usa el coeficiente Gini para estimar la calidad de los atributos. En AP existen pocos trabajos que exploran su uso, por ejemplo Przybyła y Teisseyre (2015) lo utilizan para evaluar el poder predictivo de los atributos.

A pesar de la heterogeneidad de los métodos de selección, la mayoría están soportados en inferencias estadísticas acerca de la ocurrencia de los términos en los documentos. Más importante, ellos no consideran la calidad de la información a través de diferentes sentencias en el documento. En contraste, en este trabajo de investigación se propone un esquema de selección de términos que remarca que no toda la información en un documento es igualmente relevante y que existen sentencias más informativas que otras.

3.4 Esquemas comunes de pesado de términos en AP

Los esquemas de pesado de términos usados en AP provienen del área de TC, por ejemplo: booleano, TF y TF-IDF. También se han diseñado nuevos esquemas en TC, tal es el caso del trabajo desarrollado por Debole y Sebastiani (2004) donde se propone reemplazar el componente IDF por una función de evaluación basada en la categoría del término, la cual ha sido previamente usada en la fase de selección de términos. Otros propuestas han surgido tratando problemas específicos tales como el desbalance de las categorías (Liu et al., 2009). Recientemente, el uso de enfoques genéticos fueron propuestos para aprender automáticamente esquemas de pesado de términos (Escalante et al., 2015).

Si bien se han propuesto diversos esquemas en TC, en AP existe una tendencia hacia el uso de métodos comunes como los mencionados anteriormente. Es importante notar que el factor común es el uso de información estadística de las ocurrencias de los términos en el documento, pero no se consideran atributos cualitativos de esas ocurrencias. De hecho, la frecuencia normalizada es el esquema mayormente usado en AP (Koppel et al., 2002; Schler et al., 2006; Rao et al., 2010; Schwartz et al., 2013). Esto puede ser corroborado en recientes foros de AP, donde muchos de los trabajos únicamente exploran TF-IDF o combinaciones con la frecuencia normalizada (Agrawal y Gonçalves, 2016; Bayot y Gonçalves, 2016; Bilan y Zhekova, 2016; Gencheva et al., 2016).

En contraste, en la presente investigación se propone un novedoso esquema de pesado dirigido a enriquecer la frecuencia con la evaluación del contexto de los atributos en el documento.

3.5 Colecciones de textos etiquetados

La popularidad de las redes sociales en los años recientes ha incrementado la cantidad de textos disponibles en Internet. Debido a su fácil acceso, muchos de estos textos son utilizados como datos de entrenamiento y prueba para evaluar enfoques de predicción del perfil de autores. En algunos casos, los investigadores etiquetan manualmente los documentos de las colecciones con las categorías correspondientes a los rasgos del perfil (Mukherjee y Liu, 2010; Nguyen et al., 2013); pero en su mayoría, el etiquetado se realiza automáticamente tomando en cuenta información proporcionada por los mismos usuarios (Schler et al., 2006; Rangel y Rosso, 2013). En específico, para los experimentos realizados en la presente investigación, se utilizaron las colecciones de documentos descritas en las siguientes secciones, las cuales fueron recopiladas de redes sociales y etiquetadas automáticamente con la edad y género de los autores.

3.5.1 Blogs de Schler

El corpus base de todos los experimentos en esta tesis es el repositorio propuesto en el trabajo de Schler et al. (2006). Según su autor, este corpus

es la recopilación de blogs escritos en inglés provenientes de blogger.com recopilados en agosto de 2004.

Cada blog está etiquetado por edad y género, su longitud promedio es de 7250 palabras. Los blogs fueron concentrados en tres categorías según su edad: 10s (desde 13 hasta 17 años), 20s (desde 23 hasta 27 años) y 30s (desde 33 hasta 47 años). Los creadores del corpus señalan que se omitieron los grupos intermedios para permitir una diferencia clara, ya que existían blogs que habían estado activos durante varios años. Naturalmente, para género se tienen dos categorías: hombre y mujer. La Tabla 3.1 muestra la distribución detallada del corpus, la cual está balanceada con respecto al número de hombres y mujeres en cada grupo de edad.

Tabla 3.1: Distribución del corpus de Schler

Edad (rango)	Género		
	Mujer	Hombre	Total
10s (13-17)	4,120	4,120	8,240
20s (23-27)	4,043	4,043	8,086
30s (33-47)	1,497	1,497	2,994
Total	9,660	9,660	19,320

La colección de Schler ha sido ampliamente utilizada debido a su gran tamaño. Sus creadores reportan en su trabajo una exactitud correspondiente 80% para género y 75% para edad, mediante el uso de atributos de estilo y contenido.

Por su parte, también usando una combinación de atributos de estilo y contenido en el mismo corpus, Argamon et al. (2007) obtuvieron una exactitud correspondiente a 76.1% para género y 77.7% para edad.

Por otro lado, Booker (2008) usó atributos a nivel de grupo aplicando métodos de análisis de tópicos. Sus resultados corresponden a 72.83% para edad y 75.04% para género.

Recientemente, López-Monroy et al. (2015) también reportaron resultados en este corpus. Su propuesta se basó en el uso de atributos de segundo orden reportando una exactitud de: 77.68% para edad y 82.01% para género.

3.5.2 PAN-AP-2014 corpus

Este es el corpus de entrenamiento para la tarea AP en la competencia del PAN 2014, se compone de cuatro dominios: blogs, revisiones de hoteles de TripAdvisor (de aquí en adelante, denotadas como Reviews), documentos de redes sociales (generalizados como Social Media) y publicaciones de Twitter (denotadas como Twitter). Cada documento pertenece a un autor y está etiquetado por género así como por edad. En el caso de la edad, existen 5 categorías: 18-24, 25-34, 35-49, 50-64 y $\geq$65. Los detalles de las colecciones mencionadas se muestran en la Tabla 3.2.

Es importante notar que todas estas colecciones están balanceadas con respecto al género, pero desbalanceadas con respecto a la edad. Adicionalmente, presentan aspectos que dificultan la tarea con respecto al corpus de Schler: i) el tamaño de las colecciones es muy diferente, variando desde 147 blogs de usuarios hasta 7746 en la colección denominada Social Media; ii) el número de etiquetas en el caso edad es más grande (5 etiquetas).

Tabla 3.2: Distribución de los documentos que conforman la colección PAN-AP-2014

Corpus	Género	Edad					
		18-24	25-34	35-49	50-64	$\geq$65	Total
Blogs	Mujeres	3	30	27	11	2	73
	Hombres	3	30	27	12	2	74
	Total	6	60	54	23	4	147
Twitter	Mujeres	10	44	65	30	4	153
	Hombres	10	44	65	30	4	153
	Total	20	88	130	60	8	306
Reviews	Mujeres	180	500	500	500	400	2080
	Hombres	180	500	500	500	400	2080
	Total	360	1000	1000	1000	800	4160
Social Media	Mujeres	775	1049	1123	919	7	3873
	Hombres	775	1049	1123	919	7	3873
	Total	1550	2098	2246	1838	14	7746

Todo ello, lo expone como un corpus más difícil, pero a su vez, muy interesante para probar la generalidad y robustez de los enfoques.

Los resultados reportados por distintos autores usando estas colecciones han sido muy variados. En la Tabla 3.3 se muestran algunos trabajos que han probado sus enfoques de AP a través de estas colecciones: Álvarez-Carmona et al. (2016) explotó el uso de representaciones basadas en tópicos usando (LIWC) y (LSA); Weren et al. (2014) implementó ideas de recuperación de información (IR); López-Monroy et al. (2013) exploró el uso de atributos de segundo orden (SOA); López-Monroy et al. (2015) usó representaciones basadas en subperfiles (SSR).

Cabe señalar que el enfoque con los mejores resultados en las colecciones de las secciones previas corresponde al enfoque SSR. Por su parte, SOA es un

Tabla 3.3: Resultados del estado del arte usando la colección PAN-AP-2014

Rasgo	Trabajo	Reviews	Twitter	Blogs	Social Media
Edad	LSA	34.00	39.00	48.00	36.00
	LIWC	29.00	47.00	42	34
	IR	37.62	52.61	45.58	42.51
	SOA	33.92	47.97	48.07	37.00
	SSR	36.9	49.01	53.06	38.06
Género	LSA	65.00	66.00	70.00	52.00
	LIWC	62.00	71.00	60.00	50.00
	IR	71.03	78.76	82.99	57.04
	SOA	68.05	71.92	77.96	55.36
	SSR	69.27	71.69	80.95	55.39

antecesor de SSR con los mismos principios[4]. Por ello, SSR es el método de referencia principal en esta investigación.

3.6 Resumen

En este capítulo se presentó el estado del arte relacionado con las contribuciones de esta tesis. En específico, se analizaron y discutieron las características de los enfoques que la comunidad científica ha propuesto para predecir rasgos del perfil de autores como personalidad, nivel de educación e idioma nativo. De forma detallada, se abordó la predicción de la edad y el género de los autores, por ser el contexto donde se valida el enfoque propuesto.

En este contexto, se observó una tendencia muy marcada para proponer combinaciones adecuadas de atributos de contenido y estilo que permitan

[4]Aunque Weren et al. (2014) reporta en el conjunto de entrenamiento algunos resultados mejores, en el conjunto de prueba sus resultados fueron menores al enfoque SOA. Esto sugiere que las medidas de recuperación de información causaron un sobreajuste al corpus de entrenamiento.

clasificar perfiles de autores eficientemente. Sin embargo, la presente tesis no está orientada a proponer una combinación más, sino, a enriquecer los paradigmas de uso, selección y pesado de términos mediante un sustento psicológico que indica que la información personal contenida en los textos concentra términos que pueden revelar el perfil de los autores. Por esta razón, en este capítulo se ha discutido una serie de trabajos desde la perspectiva psicológica que relacionan el perfil de los autores con el uso del lenguaje, particularmente, con la utilización de los pronombres personales.

Asimismo, se discutió el marco literario correspondiente a los esquemas de selección y pesado de términos comunes en la identificación del perfil de autores. En la literatura se ha notado el uso de técnicas comunes provenientes de TC, las cuales están basadas en la frecuencia de los términos en los documentos. En contraste, en esta tesis se proponen nuevas estrategias de selección y pesado de términos cuyo objetivo es enriquecer los esquemas basados en frecuencia mediante la evaluación del contexto de las ocurrencias de los términos. Las propuestas constituyen una alternativa para hacer frente a los desafíos adheridos a los textos provenientes de redes sociales.

De manera general, el enfoque propuesto en esta tesis se fundamenta en la riqueza discriminativa de la información personal expuesta en los documentos. En el siguiente capítulo se analiza la contribución de este tipo de información en la identificación del perfil de autores.

Capítulo 4

El rol de las frases personales en la identificación del perfil de autores

En este capítulo, se evalúa el rol de las frases que contienen pronombres personales en la identificación del perfil de autores. Primero, se explora el tipo de información transmitida en el contexto de los pronombres personales y su conveniencia para generar patrones que ayuden a predecir el perfil de los autores. Posteriormente, se describe una serie experimentos cuyos resultados muestran que el contexto de los pronombres personales en los documentos conforma la *esencia* de los textos para AP.

4.1 El valor del contexto de los pronombres personales

Los pronombres personales (PP) corresponden a elementos gramaticales que hacen referencia a objetos, personas o animales sin nombrarlos. La Tabla 4.1 muestra la clasificación de tales pronombres. Los pronombres se presentan en inglés por ser el idioma de los textos utilizados en esta investigación.

Tabla 4.1: Clasificación de los pronombres personales

Persona	Caso			
	Subjetivo	Objetivo	Posesivo	Reflexivo
Singular				
1ra.	*I*	*Me*	*Mine*	*Myself*
2da.	*You*	*You*	*Yours*	*Yourself*
3ra.	*He/She/It*	*Him/Her/It*	*His/Hers/Its*	Himself/Herself/Itself
Plural				
1ra.	*We*	*Us*	*Ours*	*Ourselves*
2da.	*You*	*You*	*Yours*	*Yourselves*
3ra.	*They*	*Them*	*Theirs*	*Themselves*

Si bien, los PP juegan un rol importante en el lenguaje porque sustituyen a los sustantivos, su contexto – términos en la misma frase – también es relevante, pues es rico en verbos y adjetivos que expresan la información acerca de dichos sustantivos. Muy interesante es cuando los pronombres hacen referencia a las personas que escriben un texto, pues la información de su contexto puede generalizarse en patrones que infieren el perfil de los autores. Desde una perspectiva psicológica, el uso de PP, sobre todo en primera persona del singular, ha sido asociada con características que definen la identidad (Pennebaker, 2011) (ver Sección 3.2). De ahí que, en esta tesis, se estudia

la riqueza del contexto de los PP en AP. Específicamente, la idea base de la presente investigación está motivada en dos hallazgos psicológicos. Primero, que el uso de los pronombres personales aportan información valiosa para revelar características del perfil de los autores (Pennebaker, 2011). Segundo, que las personas tienden a ser más honestas cuando se refieren a ellas mismas (usan más referencias a ellas mismas cuando están diciendo la verdad) (Newman et al., 2003). Ambos aspectos son de gran ayuda para reducir la falta de fiabilidad en los dominios sociales.

Particularmente, se estudian frases que contienen PP de naturaleza inclusiva, es decir, PP en primera persona del singular. En esta tesis, este tipo de frases son llamadas *Frases Personales* y han sido denotadas con las siglas FP. Los pronombres en segunda y tercera persona no fueron considerados porque ellos sugieren que el escritor está hablando acerca de algo/alguien más, sin incluirse.

La relevancia hipotetizada en esta tesis sobre las FP en AP es consecuencia del tipo de información contenida en ellas y de la claridad de su contenido. Para ilustrar esta idea, la Tabla 4.2 muestra algunas frases personales extraídas de blogs provenientes de la colección de Schler (Sección3.5.1). Estas frases describen actividades que las personas comúnmente realizan al despertar en la mañana. Como puede notarse en la Tabla 4.2, cada persona tiene su propio estilo de escritura e intereses temáticos. Sin embargo, mediante un análisis de qué se está escribiendo y cómo se está escribiendo, es posible encontrar patrones discriminativos de perfiles (e.g., hombres versus mujeres, o jóvenes versus adultos). Por ejemplo, las siguientes deducciones o generalizaciones se

Tabla 4.2: Ejemplos de frases personales

Fragmento de texto	Género	Edad
"And then I woke up at 11:00 & took a *shower* & got *dressed*. Then I was gonna fix my *hair* & put on my *makeup* & *mom* said there was no use in goin because it was late anyway.. So I didn't go"	Mujer	15
"I woke up Sunday morning and *cleaned* up the *house*. I have decided not to run away, just yet. Once the *house* was *cleaned* I took a long *bath* and *washed* my *hair* and gave it an intensive *conditioning treatment*."	Mujer	41
"I woke up, *ate*, and helped *Dad* in the *basement*. Then at *lunchtime* I *ate* again. At two I had this thing at the *library* where they showed you how to make stuff out of duck tape. Most of it I already knew."	Hombre	13
"Wow what a day! I woke up about 11:30 to a great *breakfast* of *tacos!! Beef, egg, cheese* and *salsa sauce* to be precise, *yummmm!*"	Hombre	15
"I woke up this morning feeling great. I went to the *kitchen*, *fried* me a *hamburger patty*, and some *eggs*. There were a few *dishes* that needed to be *washed* so I *washed* them. I came back up stairs, picked up my *room*, and made my *bed*. It is great to be alive and sober."	Hombre	44

obtienen de los ejemplos mostrados en la Tabla anterior: los hombres hablan más de comida que las mujeres, mientras las mujeres se expresan más acerca del cuidado personal, específicamente de su cabello. Por otro lado, también puede ser notado que los jóvenes tienden a mencionar a sus padres o a escribir informalmente ("..", "&" or "!!").

En contraste, cuando las personas usan frases sin pronombres personales, las cuales son referidas como Frases No Personales (FNP), están hablando acerca de objetos u otros individuos; por consiguiente, estas frases son imprecisas para capturar información del perfil de sus autores. Para ilustrar el ligero valor

de los contextos no personales, se muestran algunos ejemplos en la Tabla 4.3, donde los autores son los mismos de la Tabla 4.2.

En uno de los ejemplos de la Tabla 4.3, una mujer de 41 años está expresando información sobre un varón (tal vez su hijo, el cual es un tópico común en este rango de edad), pero ella claramente no está revelando parte de su perfil, ya que está hablando acerca de los intereses o preferencias de otra persona más que de los propios (quizá a ella no le guste la ciencia o las matemáticas). Es importante notar que en la redacción retiene su estilo de escritura, por ejemplo, continúa usando "..." como en la Tabla 4.2. De esta manera, se aprecia que las características de estilo son igualmente observadas en frases personales y no personales. Por ejemplo, la mujer de 15 años usa la expresión ".." en ambos tipos de frases. Con estos ejemplos, se muestra que las FNP son moderadamente valiosas para la identificación de perfiles, pues son menos informativas que las personales.

Tabla 4.3: Ejemplos de frases no personales correspondientes a los mismos usuarios de la Tabla 4.2

Fragmento de texto	Género	Edad
"It's a pretty good movie. It's all about hockey though.. kinda boring"	Mujer	13
"The boy is going on an excursion to a science centre. He is so excited. He loves anything sciencey and mathematical. "	Mujer	41
"The Guy had oleyed up to grind and when he landed on the rail cracked right in two."	Hombre	13
"Did my sister yell at him? Smack his ass? Nope. She said "he didn't do it on purpose!" and came to his defense, comforting him "	Hombre	35

Si bien, con base en estas observaciones, se puede concluir la usabilidad de ambos tipos de frases, también se ha destacado la relevancia de las FP. De ahí

que las siguientes secciones se concentren en estudiar el rol de este tipo de frases en AP.

4.2 Evaluando el rol de las frases personales en la identificación del perfil del autor

Esta sección presenta la metodología diseñada para investigar el rol de las FP en AP. El interés es responder las siguientes preguntas de investigación:

- ¿Es toda la información en un documento igualmente relevante para AP? Particularmente, ¿son las frases pesonales más discriminativas que las otras frases?
- ¿Las frases personales conteniendo pronombres en primera persona del singular y del plural son igualmente relevantes para AP? ¿Son complementarias o redundantes?
- ¿Las frases personales exponen mejor el estilo de escritura o intereses temáticos de los autores?
- ¿Son las frases personales igualmente relevantes en diferentes medios sociales?

Para responder las preguntas es importante comparar el desempeño de la clasificación de perfiles de usuarios cuando se usan únicamente FP contra el uso de los documentos enteros e incluso contra el uso de FNP. Como consecuencia, los resultados de la clasificación permitirán deducir la importancia de las FP en la tarea. Por lo tanto, en la Sección 4.2.1 se describe

el proceso que filtra las FP de un documento. Posteriormente, en la Sección 4.2.2 se detalla la configuración del proceso de clasificación usado en todos los experimentos de la Sección 4.3.

4.2.1 Proceso de filtrado

El proceso de filtrado considera la extracción de todas las FP que aparecen en cada documento de un corpus dado, tal como se ilustra en la Figura 4.1. Como se muestra en la Figura, primero cada documento es dividido en sentencias. Enseguida, se seleccionan las sentencias que incluyen algún PP en primera persona. El resto de las frases son descartadas. En los experimentos, se hace referencia a esos subconjuntos de frases como *corpus filtrado* y *corpus complemento*, respectivamente. Es importante notar que podrían existir documentos sin FP, lo cual propiciaría archivos filtrados vacíos. En tales situaciones el documento original es usado en su lugar. De esta manera, se evitan decisiones azarosas de los clasificadores cuando no hay información para decidir.

Las FP contenidas en el corpus filtrado conforman la **esencia** de los documentos para AP. La idea es clasificar con dicha esencia mostrando así que constituye la información más valiosa para la tarea. Por consiguiente, el proceso de filtrado fue integrado dentro del método clásico de solución de las tareas del análisis de autoría, tal como se muestra en la Figura 4.2. Observe que el proceso de filtrado se aplica a los textos de entrenamiento así como a los que se usarán en la fase de prueba, los procesos siguientes son aplicados

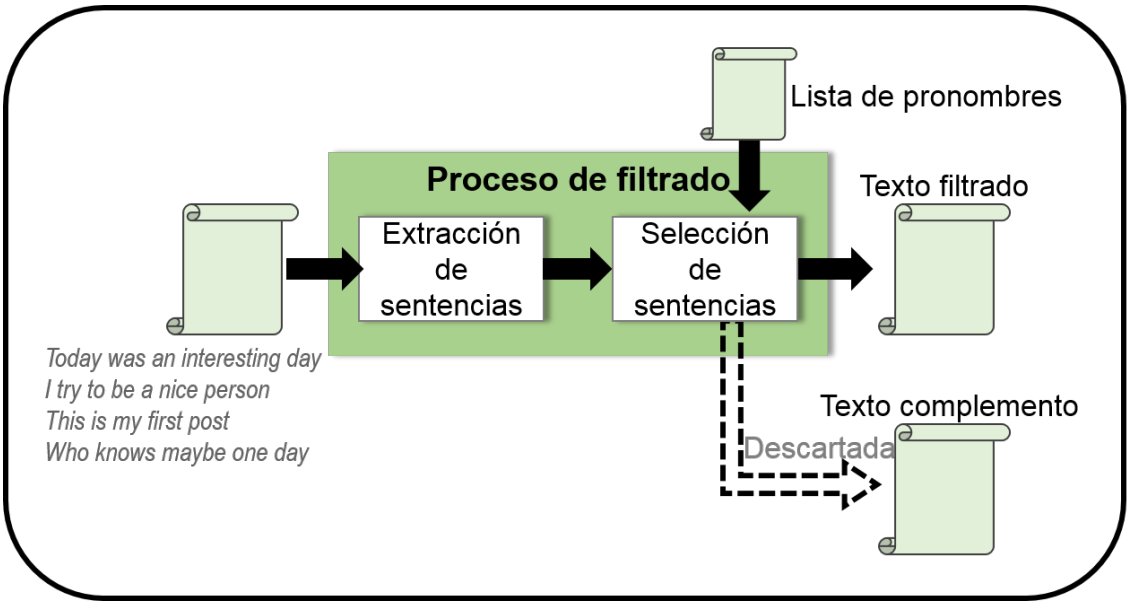


Figure 4.1: Proceso de filtrado

en la forma tradicional. Es importante notar que la representación proveniente de los textos filtrados alimenta un proceso de aprendizaje y clasificación. La configuración de tal proceso es explicada a continuación.

4.2.2 Proceso de clasificación

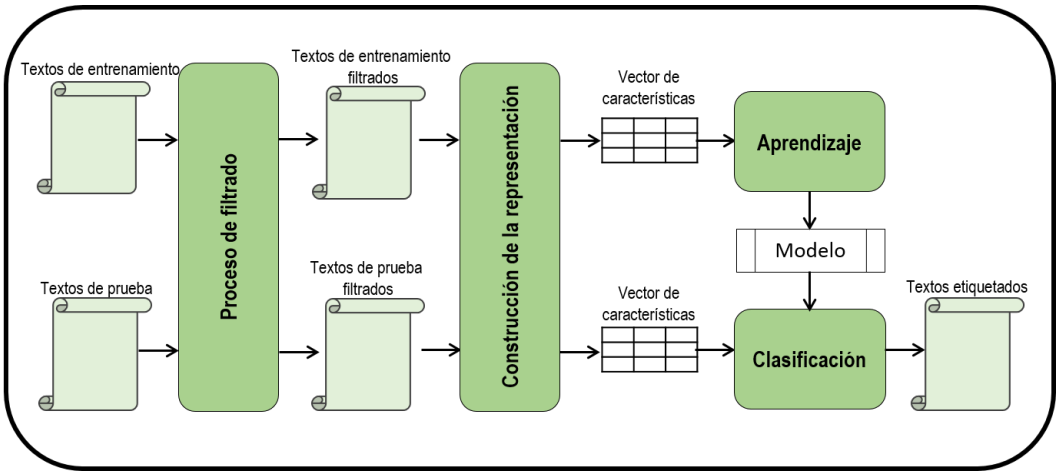
Los experimentos mostrados en este capítulo siguen un marco estándar de clasificación para AP:

- Términos. Se utilizó el conjunto de atributos descrito en (Schler et al., 2006; López-Monroy et al., 2015): 1000 palabras de contenido[1] con la más alta ganancia de información, palabras de función[2], signos de

[1] Las palabras de contenido fueron extraídas mediante expresiones regulares (López-Monroy et al., 2015).

[2] Se usó una lista de 467 palabras de función (Argamon et al., 2003).

Figure 4.2: Integración del módulo de filtrado



puntuación, palabras coloquiales (*slang words*) y etiquetas de partes de la oración (POS)[3].

- Representación. Usando todas las características mencionadas en el inciso anterior, se construye una bolsa de palabras. Los pesos de los términos corresponden a su frecuencia normalizada (TF) con respecto al número de términos en el documento.

- Clasificador. Para clasificar los documentos, se utilizó el clasificador SVM de la librería LIBLINEAR (Fan et al., 2008) usando parámetros de optimización por defecto.

- Evaluación. Se aplicó validación estratificada de 10 capas en cada experimento, y se usó la exactitud como principal medida de evaluación. Para evaluar la diferencia estadística entre las diferentes configuraciones de los conjuntos de datos (original, filtrado y complemento) se aplicó la prueba pareada t sobre las 10 capas con un nivel de significancia de 0.05.

[3]Las etiquetas POS fueron obtenidas usando el etiquetador de Stanford.

Esta prueba fue usada en todos los experimentos de este capítulo con base en la exactitud (ver Sección 2.1.4.3)

4.3 Experimentos y resultados

En esta sección se describen los experimentos realizados para responder a las preguntas de investigación expuestas en la Sección 4.2. Los experimentos emplean la metodología y características mencionadas anteriormente. La siguiente lista muestra los rasgos explorados y sus categorías:

- Edad: con tres categorías (10s, 20s, 30s)
- Género: con dos categorías (Hombre, Mujer)

4.3.1 Experimento 1: relevancia de frases personales

El objetivo de este experimento consiste en determinar el valor de las FP para identificar el perfil de autores. Basado en la idea que las personas expresan sus intereses y estilo de escritura cuando hablan acerca de ellas mismas, este experimento está enfocado en evaluar el rol de las frases que contienen *PP* en primera persona del singular.

Para llevar a cabo esta evaluación, se usó la colección de datos de Schler, descrita en la Sección 3.5.1. Primero, se filtraron las FP que contienen alguno de los pronombres siguientes: *I, me, mine, my, myself*, así como también la cadena *im*, ya que su uso es muy común en textos de redes sociales. Las estadísticas de filtrado se muestran en la Tabla 4.4, en donde se observa que el

corpus filtrado representa el 48.12% de la información de la colección original y tiene menos información que su corpus complemento. También se muestra el número de archivos vacíos que resultaron del proceso de filtrado y que fueron sustituidos por el archivo original. La cantidad de este tipo de archivos refleja que en la mayoría de los documentos existe información personal. A su vez, indica que la información sumada por el reemplazo de los archivos no es significativa.

Tabla 4.4: Resultados en las colecciones filtradas y complemento correspondientes a blogs

Corpus	Número de Sentencias	Archivos Vacíos	Edad	Género
Original	9,155,301	0	77.49	80.07
Filtrado	4,405,783	69	76.09	79.63
Complemento	5,510,302	131	69.98	72.59

Para evaluar la relevancia de las FP, se comparó la exactitud de la clasificación de edad y género cuando se usan las colecciones generadas y que corresponden a la primer columna de la Tabla 4.4. Las últimas dos columnas muestran los resultados obtenidos. Es importante destacar que los resultados obtenidos usando el corpus filtrado son significativamente mejores que esos correspondientes al corpus complemento, aunque hay menos información en el primero. Esto indica que la información personal de los autores es, en efecto, más importante que la información no personal. Además, esos resultados también muestran que usando únicamente las FP es posible lograr un desempeño muy similar que usando los documentos completos. En la predicción de género no hay una diferencia significativa entre los resultados usando el corpus filtrado o el original.

En concreto, los resultados, por un lado, confirman la relevancia de las FP para AP. Por el otro lado, aseguran que las FP conforman la esencia de los documentos para esta tarea.

4.3.2 Experimento 2: el valor agregado de las frases personales en plural

Este experimento examina el rol de las frases con pronombres en primera persona del plural en AP. Particularmente, se enfoca en investigar si estas frases, las cuales tienen naturaleza inclusiva y expresan información acerca de los usuarios como parte de un grupo, podrían mejorar la clasificación.

Al igual que en el experimento previo, se utilizó el corpus de Schler como colección de referencia. Sin embargo, en este caso, se consideraron frases con pronombres en la primera persona del singular, así como del plural. Por consiguiente, en el proceso de filtrado para considerar las frases en plural se extrajeron sentencias conteniendo alguno de los siguientes pronombres: *we*, *us*, *our*, *ours*, *ourselves*. Algunas estadísticas obtenidas son mostradas en la Tabla 4.5.

Tabla 4.5: Resultados usando frases personales en plural

Corpus	Número de sentencias	Archivos vacíos	Edad	Género
Original	9,155,301	0	77.49	80.07
Filtrado singular-plural	4,943,390	33	76.99	79.82
Filtrado plural	908,815	1075	67.00	70.35
Filtrado Singular	4,405,783	69	76.09	79.63

Es importante observar, que existen considerablemente menos frases con pronombres en primera persona del plural, que con pronombres en primera persona del singular, lo cual podría ser explicado por el tipo de información compartida en blogs. En adición, puede ser notado que su combinación únicamente causó un incremento de 537,607 frases (5.9%) sobre el corpus filtrado singular, indicando la frecuente co-ocurrencia de pronombres de la primera persona del singular y del plural en publicaciones de redes sociales.

La Tabla 4.5 muestra la exactitud obtenida de la clasificación usando diferentes configuraciones del corpus filtrado. Los resultados correspondientes al uso de las frases personales en singular mejoran el desempeño considerablemente con respecto a las frases en plural. Las diferencias fueron de 9.1% y 9.3% para edad y género, respectivamente. Estas diferencias podrían ser atribuidas a la diferencia en los tamaños de las colecciones, pero también sugieren que las frases personales en plural cambian su enfoque de los intereses particulares de las personas a los intereses de los grupos.

Por otro lado, la prueba de significancia estadística indicó que las diferencias de exactitud observadas entre el corpus filtrado singular-plural y el corpus filtrado singular no fueron estadísticamente significantes para las tareas de predicción de edad y género. Estos resultados permiten concluir que las frases personales en plural no tienen una relevancia especial para AP. Más aún, los resultados también corroboran la utilidad destacada de las frases personales en singular para esta tarea. Por lo tanto, es importante señalar que en esta tesis las siglas FP hacen referencia a frases personales en singular.

4.3.3 Experimento 3: información del contenido y estilo en frases personales

Los experimentos previos han mostrado el importante rol de las FP para la identificación del autor. El propósito de este experimento fue entender el poder de discriminación de estas frases. Particularmente, se quiere determinar la contribución de información de contenido y estilo de esas frases para AP.

Particularmente, en este experimento, se dividieron los términos referidos en la Sección 4.2.2 en tres conjuntos disjuntos: palabras, las cuales representan información de contenido, palabras de función y POS, que señalan información del estilo. Para evaluar la relevancia de cada tipo de términos, se comparó su exactitud de clasificación cuando se usa el corpus filtrado y el complemento. La Tabla 4.6 muestra los resultados.

Tabla 4.6: Resultados según el tipo de términos

Tipo de términos	Corpus	Exactitud	
		Edad	Sexo
Palabras	Original	76.06	78.12
	Filtrado	75.04	78.08
	Complemento	68.49	71.19
Palabras de función	Original	68.56	73.05
	Filtrado	67.00	70.78
	Complemento	61.31	67.56
POS	Original	63.09	6811
	Filtrado	62.87	66.35
	Complemento	59.79	65.68

Los resultados confirman conclusiones de trabajos previos, las cuales han señalado que la información de contenido es más relevante que la información

de estilo para AP (Schler et al., 2006). Los resultados también muestran que la diferencia de desempeño entre el corpus original y filtrado es más bajo en el espacio de las palabras, demostrando que los intereses temáticos son adecuadamente capturados en frases personales.

Por otro lado, comparando los resultados del corpus filtrado y el complemento, es posible observar una diferencia promedio de 6.7% en favor del corpus filtrado usando las palabras como características, mientras las diferencias cuando se utilizaron palabras de función y etiquetas POS corresponden a 4.4% y 1.9%, respectivamente.

Estos resultados sugieren que el valor de las FP recae principalmente en el aspecto de contenido más que en la información acerca del estilo. Se puede concluir que la información del estilo de redacción de los autores podría ser igualmente capturada tanto en FP como en FNP, ya que ambos tipos de frases son escritas por el mismo autor. Sin embargo, los intereses temáticos de los autores son mejor capturados en las FP.

4.3.4 Experimento 4: información necesaria

Este experimento está orientado a determinar si una pequeña muestra de documentos con alta concentración de FP es suficiente para lograr resultados comparables al uso todos los documentos. Para ello, se emplearon subconjuntos de documentos del corpus de Schler que contenían mayor número de FP: 100, 200, 500, 1,000, 2,000, 4,000, 6,000, 10,000, 14,000 y 16,000.

Los resultados se muestran en la Figura 4.3. La curva de color rosa confirma que un número pequeño de instancias (documentos) con gran cantidad de FP generan un desempeño aceptable y muy superior al uso del mismo número de documentos pero con menor número de FP (curva verde). Por ejemplo, con sólo 4000 instancias que representa un 20.70% del total, obtenemos 74.71% que es sólo 5.36 puntos menor que la exactitud usando el total de instancias. Naturalmente, mientras aumenta el número de instancias consideradas, las dos curvas convergen. Los resultados determinan un criterio de construcción

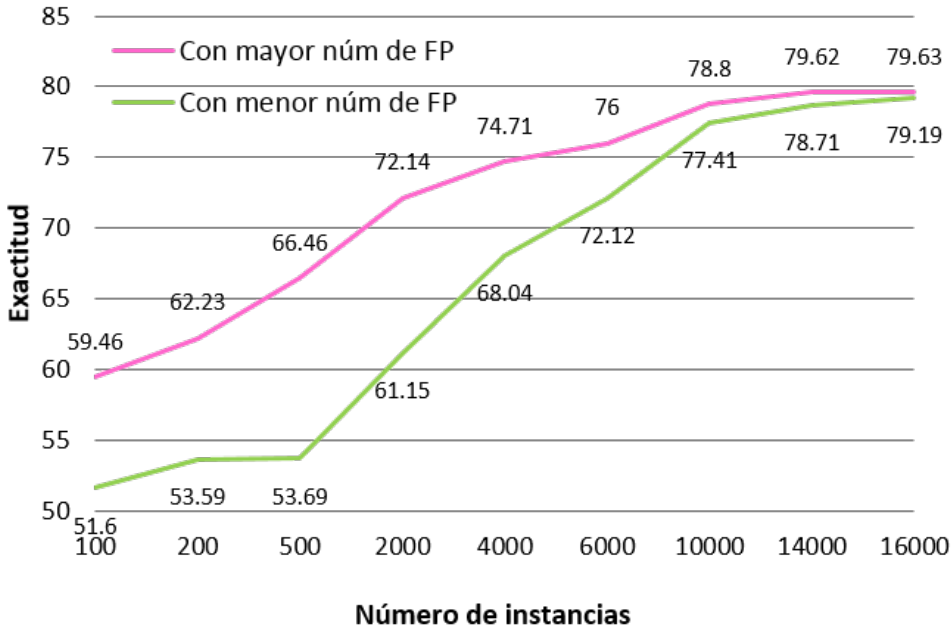


Figure 4.3: Clasificación de género usando las instancias con mayor número de frases personales.

de corpus que permite ahorrar tiempo de etiquetado. Específicamente, las instancias a etiquetarse corresponderían a aquellas con mayor presencia de FP, pues son las que concentran mayor cantidad de información valiosa para la tarea.

4.3.5 Experimento 5: frases personales en diferentes dominios de redes sociales

El propósito de este experimento es evaluar la relevancia de las frases personales en AP usando documentos provenientes de diferentes redes sociales. De esta manera, se analiza la generalidad de los hallazgos encontrados previamente y el grado de independencia de la propuesta hacia el dominio de aplicación. Para este experimento se usó el corpus PAN-AP-2014 (ver Sección 3.5.2) que contiene diversas colecciones de documentos. El corpus filtrado fue construido seleccionando las frases que contienen PP en singular como se detalló en la Sección 4.2.1. La Tabla 4.7 muestra algunos datos obtenidos usando las colecciones señaladas.

Tabla 4.7: Estadísticas del proceso de filtrado en el corpus PAN-AP-2014

Colección	Frases en el corpus original	Frases en el corpus filtrado	Archivos vacíos
Blogs	22,994	5,565	10
Twitter	318,691	49,540	7
Reviews	52,833	19,248	1,377
Social Media	3,207,509	736,615	1,349

En todas las colecciones se trataron dos problemas de clasificación: predicción de edad con cinco categorías (18-24, 25-34, 35-49, 65 o más) y predicción de género con dos categorías (hombre, mujer). Los resultados se muestran en la Tabla 4.8 y confirman valores de exactitud similares cuando el corpus filtrado o el original están siendo utilizados. Pero, es muy importante remarcar que las

colecciones filtradas representan un pequeño subconjunto (de 15% a 36%) del corpus original.

Tabla 4.8: Resultados usando sólo sentencias personales en las colecciones del PAN 2014

Colección	Corpus	Exactitud		% en el corpus filtrado
		Edad	Género	
Blogs	Original	36.56	68.42	24.20% (de 22,944 frases)
	Filtrado	43.92	62.14	
Twitter	Original	35.33	71.33	15.54% (de 318,691 frases
	Filtrado	37.49	59.55	
Reviews	Original	30.84	67.24	36.43% (de 52,833 frases)
	Filtrado	29.21	65.21	
Social Media	Original	34.84	53.64	22.97% (de 3,207,509 frases)
	Filtrado	33.99	52.68	

Particularmente, las pruebas de significancia estadística indicaron que los resultados para la predicción de edad fueron comparables a través de todos los dominios considerados, mientras para la clasificación de género, se encontró una diferencia estadísticamente significante para Twitter y Blogs. Sin embargo, es importante notar que para esas dos colecciones se obtuvieron mejores resultados en la predicción de edad usando el corpus filtrado que empleando el corpus original, lo cual causa un desempeño total comparable.

Los resultados en la Tabla 4.8 son muy alentadores, principalmente porque la predicción de edad en esas colecciones considera cinco categorías con valores consecutivos y, por lo tanto, representa un problema de clasificación más difícil que cuando se usa el corpus de Schler. En general, estos resultados soportan la relevancia de las frases personales. Incluso también reafirman su rol como la esencia de los documentos para AP.

4.4 Resumen y discusión

En este capítulo se investigó la relevancia de las frases personales en la tarea de AP. La idea base considera que las personas reflejan mejor sus características y estilo de escritura cuando escriben acerca de ellas mismas. Si bien, existen enfoques actuales que usan algunos pronombres personales como atributos de la representación, en esta investigación se exploró su contexto (términos ocurriendo en la misma frase) en la tarea.

Los experimentos reportados claramente indican que las frases personales tienen información de alto valor para predecir la edad y el género de usuarios de redes sociales. Considerando únicamente este tipo de frases, se obtienen reducciones de hasta 60% de la información en los documentos y un desempeño comparable al uso de toda la información. Por consiguiente, las frases personales pueden ser consideradas la *esencia* de los documentos para la identificación del perfil de los autores, pues es en la información personal donde se concentran los datos discriminativos de los perfiles.

El término *esencia* ha sido utilizado en el contexto de clasificación tradicional de textos para indicar que no todas las partes de un documento son relevantes (Mihalcea y Hassan, 2005; Shen et al., 2007; Anguiano-Hernández et al., 2010). Por ejemplo, Mihalcea y Hassan usan este término para nombrar hacer referencia a los resúmenes de los documentos. Así pues, clasificando la esencia y no los documentos enteros, se logra remover las secciones menos importantes o ruidosas antes de la clasificación generando buenos resultados.

Particularmente, a la fecha, no se había explorado el uso de frases personales como la esencia de los documentos en AP.

Por otro lado, en este capítulo se respondieron algunas preguntas de investigación expuestas al inicio del mismo, encontrando que:

- No toda la información de un documento es igualmente relevante para AP. Las frases personales son más discriminativas que las no personales.
- A pesar de la naturaleza inclusiva de las frases personales en plural, ellas no tienen especial relevancia para AP y su información tampoco es complementaria de las frases personales en singular.
- Las frases personales capturan mejor la información de contenido (intereses de los usuarios), mientras la información de estilo puede ser igualmente extraída de frases personales como no personales
- La relevancia de las frases personales es una característica que fue observada en diferentes dominios sociales.

Asimismo, considerando la calidad de los documentos a través de su esencia, se establece un criterio para la construcción de colecciones de documentos etiquetados para AP. Los esfuerzos y tiempo de etiquetado estarían concentrados a analizar únicamente documentos cuya esencia sea significativa.

Una vez confirmado el importante rol de las frases personales, surge el interés de medir, cuánta información personal está expresando cada término. Para lograr tal fin, se propone una novedosa medida descrita en el siguiente capítulo.

Capítulo 5

Índice de expresión personal

Este capítulo está dirigido a determinar el grado de asociación de los términos con la información personal de los autores. Considerando que las frases personales son los contextos más valiosos de los documentos para inferir rasgos de los autores, se proponen tres nuevas medidas: precisión personal, cobertura personal y el índice de expresión personal. Tales medidas evalúan el tipo de contexto donde ocurren los términos y asignan especial importancia a aquellos términos frecuentes en frases personales.

5.1 Criterios generales

Se ha mencionado previamente que la mayoría de los términos en las frases personales exponen información relacionada con el perfil del autor. Para determinar la cantidad de información personal revelada por cada término, es necesario considerar el contexto de los términos, es decir, el tipo de frases donde aparecen. Un documento d_j está formado por un conjunto de frases S_j, el cual a su vez está compuesto por los subconjuntos P_j y N_j, que representan los subconjuntos de frases personales y no personales, respectivamente, tal como se muestra en la Figura 5.1. Por lo tanto, un término t_i puede aparecer en el subconjunto P_j y/o en N_j.

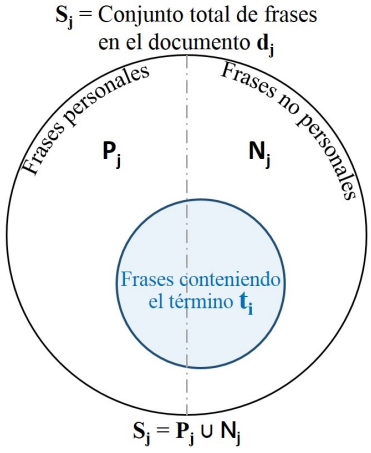


Figure 5.1: Conjuntos de frases existentes en un documento

Considerando tal notación, así como la función $\#(t_i, X)$ que indica el número de frases donde el término t_i aparece en el conjunto X del documento d_j (X corresponde a S_j, N_j o P_j), se proponen tres medidas que retoman ideas provenientes de recuperación de información (Salton y Buckley, 1988):

precisión personal, *cobertura personal* y el *índice de expresión personal (Personal Expression Index, PEI, por sus siglas en inglés)*.

5.1.1 Precisión personal (ρ)

Es una medida que estima la concentración de información personal revelada en el contexto de un término. Esta medida corresponde a la expresión 5.1 y es definida como el porcentaje de frases personales conteniendo el término t_i dentro del documento d_j:

$$\rho(t_i, d_j) = \frac{\#(t_i, P_j)}{\#(t_i, S_j)}. \tag{5.1}$$

En otras palabras, ρ corresponde a la probabilidad de ocurrencia del término en las frases personales de un documento.

5.1.2 Cobertura personal (τ)

Es una medida que cuantifica la porción de frases personales de un documento (i.e., la porción de su “esencia") cubierta por el término t_i y corresponde a la ecuación 5.2:

$$\tau(t_i, d_j) = \frac{\#(t_i, P_j)}{|P_j|}. \tag{5.2}$$

Puede ser interpretada como la probabilidad condicional de la ocurrencia de un término dado el conjunto de frases personales.

5.1.3 Índice de expresión personal

Aunque ρ y τ son medidas cuyo valor incrementa cuando el número de ocurrencias en las frases personales es más grande, su comportamiento puede ser opuesto. Por ejemplo, un término apareciendo una sola vez en un documento y particularmente en una frase personal, obtendría un valor muy alto de precisión personal (ρ), pero no necesariamente alta cobertura (τ), principalmente, porque el documento puede estar formado por varias frases personales. Por el contrario, un término apareciendo en la única frase personal de un documento conseguiría el más alto valor para ρ, independientemente de sus ocurrencias en frases no personales.

Desde luego, es natural preguntarse: ¿cuál medida es más importante para la tarea, la precisión o la cobertura personal? En realidad, ambas son indispensables. De ahí que, se propone estimar el balance de tales medidas en un solo valor a través de otra medida propuesta en este trabajo de investigación: el índice de expresión personal (PEI).

La medida PEI, presentada en la ecuación 5.3, es una media armónica de ρ y τ e indica que entre más frecuente es la ocurrencia de un término en frases personales y menos frecuente en las frases no personales, el término es más revelador del perfil del autor:

$$PEI(t_i, d_j) = 2\frac{\rho(t_i, d_j)\,\tau(t_i, d_j)}{\rho(t_i, d_j) + \tau(t_i, d_j)}. \tag{5.3}$$

Entonces, los términos más valiosos son aquellos con alta precisión personal así como alta cobertura. Por lo tanto, un término que ocurre exclusivamente en todas las frases personales de un documento producirá el máximo valor posible (PEI=1), lo cual indica que el término es parte del contexto personal del autor y que existe una alta probabilidad de que se encuentre revelando algún rasgo de su perfil. En cambio, valores bajos de PEI indican que el término no es un elemento representativo del perfil del autor.

Análogamente a PEI, en esta tesis se presenta también el concepto opuesto, el cual es descrito a continuación.

5.1.4 Índice de expresión no personal

El índice de expresión no personal (*Non Personal Expression Index*, NEI, por sus siglas en inglés) considera las ocurrencias de los términos en el subconjunto de frases no personales (N_j) y está dirigida a capturar el nivel de asociación de cada término a la expresión de información no personal. Por lo tanto, NEI es una combinación de precisión no personal ($n\rho$) y cobertura no personal ($n\tau$), tal como se muestra en la siguiente ecuación:

$$NEI(t_i, d_j) = 2\frac{n\rho(t_i, d_j)\, n\tau(t_i, d_j)}{n\rho(t_i, d_j) + n\tau(t_i, d_j)}, \tag{5.4}$$

donde:

$$n\rho(t_i, d_j) = \frac{\#(t_i, N_j)}{\#(t_i, S_j)}, \tag{5.5}$$

$$n\tau(t_i, d_j) = \frac{\#(t_i, N_j)}{|N_j|}. \tag{5.6}$$

Es importante observar que el numerador en las ecuaciones 5.5 y 5.6 corresponde al número de frases donde ocurre el término dentro del conjunto N_j. Por lo tanto, los términos con altos valores de NEI representan información difícilmente valiosa para revelar rasgos del perfil de los autores.

5.2 Ejemplo del uso del índice de expresión personal

En la Tabla 5.1 se muestra una publicación de una fémina cuya edad se ubica en el intervalo de 50-64 años. Esta publicación[1] fue extraída de la colección PAN-AP-corpus 2014 en la sección de críticas de hoteles (Reviews). Las palabras con mayor PEI están marcadas con negrita.

Tabla 5.1: Ejemplos de términos con índice de expresión personal mayor a 0

*Loved this **Hotel** This was a **trip** to my **retirement** and 3 other **friends** came along we all **loved** this **Hotel**. It's why I always try to stay at **Marriotts**. **Yhe Hotel** was right on the **beach**, very **nice** clean **beach** I could have **stayed right** there the **whole trip**. The resturant food was great, and the rooms were clean, roomy, and very comfortable. I will tell all my friends about this Hotel and I will **celebrate** my **Birthday** there next year.*

[1] La redacción e inclusive errores gramaticales corresponden a la publicación original.

Para este ejemplo, los valores de PEI corresponden a la lista de pesos de la Tabla 5.2. Observe que los términos con valores de PEI mayores que cero, son palabras que señalan intereses de personas que comparten rasgos socio-demográficos de la autora. De ahí, PEI podría sugerir aquellas palabras que describen el perfil del autor. Por ejemplo, *retirement* es un término característico de las personas pertenecientes a ese grupo de edad. Así también, *niced, loved* son palabras asociadas frecuentemente a mujeres. Por el contrario, el término *food* es generalmente usado por el género masculino (ver Sección 4.1), lo cual explica el valor nulo para PEI y un valor igual a la unidad para NEI.

Es necesario señalar que PEI es una estimación del grado de asociación de los términos a los intereses de los autores, información que es de gran ayuda para distinguir perfiles. Por lo tanto, en los siguientes capítulos, se explotará el uso PEI e incluso NEI para crear esquemas de selección y pesado de términos que se beneficien de tal información.

5.3 Resumen y discusión

En este capítulo se presentó el concepto PEI, una medida propuesta en esta investigación para cuantificar la asociación de un término con la información personal de los autores. La medida corresponde a la media armónica de dos factores también propuestos en esta tesis: precisión y cobertura personal. Los valores de esta medida varían en un rango de 0 a 1. Donde el 0 indica que el término no estuvo presente en los contextos personales del documento, por lo

Tabla 5.2: Ejemplos de términos con valores de PEI y NEI

Palabra	PEI	NEI
hotel	8.86	0
trip	0.67	0
friends	0.67	0
celebrate	0.67	0
retirement	0.4	0
right	0.4	0
year	0.4	0
yhe	0.4	0
Marriotts	0.4	0
stayed	0.4	0
beach	0.4	0
nice	0.4	0
stay	0.4	0
birthday	0.4	0
loved	0.4	0
try	0.4	0
whole	0.4	0
clean	0.33	0.67
comfortable	0	1
rooms	0	1
roomy	0	1
food	0	1
great	0	1
resturant	0	1

que con alta probabilidad no está expresando información del perfil del autor. Por el contrario cuando su valor es 1, señala que el término está muy asociado al contexto personal y por lo tanto, será un atributo valioso para la predicción.

En este capítulo se ha mostrado que las palabras con valores altos de PEI corresponden términos que expresan intereses de los autores que pueden revelar su perfil. Lo anterior no quiere decir que únicamente los términos con un alto valor de PEI sean predictivos. Existen otros términos cuya forma de

uso permite distinguir perfiles, tal es el caso de las palabras de función cuya frecuencia de uso expresa información acerca del estilo de redacción de los autores más que de sus intereses incluso, tendrán un bajo valor de PEI. En consecuencia, la medida PEI indica cuáles términos tienen alta probabilidad de revelar los rasgos de los autores independientemente de su frecuencia de uso.

Asimismo, se diseñó la medida NEI como un concepto análogo a PEI que determina el grado de asociación de los términos con la información no personal de los autores. Mediante ambas medidas se logra estimar la calidad de las frases (contexto) donde ocurren los términos. Esta información cualitativa es considerada en la presente tesis para construir la representación de los documentos como se describe en los siguientes capítulos.

Capítulo 6

DPP: un nuevo esquema de selección de términos

La selección de términos es un proceso muy importante en las tareas de categorización supervisada. En esta tesis, se propone un novedoso esquema para seleccionar los términos más relevantes a partir del conjunto total de ellos. El esquema propuesto es llamado *pureza personal discriminativa*. En este capítulo, primero, se presentan las características generales y composición del esquema. Posteriormente, se muestran los experimentos que permiten validar su usabilidad en AP dentro del contexto de las redes sociales.

6.1 Definición del esquema propuesto DPP

Las técnicas de selección de términos reducen el conjunto completo de términos T a un subconjunto T' que contiene aquellos más relevantes para representar a los textos etiquetados. En esta tesis se diseñó el método *pureza personal discriminativa* (*Discriminative Personal Purity*, DPP, por sus siglas en inglés), el cual no sólo considera la distribución de los términos a través de las categorías, como la mayoría de las medidas tradicionales lo hace (e.g. ganancia de información), si no que también toma en cuenta el tipo de frases donde los términos aparecen. Integrando así, información cualitativa del contexto de los términos. Para ello, se incorpora el uso de PEI (descrito en el Capítulo 5), el cual es aprovechado con el objetivo de elegir los términos relacionados con los perfiles de los autores.

Formalmente, la función DPP es definida en la ecuación 6.1, como un producto de dos factores:

$$DPP(t_i) = \max_{k=1}^{|C|} \{PP_k(t_i)\} \; gini(t_i), \tag{6.1}$$

1. Un factor descriptivo, definido como el máximo valor de la función PP_k, la cual captura la capacidad de un término para describir información personal de los autores perteneciendo a una categoría específica c_k.

2. Un factor discriminativo, basado en el coeficiente Gini con el objetivo de calificar la habilidad de un término para discriminar las diferentes categorías (perfiles) de autores.

A continuación se describen ambos factores.

6.1.1 Pureza personal categórica como factor descriptivo

La pureza personal categórica de un término t_i en un categoría c_k, definida como $PP_k(t_i)$, evalúa la información personal capturada por el término en los documentos pertenecientes a esa categoría. Formalmente, PP_k está representada por la ecuación 6.2. Particularmente, es calculada como el cociente acumulativo de PEI entre NEI de todos los términos pertenecientes a los documentos de la categoría c_k:

$$PP_k(t_i) = \log_2 \left(2 + \frac{1}{2} \sum_{d_j \in c_k} \frac{PEI(t_i, d_j) + 1}{NEI(t_i, d_j) + 1} \right). \tag{6.2}$$

De esta manera, un término con valores de PEI mayores que NEI será premiado. Adicionalmente, la fórmula 6.2 tiene las siguientes características:

- Usa un tipo de suavizado aditivo (*Laplace smoothing*) para eliminar problemas relacionados con la división por cero. El cociente más pequeño es logrado cuando $PEI(t_i, d_j) = 0$ y $NEI(t_i, d_j) = 1$.
- Utiliza una multiplicación por la constante $1/2$ para asegurar valores mayores que 0 y menores que uno. Específicamente, en el intervalo [1/4,1].
- Aplica el logaritmo base 2 con la finalidad de reducir la amplia escala preservando el orden de las sumas, así como para tratar el problema de desbalance (es muy común tener multitud de documentos de algunos

perfiles pero muy pocos de otros) ya que el número de documentos pertenecientes a cada categoría puede estar desbalanceado[1].

- La constante 2 es sumada para garantizar valores mayores que 1 después del cálculo del logaritmo, especialmente, cuando el valor de la suma acumulativa es menor que 1.

6.1.2 Coeficiente Gini como factor discriminativo

El segundo componente en la ecuación 6.1, factor discriminativo denotado como $gini(t_i)$, estima la capacidad de un término para discriminar documentos de las diferentes categorías de autores. Para lograr tal fin, se evalúa la distribución de los términos en todas las categorías. Por ejemplo, la presencia concentrada de un término en sólo una de las categorías señala su pertinencia para lograr la discriminación. Por el contrario, las ocurrencias de un término igualmente distribuidas en todas las categorías indica un bajo nivel de discriminación.

Este segundo factor es determinado a través del coeficiente Gini, una medida que captura, en un solo valor, el nivel de concentración o desigualdad de cualquier distribución

[1]Otras técnicas podrían ser usadas, por ejemplo, dividir por el número de documentos en la categoría en lugar de aplicar logaritmo. Sin embargo, dos razones fueron consideradas para preferir logaritmo: 1) el balance/desbalance de categorías es considerado en el componente discriminativo 2) la intención es enfatizar términos con alta suma de valores independientemente del número de documentos en las categorías, principalmente porque no todos los textos tienen el mismo nivel personal.

Para estimar el coeficiente Gini, se analizó la distribución generada por un término t_i en el conjunto de categorías $C = \{c_1, ..., c_{|C|}\}$. Es decir: $\left\{\frac{\#(c_1,t_i)}{\#(c_1)}...\frac{\#(c_{|C|},t_i)}{\#(c_{|C|})}\right\}$, donde cada elemento representa la frecuencia relativa del número de documentos de la categoría donde el término t_i está apareciendo, dividida por el número total de documentos de la categoría.

Si bien el índice Gini ha sido calculado como una porción de área de Lorenz (Lorenz, 1905), existen formas alternativas para estimarlo. En esta investigación se aplicó la fórmula 6.3, mostrada por Dixon et al. (1988), cuyos rango de valores va desde 0 hasta 1, indicando completa igualdad o desigualdad, respectivamente:

$$gini(t_i) = \frac{1}{\mu \cdot |C|(|C|-1)} \sum_{k=1}^{|C|} (2k - |C| - 1) \frac{\#(c_k, t_i)}{\#(c_k)}, \tag{6.3}$$

donde $\#(c_k)$ y $\#(c_k, t_i)$ indican el número de documentos de la categoría c_k y el número de documentos de esa categoría conteniendo el término t_i respectivamente. Por lo tanto, $\frac{\#(c_k,t_i)}{\#(c_k)}$ representa la frecuencia relativa de t_i en c_k. Por su parte, μ denota la media de la distribución de las frecuencias relativas. Además, según Dixon et al. (1988), la ecuación 6.3 requiere que las categorías estén ordenadas de acuerdo con su frecuencia relativa, del menor al mayor valor (ascendentemente): $\frac{\#(c_1,t_i)}{\#(c_1)} \leq \frac{\#(c_2,t_i)}{\#(c_2)} \leq ... \frac{\#(c_{|C|},t_i)}{\#(c_{|C|})}$.

6.2 Experimento: evaluación de DPP

El propósito de este experimento es analizar la contribución del esquema propuesto DPP en la tarea de AP usando las colecciones descritas en la Sección 3.5. Para este propósito, se comparó el desempeño de la tarea usando dos esquemas de selección de términos: i) la pureza personal discriminativa (DPP) y ii) la tradicional ganancia de información (IG). Para ambos casos, se consideraron tres diferentes esquemas de pesado: TF (frecuencia normalizada), booleano y TF-IDF.

Los resultados de la comparación se muestran en la Figura 6.1 donde las columnas indican las exactitudes obtenidas cuando IG fue usada para seleccionar términos. A su vez, los guiones representan las exactitudes alcanzadas cuando DPP es usada como estrategia de la selección de términos. Estos resultados muestran mejoras importantes para todos los esquemas de pesado cuando DPP es usada, especialmente para el caso edad. En cambio, se percibe un desempeño similar para el problema de clasificación de género.

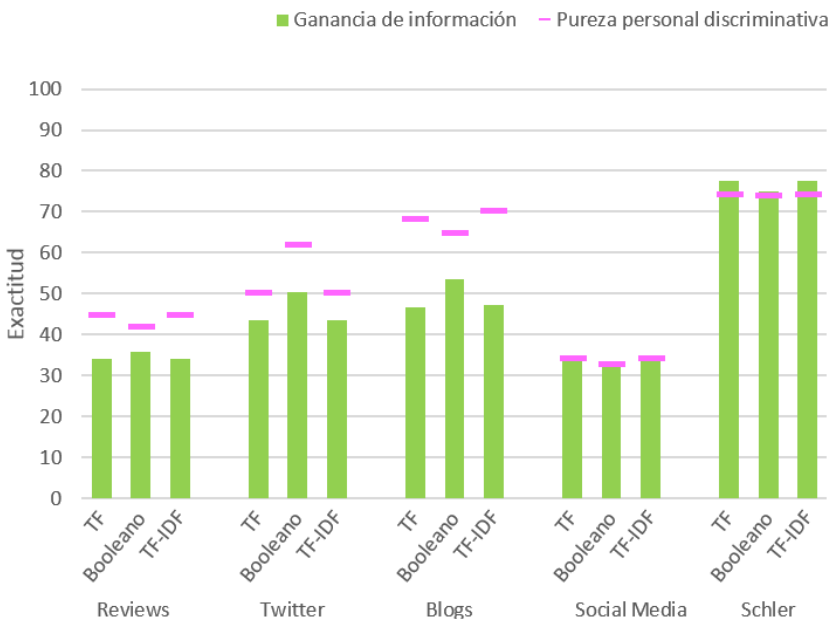
Los resultados sugieren que los términos personales son altamente relevantes para predecir la edad, los cuales son, en cierta medida, homogéneos con respecto al género de los usuarios (hombres y mujeres). Similarmente, el bajo desempeño (mínimas ganancias o incluso pérdidas) para la clasificación de género es un comportamiento esperado, pues es más difícil encontrar características textuales que apliquen a perfiles muy heterogéneos (por ejemplo, términos que caracterizan a mujeres de cinco diferentes rangos de

edad), especialmente cuando hay pocos datos para extraer tal evidencia (e.g., los conjuntos de datos correspondientes a Twitter y Blogs son muy pequeños).

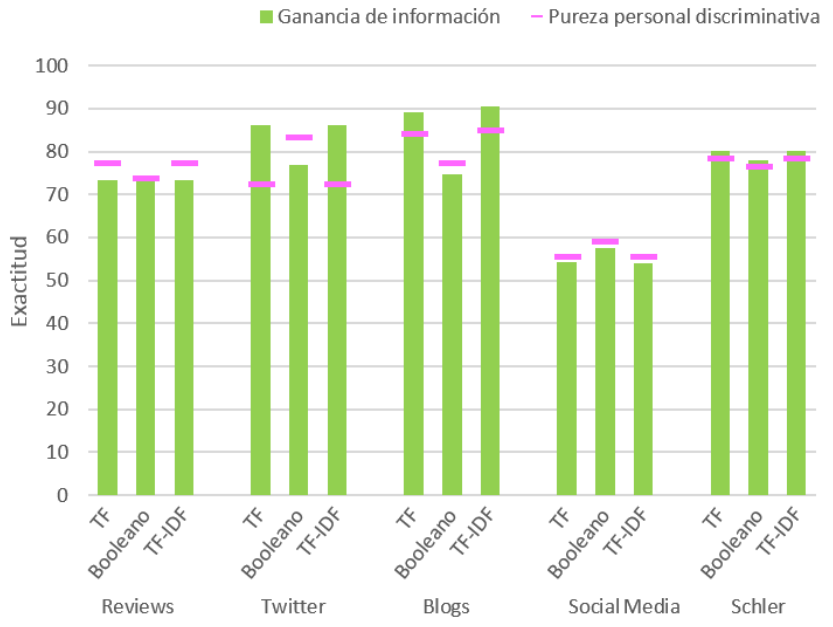
En general, los resultados de la Figura 6.1 indican que DPP es mejor que IG. De acuerdo con la prueba del rango con signo de Wilcoxon usando la exactitud y un nivel de significancia de 0.05 en los 15 conjuntos de datos (3 esquemas por cada una de las 5 colecciones), DPP es significativamente mejor que IG para el caso edad. Por lo tanto, se considera que DPP es un elemento clave para discriminar edad. En cambio, en el caso género los resultados son comparables.

Para profundizar el análisis del esquema DPP, la Figura 6.2 muestra algunas nubes de palabras con los 100 términos léxicos (palabras) mejor calificados por DPP de la clasificación de género en las colecciones de Reviews y Twitter, así como del caso edad en la colección de Blogs. En las nubes de palabras, el tamaño de la fuente corresponde a la posición del término dentro de las 100 mejores. El color verde y orientación horizontal representan las palabras seleccionadas por ambos esquemas: IG y DPP; mientras el color rojo y orientación vertical identifica las palabras seleccionadas únicamente por DPP.

A partir de la Figura 6.2, se observa que la mayoría de los términos más relevantes seleccionados por el enfoque propuesto fueron también elegidos por IG. Si bien, varios términos seleccionados únicamente por DPP no son frecuentes, ellos parecen ser términos intuitivos conocidos en la literatura de AP. Por ejemplo, para identificar adultos hay términos valiosos como: *newspaper*, *doctors* y *treatments*. A su vez, para la identificación de género las palabras



(a) Edad



(b) Género

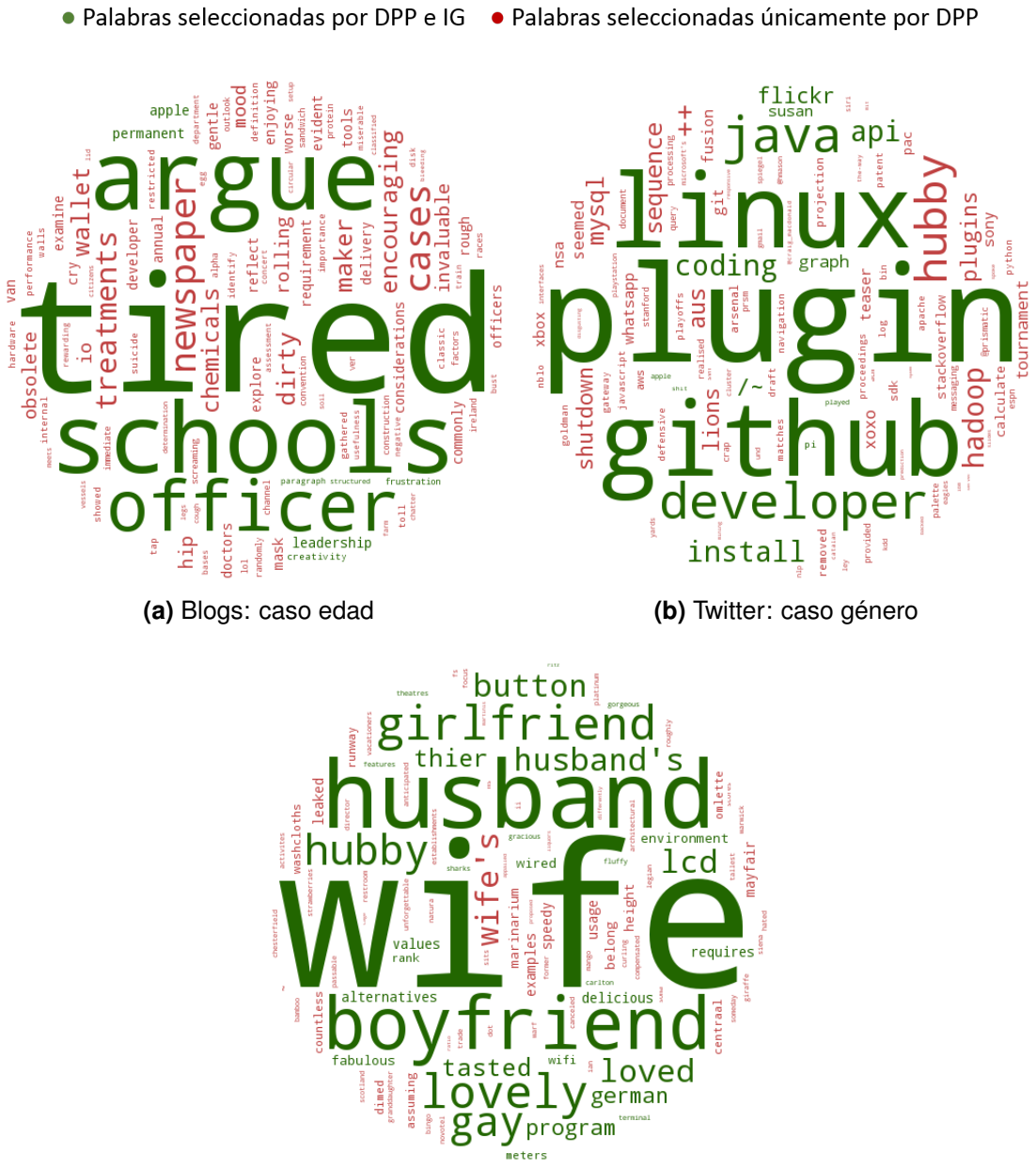
Figure 6.1: Comparación de la pureza personal discriminativa y ganancia de información.

mysql, haddop o *plugins* son muy cercanas a tópicos de tecnología, los cuales han sido asociados con el género masculino. Finalmente, palabras como *xoxo, aws, hubby* han mostrado ser de gran ayuda para la identificación de mujeres. Es importante notar que DPP califica cada término existente en el vocabulario asignando valores mayores que cero. De esta manera, DPP enriquece la selección incluyendo varios términos relacionados con expresiones personales que ganancia de información califica como no informativos (valor igual a cero).

6.3 Resumen y discusión

En este capítulo se describió la técnica DPP, una propuesta diseñada para seleccionar términos en textos provenientes de redes sociales y hacer frente a los desafíos de este tipo de comunicación. Varias técnicas de selección de términos han sido empleadas en AP, desde un umbral de frecuencia hasta estrategias más sofisticadas como ganancia de información, información mutua y χ^2. Sin embargo, de acuerdo con Debole y Sebastiani (2004), la mayoría identifica los mejores términos como aquellos distribuidos de forma diferente en el conjunto de ejemplos positivos y negativos de la categoría. Sin embargo, cuando se tratan textos de redes sociales, existen grandes desafíos que limitan el alcance de los métodos basados en el uso de la frecuencia y/o distribución.

Por su parte, el DPP considera la calidad del contexto de ocurrencia de los términos más que su frecuencia. En dicha estrategia, la selección asume que cada término tiene un valor descriptivo y discriminativo. Específicamente, el factor descriptivo de un término depende directamente del tipo de información



(a) Blogs: caso edad

(b) Twitter: caso género

(c) Reviews: caso género

Figure 6.2: Las 100 palabras mejor calificadas por DPP.

expresada por el autor. Concretamente, depende de la relación entre el valor PEI y NEI (medidas propuestas en esta investigación). Por otro lado, el factor discriminativo hace uso de una implementación del coeficiente Gini, el cual ha sido ampliamente utilizado por varios métodos de partición en árboles de decisión (por ejemplo, CART y SLIQ), selección de términos (Shankar y Karypis, 2000; Shang et al., 2007; Singh et al., 2010) e incluso en algunas estrategias de pesado de términos (Zhu y Lin, 2013). Sin embargo, en esta tesis, su uso es requerido como un valor numérico que cuantifica la desigualdad de la distribución de los documentos según las categorías.

Los resultados señalan que las técnicas tradicionales de selección, omiten términos valiosos para la representación y que forman parte del contexto personal del autor. Además, se obtuvieron mejoras importantes con respecto a esquemas tradicionales en la predicción de la edad de los autores. Los resultados sugieren que existen términos de índole personal que distinguen diferentes grupos de edades. Sin embargo, para el caso género, el vocabulario de términos de los hombres y de las mujeres es muy similar.

Capítulo 7

EXPEI: un nuevo esquema de pesado de términos

En este capítulo se describe el esquema de pesado de términos propuesto en esta investigación, el cual está motivado en el rol especial que ejercen las frases personales en AP (ver Capítulo 4). Primero, se presenta la definición formal del esquema propuesto, el cual fue denominado EXPEI y cuya idea principal es premiar los términos con alta probabilidad de revelar los rasgos del perfil de los autores. Posteriormente, se muestran los experimentos que permiten evaluar su pertinencia en AP mediante el uso de distintas colecciones de documentos provenientes de distintas redes sociales.

7.1 Definición del esquema propuesto EXPEI

En un clasificador de textos, generalmente, la fase de pesado de términos toma aquellos términos seleccionados, T', y calcula un peso $w_{i,j}$ para cada documento d_j de la colección. El peso representa cómo el término t_i describe al documento d_j. En este trabajo de investigación, se propone un novedoso esquema de pesado de términos denominado *recompensa exponencial de información personal* denotado por las siglas EXPEI (*exponential rewarding of personal information*). Como su nombre sugiere, este esquema asigna una recompensa exponencial al peso de los términos que ocurren en frases personales. De esta manera, el esquema EXPEI considera toda la información de los documentos (personal y no personal), pero enfatiza la información personal.

De forma esencial, el diseño del esquema EXPEI está basado en las siguientes ideas:

- Asignar un peso a cada término del documento independientemente de su ocurrencia en las frases personales. Este peso inicial debe estar relacionado con la frecuencia total del término en el documento.
- Incrementar el peso de los términos en proporción a su ocurrencia en el subconjunto de frases personales.
- Dar igual importancia a documentos largos y cortos. Por consiguiente, los pesos deben ser normalizados: $0 \leq w_{i,j} \leq 1$.

La expresión 7.1 representa el esquema propuesto EXPEI:

$$w_{i,j} = \left(\sqrt{TF(t_i, d_j)}\right)^{1-PEI(t_i,d_j)}, \tag{7.1}$$

donde $TF(t_i, d_j)$ representa la frecuencia normalizada del término t_i en el documento d_j calculada como $\frac{\#(t_i,d_j)}{len(d_j)}$. Aquí, la raíz cuadrada de TF es usada para incrementar cada valor permitiendo percibir los cambios causados por el exponente PEI (recompensa o premio), especialmente, cuando TF tiene un valor muy pequeño. Concretamente, el premio está basado en la medida PEI ocasionando que el peso de un término incremente de acuerdo con su asociación a la información personal.

La Figura 7.1 ilustra el comportamiento del esquema EXPEI para diferentes valores de PEI (premios). EXPEI está basado en los valores de TF asignados a los términos. Específicamente, los términos con PEI = 1, obtendrán pesos iguales a 1 (EXPEI = 1, su máximo valor posible), independientemente de su frecuencia. Por otro lado, los términos con $0 <$ PEI < 1 serán proporcionalmente premiados; este premio es más importante para aquellos términos con baja frecuencia. Finalmente, los pesos de los términos con PEI = 0 serán suavizados por EXPEI (haciéndolos un poco más grandes que sus valores de TF). De esta manera, se permite que los términos con baja frecuencia tengan la oportunidad de contribuir en la descripción del documento.

Considerando lo anterior, EXPEI asigna valores altos a términos no frecuentes pero que están muy asociados al contexto personal de los autores. Incluso, los términos muy frecuentes también son relevantes para EXPEI. Por tal motivo,

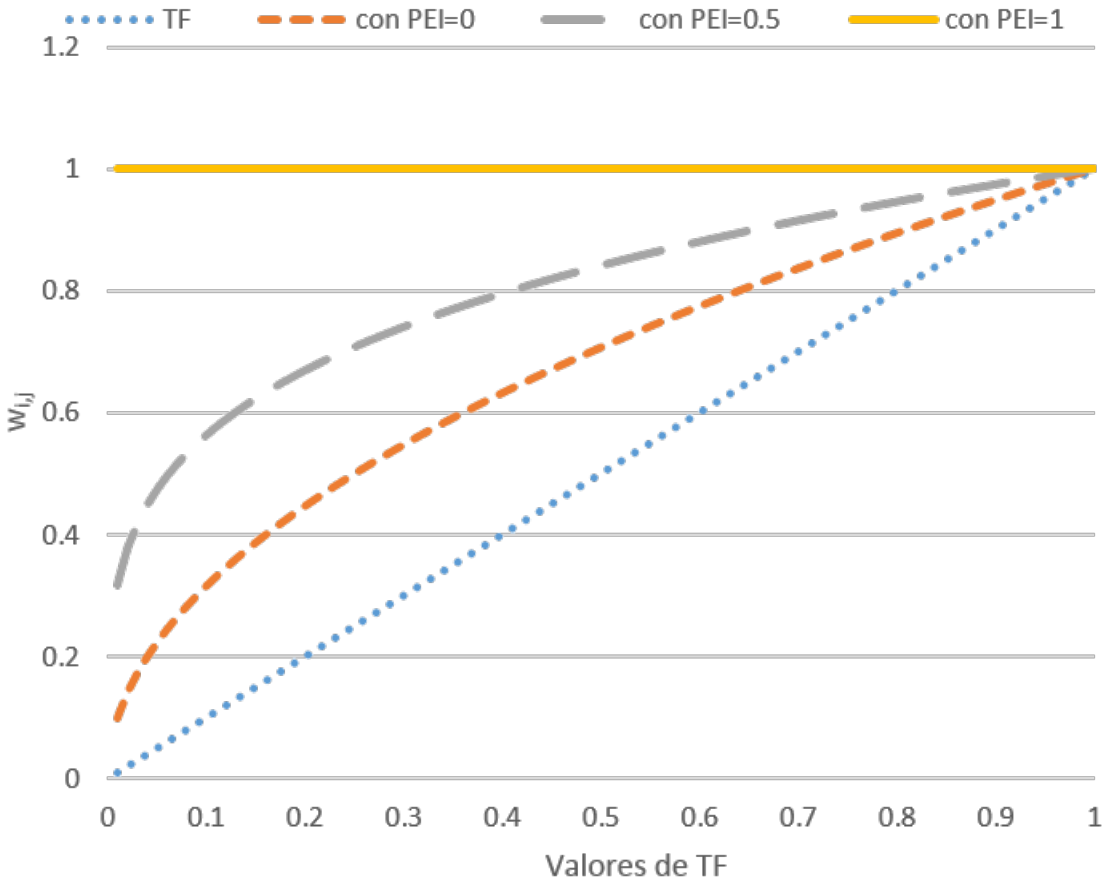


Figure 7.1: Representación del esquema EXPEI

aun cuando los términos estén relacionados a los contextos personales, no son sancionados pero tampoco premiados. De esta manera, EXPEI, sesga los pesos a los valores de PEI. Por lo tanto, EXPEI enriquece la frecuencia con información cualitativa ocasionando un deslinde parcial de la misma.

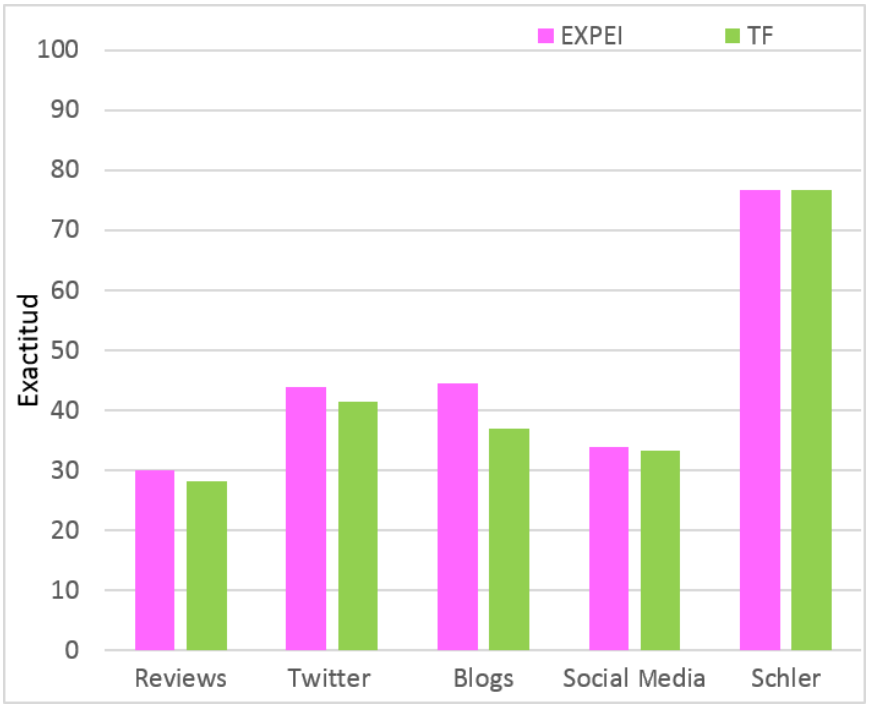
7.2 Experimento: evaluación de EXPEI

Este experimento está dirigido a conocer la contribución del esquema de pesado propuesto EXPEI en AP. Particularmente, en este experimento se

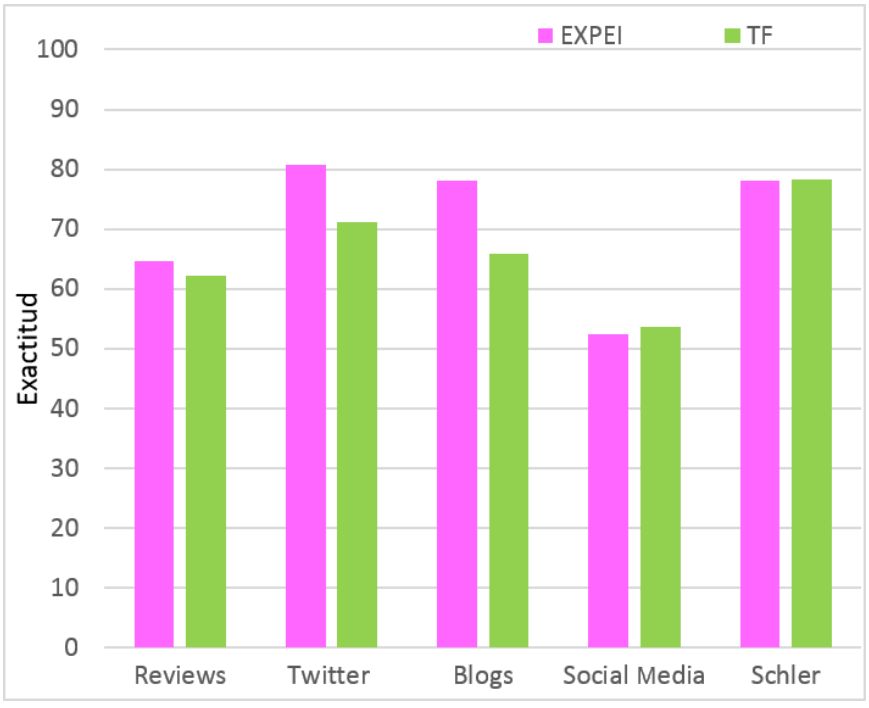
evita el uso de alguna técnica de selección de términos y se considera una representación formada por los 10,000 términos más frecuentes en el vocabulario. Para efectos de comparación, el valor de cada término en la representación es calculado por los esquemas EXPEI y TF.

Los resultados de la evaluación se muestran en la Figura 7.2, donde se confirma la relevancia del esquema de pesado propuesto. Por su parte, la prueba del rango con signo de Wilcoxon usando un nivel de significancia de 0.05, confirma mejoras significativas usando el esquema EXPEI con respecto a TF en ambas tareas de clasificación: edad y género; aunque las diferencias más notorias corresponden al caso género y en las colecciones de documentos más pequeñas. Es importante mencionar que para evaluar la significancia estadística se utilizaron los resultados de las 10 capas de la validación cruzada correspondiente a cada colección de documentos, (es decir, 50 conjuntos de datos). Los resultados sugieren que hombres y mujeres tienden a usar vocabularios similares. Sin embargo, las diferencias estriban en los términos específicos usados para describir sus intereses y asuntos personales (información personal), los cuales son mejor capturados por el esquema EXPEI que por TF.

Adicionalmente, se observan mayores ganancias en colecciones con pocos documentos y cuya longitud es corta. A partir de esta observación, se analizó la influencia de la longitud de los documentos en los resultados. Para ello, en la Figura 7.3 se graficaron correlaciones de los esquemas EXPEI y TF pesando los 10,000 términos más frecuentes de documentos que pertenecen a dos tipos



(a) Edad



(b) Género

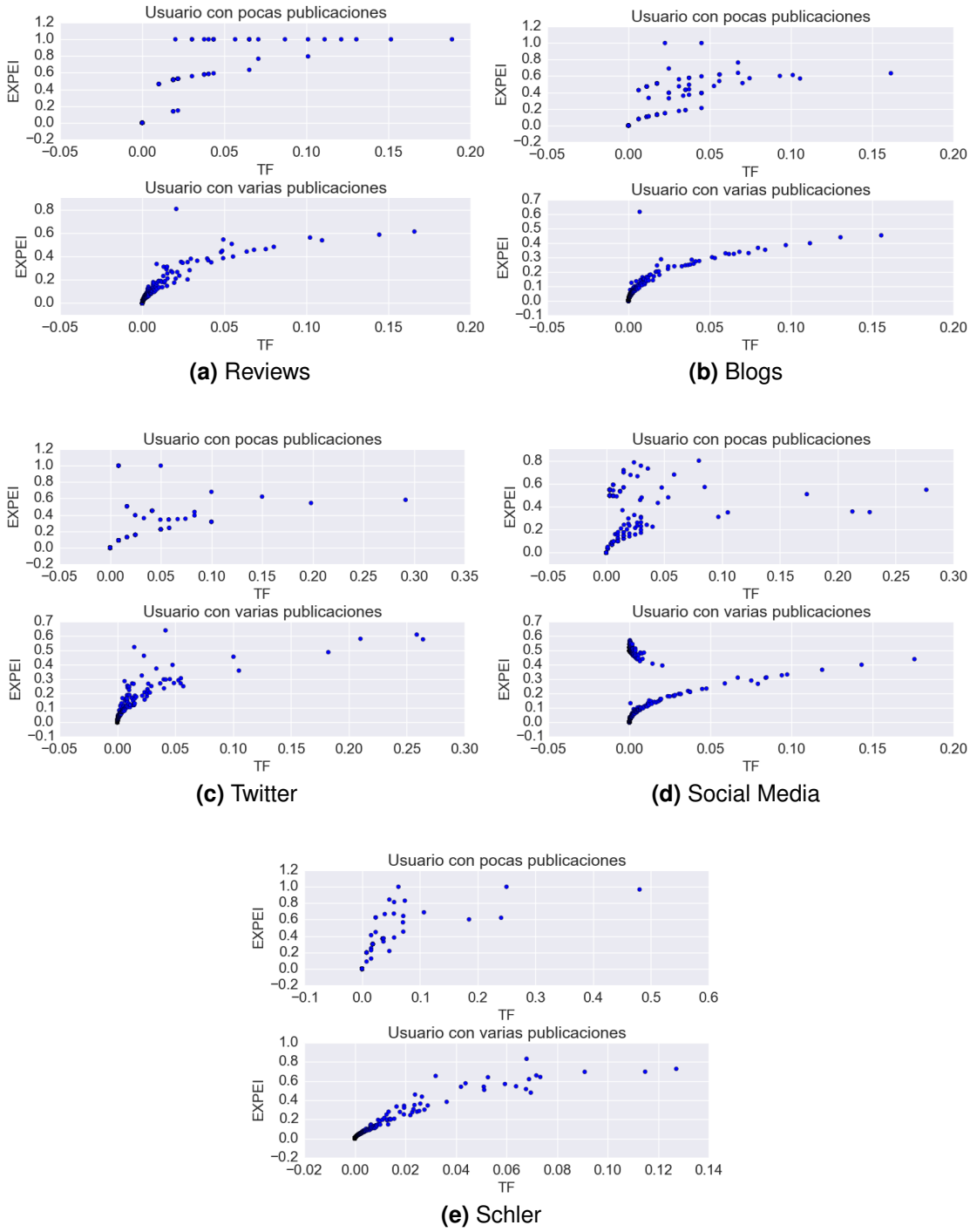
Figure 7.2: Comparación de EXPEI contra TF

de usuarios: i) con muy pocas publicaciones y ii) con un gran número (varias) publicaciones. En la figura, la inclinación formada por la concentración de los pesos indica la fuerza de la correlación entre ambos esquemas. Una correlación muy alta haría que los puntos se concentrarán en una línea inclinada. Es notorio que las representaciones están menos correlacionadas en los documentos con pocas publicaciones, esto indica que EXPEI extrae información relevante incluso cuando es poco frecuente, lo cual se traduce en una ventaja importante cuando hay poca información. Por otro lado, cuando existen más publicaciones, las correlaciones tienden a incrementar, sugiriendo que TF así como EXPEI capturan información personal que también es frecuente.

7.3 Resumen y discusión

La mayoría de los trabajos en el estado del arte han usado esquemas de pesado provenientes de recuperación de información como son: TF, TF-IDF y booleano entre otros. En este capítulo se propuso un novedoso esquema de pesado de términos, el cual ha sido referenciado como EXPEI.

El esquema EXPEI asigna los pesos de cada término en los documentos usando su frecuencia, pero enriqueciéndola mediante un exponente relacionado directamente con la medida PEI, la cual fue descrita en el Capítulo 5. De esta manera, cuando un término tiene un valor de PEI máximo (el cual corresponde a la unidad), el peso del término tendrá un valor igual a uno, independientemente del valor de la frecuencia. Por lo tanto, el beneficio más impactante se observa en los valores pequeños de TF. En contraste, si el



(a) Reviews **(b)** Blogs

(c) Twitter **(d)** Social Media

(e) Schler

Figure 7.3: Correlaciones de los esquemas EXPEI y TF

valor de PEI es igual a cero, EXPEI suaviza el valor generando pesos un poco mayores que TF, lo cual indica que el término no está asociado al contexto personal y que su frecuencia determinará su capacidad descriptiva.

Los experimentos mostraron la utilidad del esquema EXPEI, obteniendo mejoras en desempeño con respecto a esquemas tradicionales de pesado de términos. Por lo tanto, los resultados sugieren que los esquemas tradicionales, los cuales se basan en la frecuencia y/o distribución de los datos, son insuficientes para abordar el problema del perfilado de autores en redes sociales. Sobre todo, porque en esos contextos tan informales, es importante ponderar los fragmentos de texto donde aparece un término. De ahí que, el enfoque propuesto, combina información de la frecuencia del término con calificaciones del tipo de contexto donde aparece el término. De tal forma, que la frecuencia se altera produciendo pesos parcialmente deslindados de ella.

Capítulo 8

Enfoque integral

En este capítulo se plantea la contribución general de este trabajo de investigación. Específicamente, se define un enfoque integral que selecciona y pesa términos explotando los beneficios de la información personal para generar representaciones de los documentos. Por lo tanto, el enfoque integral se compone de las estrategias propuestas en los capítulos anteriores. En la primera sección del capítulo, se presentan las características del enfoque. Posteriormente, se describe la metodología experimental. Finalmente, se presentan los experimentos, resultados y comparaciones de desempeño con métodos del estado del arte.

8.1 Características del enfoque

La construcción de un clasificador de perfiles de autores involucra la transformación de documentos en una representación adecuada para los algoritmos de aprendizaje automático involucrando dos etapas principales: i) la selección de términos y ii) el pesado de términos. Teniendo en cuenta tales etapas, el enfoque integral propuesto se compone de la estrategia de selección de términos DPP (Capítulo 6), así como el esquema de pesado EXPEI (Capítulo 7). De ahí que el enfoque integral sea denominado DPP-EXPEI.

8.2 Metodología experimental

Para evaluar el enfoque integral, se diseñó una serie de experimentos cuya configuración general se muestra enseguida:

- Términos. Se consideró una combinación de atributos de contenido, estilo y sintácticos. Específicamente, como atributos de contenido se eligieron los 1000 términos con mayor pureza personal discriminativa[1] (aportación de esta tesis en el Capítulo 6). Estos términos incluyen palabras de contenido, signos de puntuación, vocabulario usual en redes sociales y emoticones. Por otro lado, como atributos de estilo se usaron palabras de función. Finalmente, se usaron etiquetas POS como elementos sintácticos.

[1] La técnica de selección DPP fue aplicada sobre los 10,000 términos más frecuentes de la colección. Se asume que términos escasamente frecuentes en la colección no serán relevantes para la clasificación.

- Pesado de términos. Cada término seleccionado fue pesado con el esquema EXPEI (aportación de esta tesis en el Capítulo 5).
- Clasificadores. Se consideró SVM como algoritmo principal de clasificación. Adicionalmente, en el segundo experimento, otros algoritmos fueron considerados: KNN y MNB. Todos ellos con las configuraciones por defecto. La evaluación se realizó mediante validación cruzada de 10 capas reportando la exactitud como principal medida de evaluación.
- Análisis de significancia. Finalmente, se aplicaron pruebas de significancia estadística (Demšar, 2006; García et al., 2010; Rodríguez-Fdez et al., 2015): la prueba del rango con signo de Wilcoxon para dos clasificadores[2] y la prueba *Friedman's Aligned Rank* para múltiples clasificadores[3].

8.3 Experimentos

Los experimentos correspondientes a la evaluación de DPP-EXPEI, son presentados como sigue: i) un análisis del enfoque prediciendo edad y género de autores; ii) una prueba sobre la robustez del enfoque ante diferentes clasificadores y iii) un estudio de la correlación existente entre los resultados y las características de las colecciones de documentos.

[2]Las pruebas de significancia estadística fueron calculadas mediante el paquete scmamp en R.

[3]El método Holm fue usado para correcciones de múltiples pruebas.

8.3.1 Experimento 1: desempeño del enfoque

El propósito de este experimento es proporcionar una perspectiva general sobre el desempeño del enfoque DPP-EXPEI. Los resultados fueron comparados con enfoques del estado del arte presentados en los siguientes trabajos, los cuales han usado las mismas colecciones de documentos (ver Sección 3.5):

- El trabajo de Álvarez-Carmona et al. (2016), donde se explotan representaciones basadas en tópicos usando análisis semántico latente **(LSA)** y conteos de palabras **(LIWC)**.
- El enfoque de López-Monroy et al. (2015), donde se describe una representación basada en encontrar subperfiles de usuario (**SSR**) modelando así, la heterogeneidad de los documentos que pertenecen al mismo perfil. En las colecciones del PAN 2013-2016, ha presentado los mejores resultados de ahí, en esta tesis fue utilizado como el método de referencia principal.
- El método de López-Monroy et al. (2014), donde se exploran atributos de segundo orden **(SOA)**.
- El trabajo de Weren et al. (2014), donde se presenta un método basado en ideas de recuperación de información **(IRF)**.
- El método de Booker (2008), donde se usan atributos a nivel de grupo **(GLA)** mediante un análisis de tópicos.
- El trabajo de Argamon et al. (2007), donde se presenta un análisis de más de 140 millones de palabras en inglés **(MW)** obtenidas de blogs.

- El enfoque de Schler et al. (2006), donde un conjunto de características estilísticas así como de contenido **(SC)** es usado para encontrar diferencias de género y edad.

Los resultados y comparaciones se muestran en la Tabla 8.1. Es posible notar un mejor desempeño en la identificación de género que en edad. Ese comportamiento es común dada la baja complejidad del problema de clasificación de género, ya que se tratan únicamente dos categorías (hombre, mujer) con un número balanceado de documentos etiquetados. Asimismo, es importante observar que el enfoque DPP-EXPEI supera a los enfoques de referencia en cada colección del PAN 2014 (únicamente en Social Media, SSR reportó una mejor exactitud). Además, el enfoque obtuvo mejores resultados que SSR – los mejores resultados reportados en el estado del arte – en tres colecciones en el caso edad. Mientras, en el caso género, se mejoran los resultados en las cuatro colecciones.

Es necesario señalar que se obtuvieron ganancias importantes, por ejemplo, en el corpus Blogs para el caso edad existe una diferencia aproximada de 22%; mientras las pérdidas obtenidas fueron muy pequeñas, la desventaja más representativa corresponde a 4.15 % en el caso edad de la colección Social Media. Por otro lado, la prueba del rango con signo de Wilcoxon fue aplicada para comparar estadísticamente SSR y DPP-EXPEI sobre los diez conjuntos de datos (uniendo edad y género). Con un nivel 0.05 de significancia, la prueba indica que DPP-EXPEI es significativamente mejor que SSR.

Tabla 8.1: Resultados usando los esquemas propuestos: pureza personal discriminativa para seleccionar términos y EXPEI para pesarlos

		Colecciones PAN 2014				Schler
	Enfoque	Reviews	Twitter	Blogs	Social Media	corpus
Edad	DPP-EXPEI	**44.83**	**61.44**	**75.34**	33.91	75.9
	LSA	34	39	48	36	-
	LIWC	29	47	42	34	-
	SSR	36.9	49.01	53.06	**38.06**	**77.68**
	SOA	33.92	47.97	48.07	37	-
	IRF	37.62	52.61	45.58	42.51	-
	GLA	-	-	-	-	72.83
	MW	-	-	-	-	77.4
	SC	-	-	-	-	76.01
Género	DPP-EXPEI	**76.42**	**81.5**	**84.25**	**58.57**	79.43
	LSA	65	66	70	52	-
	LIWC	62	71	60	50	-
	SSR	69.27	71.69	80.95	55.39	**82.01**
	SOA	68.05	71.92	77.96	55.36	-
	IRF	71.03	78.76	82.99	57.04	-
	GLA	-	-	-	-	75.04
	MW	-	-	-	-	80.5
	SC	-	-	-	-	80.01

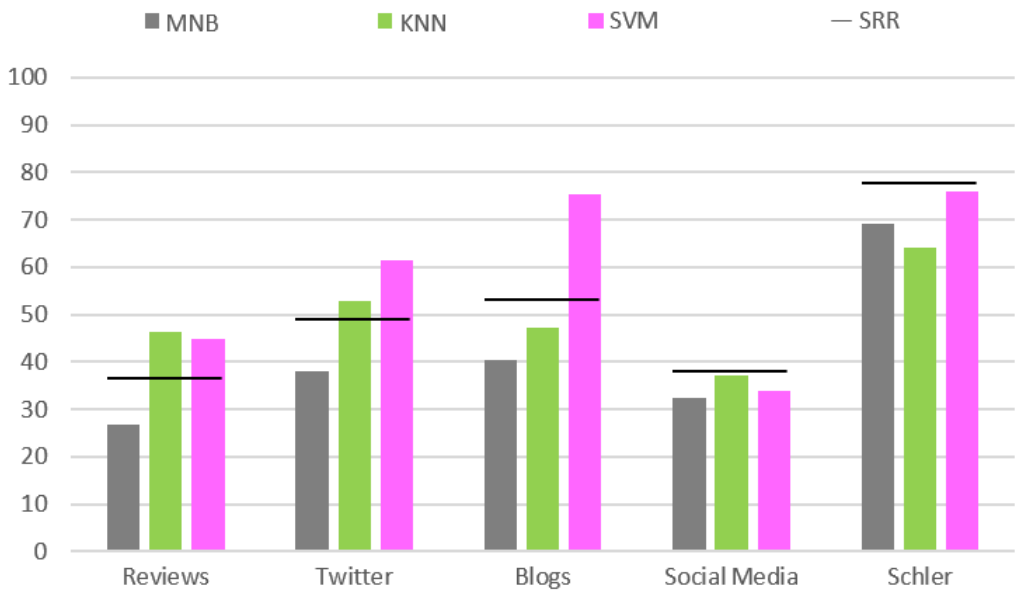
A partir de esos resultados, es posible concluir que DPP-EXPEI enfoca la atención en los términos más relevantes (intereses personales) cuando se tiene menos información (i.e., colecciones pequeñas). Sin embargo, cuando hay más información textual disponible (colecciones más grandes) DPP-EXPEI tiene menos impacto, principalmente, porque los términos frecuentes tienden a exponer directamente tales intereses personales.

8.3.2 Experimento 2: robustez ante diferentes algoritmos de clasificación

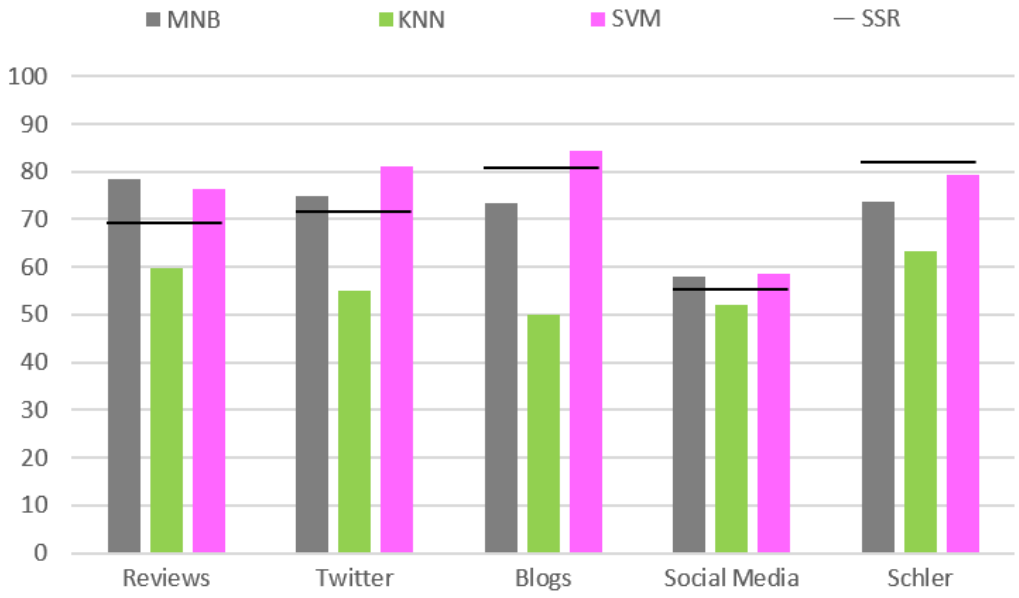
El experimento previo mostró mejoras cuando DPP y EXPEI son explotados usando SVM como algoritmo de clasificación. Por su parte, este experimento analiza la robustez del enfoque propuesto cuando otros clasificadores son usados. Los resultados se presentan en la Figura 8.1, donde se exhibe una comparación de diferentes clasificadores con parámetros por defecto (*default*): MNB, KNN (K=1) y SVM. También se muestra el enfoque SSR para efectos de comparación.

En la Figura 8.1 se observa que la propuesta definitivamente está tomando ventaja de SVM para la predicción de edad y género. Sin embargo, es interesante ver las mejoras obtenidas por KNN para el caso edad en las colecciones Reviews y Twitter. Por lo contrario, en el caso género, Naïve Bayes se ha desempeñado mejor que KNN y sobrepasa el desempeño de SSR en tres colecciones (Reviews, Twitter y Social Media).

Para validar los resultados, se evaluó la significancia estadística con la prueba *Friedman Aligned Rank* usando un nivel de significancia correspondiente a 0.05. La prueba fue aplicada para comparar la exactitud de los tres algoritmos de clasificación sobre las cinco colecciones de documentos. Para el caso edad, la prueba indica que SVM es significativamente mejor que Naïve Bayes, pero no hay una diferencia significativa con KNN. Por el otro lado, para el caso



(a) Edad



(b) Género

Figure 8.1: Comparación de diferentes algoritmos de clasificación

género, el desempeño de SVM es significativamente mejor que KNN, pero no hay diferencia significativa con Bayes.

En general, los resultados indican que el clasificador es importante para explotar correctamente la propuesta. Aunque se obtienen resultados aceptables con diferentes algoritmos, la mejor combinación es DPP-EXPEI con un clasificador basado en SVM.

8.3.3 Experimento 3: el rol de las características de la colección

El propósito de este experimento es analizar el rol o influencia de diferentes características de las colecciones en el desempeño del enfoque propuesto. En particular, se analizó la correlación entre propiedades específicas de las colecciones y la mejora de exactitud del método sobre el enfoque de referencia SSR (López-Monroy et al., 2015). Para ello, se aplicó el coeficiente de correlación de Spearman (Spearman, 1987). Las propiedades de las colecciones analizadas son las siguientes:

- **Número de ejemplos:** Representa el número total de documentos en la colección de textos.
- **Proporción de tipos de tokens** (*Type Token Ratio*, TTR, por sus siglas en inglés): Es una medida de la riqueza del vocabulario, la cual corresponde al promedio de las diferentes unidades léxicas (tokens) existentes en un documento (Laufer y Nation, 1995).

- **Densidad léxica:** Es calculada como el promedio de unidades léxicas (no incluyen palabras de función) sobre el total del número de unidades en el texto. También es una medida de riqueza del vocabulario (Laufer y Nation, 1995).

- **Desbalance de clases:** Es una medida supervisada estimada como la desviación estándar de las diferencias entre el número de documentos real y el ideal para cada categoría. El número de documentos ideal por categoría es definido como el cociente del número de documentos en la colección y el número de categorías. Entre más alto es este valor, más desbalanceada es la colección (Rosso et al., 2009)[4].

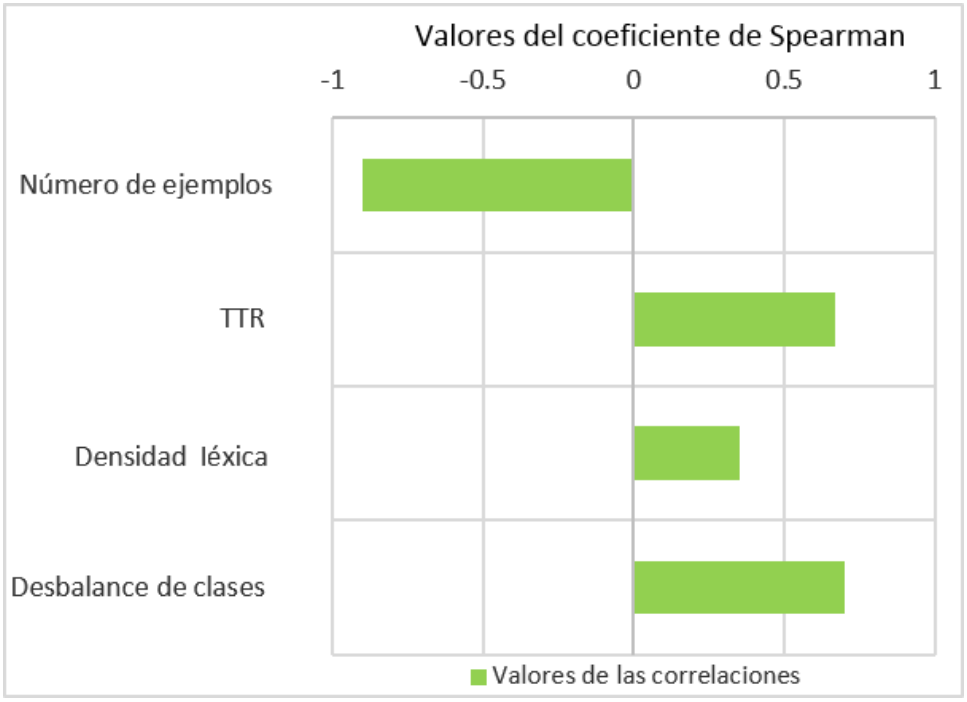
La Figura 8.2 muestra los valores de las correlaciones. A partir de la figura, se pueden señalar aspectos interesantes acerca del enfoque:

- El enfoque propuesto se desempeña mejor que el método de referencia cuando el número de ejemplos es pequeño. Esta propiedad es muy importante ya que en muchos escenarios del mundo real existe carencia de ejemplos para entrenamiento.

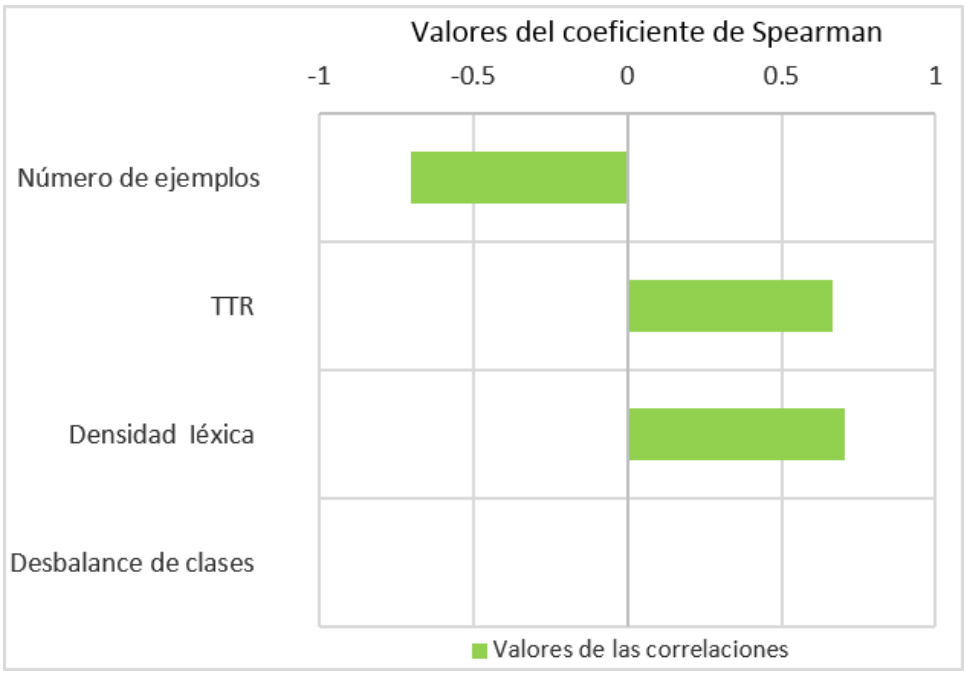
- Entre más alto es el valor TTR, mayor es la diferencia positiva con el método de referencia. Lo anterior sugiere, que el uso de un vocabulario diverso, característica común en las redes sociales, beneficia al enfoque propuesto.

- La densidad léxica también contribuye al desempeño del enfoque. Eso significa que, los documentos con un mayor número de sustantivos,

[4]Para la comparación, los valores son normalizados dividiendo por el número de documentos en la colección.



(a) Edad



(b) Género

Figure 8.2: Correlaciones del enfoque integral con características de las colecciones

verbos y adjetivos tienden a favorecer el desempeño del enfoque, principalmente, porque expresan información personal (intereses, preferencias, hábitos, etc.) de los usuarios. De esta manera, el enfoque se está beneficiando de la variabilidad del lenguaje existente en las redes sociales.

- Cuando la colección es altamente desbalanceada[5] el enfoque propuesto tiende a mejorar los resultados del método de referencia. Esta correlación es muy importante porque el desbalance de clases es un gran desafío que interesa a la comunidad científica, sobre todo en dominios de redes sociales.

En resumen, el enfoque es apropiado para la tarea de identificación del perfil de autores cuando existan pocos ejemplos con alta densidad léxica y colecciones de entrenamiento desbalanceadas, condiciones desafiantes en AP. Por el contrario, se recomienda evitar el uso del enfoque propuesto cuando exista un gran número de ejemplos.

8.4 Resumen y discusión

En este capítulo se presentó un novedoso enfoque para AP que integra los esquemas de selección y pesado de términos propuestos en los Capítulos 6 y 7, respectivamente. De esta manera, se conjugan las ventajas de ambas técnicas para lograr un enfoque integral que mejora el desempeño de la predicción de rasgos del perfil de autores en redes sociales. Concretamente,

[5]Para el caso género, el desbalance tiene un valor 0 porque los documentos están balanceados.

a través de ambas técnicas, el enfoque toma como base la evaluación del tipo de información que revela cada término. Es importante notar que, como se muestra en la Tabla 8.1, el esquema de pesado EXPEI es crucial para mejorar el desempeño del esquema DPP.

A partir de los resultados, es posible concluir que el enfoque integral DPP-EXPEI permite enfocar la atención en los términos más relevantes (intereses personales) cuando se tiene menos información (i.e., colecciones pequeñas). Sin embargo, cuando hay más información textual disponible (colecciones más grandes) DPP-EXPEI tiene menos impacto, principalmente, porque los términos frecuentes tienden a exponer directamente tales intereses personales.

También, se observó que los algoritmos de clasificación impactan en el desempeño del enfoque. Similar a los enfoques actuales, se notó una ventaja significativa de las SVM sobre otros algoritmos de clasificación.

Asimismo, se observó que el enfoque propuesto genera ganancias de exactitud muy importantes con respecto a otros métodos del estado del arte que usan las mismas colecciones de documentos. Particularmente, se notó una mejora de hasta 22% sobre el trabajo de López-Monroy et al. (2015), cuyo enfoque ha tenido éxito en la tarea e incluso ha sido usado en nuevas propuestas sometidas en los foros del PAN.

Finalmente, se estudió la influencia de las propiedades de las colecciones de documentos en los resultados. Se determinó que el enfoque integral

obtiene mayores ventajas con respecto al estado del arte cuando existen: pocos ejemplos de entrenamiento, alto desbalance de clases y valores altos de densidad léxica.

Conclusiones y trabajo futuro

En este trabajo de investigación se aborda la tarea conocida como *Identificación del perfil de autores*, la cual consiste en analizar los textos para predecir las características socio-demográficas de los autores. Generalmente, se ha tratado desde la perspectiva de categorización supervisada de textos. Las contribuciones se han dirigido a encontrar un conjunto adecuado de atributos que capture las diferencias y similaridades de grupos de autores que comparten rasgos comunes como edad y género. Diferente de la tendencia general, en este trabajo de investigación se propuso un nuevo enfoque para clasificar perfiles de autores en redes sociales cuya esencia indica que la información contenida en los textos no es igualmente relevante para la tarea.

La idea base de la investigación expresa que las frases con pronombres en primera persona del singular (frases personales) exponen información valiosa que revela el perfil de los autores. En este capítulo se presentan las conclusiones generadas, contribuciones específicas y direcciones futuras.

Conclusiones

Mediante el enfoque propuesto, las hipótesis planteadas en esta tesis fueron confirmadas. Asimismo, se concluyen dos aspectos primordiales:

- La información de tipo personal es más valiosa que el resto de la información encontrada en un texto para identificar el perfil de autores en redes sociales.
- Enfatizar el valor de los términos contenidos en las frases personales considerando la información en frases no personales enriquecen los resultados de la clasificación de perfiles.

De forma específica, el rol de las frases personales en la identificación del perfil del autor fue estudiado. Se descubrió que las frases personales contenidas en los documentos representan la esencia de los mismos (la parte más importante) para la tarea.

Con base en dicha esencia, se estableció un criterio de construcción y etiquetado de corpus para AP, una tarea altamente costosa en términos de esfuerzo y tiempo. El criterio facilita la construcción, pues propone evaluar la calidad de los documentos según su esencia con el objetivo de filtrar o limitar el número de documentos etiquetados.

Adicionalmente, para tomar ventaja del rol de las frases personales, se diseñaron novedosas estrategias que enfatizan el valor de los términos que existen en las frases personales: un método de selección de términos

denominado pureza personal discriminativa (DPP) y un esquema de pesado de términos (EXPEI). Las estrategias diseñadas se basan en una medida llamada índice de expresión personal (PEI), la cual fue propuesta para cuantificar el grado de asociación de los términos a la información personal del autor.

La evaluación experimental permitió validar las representaciones generadas por los esquemas de selección y pesado propuestos para identificar perfiles de autores en textos provenientes de medios sociales. La comparación con enfoques actuales del estado del arte mostró una fuerte evidencia acerca de la usabilidad de los esquemas propuestos. En este contexto, se obtuvieron las siguientes conclusiones:

- DPP y EXPEI son estrategias que evalúan las características cualitativas de las ocurrencias de los términos en los documentos. Actualmente, los métodos de selección y pesado de términos usados, que en su mayoría provienen de la recuperación de información, están soportados en inferencias estadísticas acerca de las ocurrencias de los términos en los documentos, sin considerar el tipo de contexto de los términos.
- Las estrategias propuestas de selección y pesado aplicadas de forma individual, generan un desempeño competitivo con respecto a métodos del estado del arte. Sin embargo, cuando ambas técnicas son consideradas en forma conjunta, se obtienen ganancias significativas con respecto al enfoque que ha reportado el desempeño más alto en las mismas colecciones de documentos, en promedio 7.34% en la identificación de edad y 5.76% en género.

- Para la predicción de género, la etapa de pesado de términos fue más relevante. Este comportamiento indica que para predecir género, los pesos de los términos son muy importantes.
- Para la predicción de edad, la fase de selección de términos resultó muy importante. Por lo tanto, aquí la presencia o ausencia (ver el desempeño del pesado booleano en el experimento 6.3) de tópicos específicos son suficientes para determinar el perfil apropiado.
- La pertinencia del método fue corroborada para trabajar en dominios de redes sociales, en donde es muy común ver datos desbalanceados, variabilidad del lenguaje y alta riqueza del vocabulario.

Contribuciones específicas

La presente investigación aporta las siguientes contribuciones, las cuales están asociadas a los objetivos:

- Un estudio sobre la relevancia de las frases personales en la identificación del perfil de autores, sustentado en estudios psicológicos que relacionan el uso de pronombres personales con las características de las personas.
- Un criterio para construcción de corpus, el cual filtra los documentos a etiquetarse de acuerdo con el tipo de frases existentes en su contenido.
- Una medida (PEI) para estimar la asociación de los términos con la información personal.

- Un esquema de pesado de términos (EXPEI) para seleccionar los términos más valiosos de acuerdo con su información personal.
- Un esquema de selección de términos (DPP) que distingue los valores de los términos ocurriendo en frases personales y no personales. A nuestro conocimiento, esta tesis es el primer trabajo que propone un nuevo método de selección de términos para AP que está alejado del uso común de la frecuencia de los términos.
- Un estudio comparativo del enfoque integral (seleccionando términos con DPP y pesándolos con EXPEI): i) usando diferentes algoritmos de aprendizaje, ii) empleando distintos dominios de redes sociales y iii) contrastando con principales métodos del estado del arte obteniendo mejoras en la mayoría de los resultados.

En suma, se contribuye con un novedoso enfoque integral para la identificación del perfil de autores en redes sociales cuya selección y pesado de términos se sustenta en la relevancia de las frases personales.

Trabajo futuro

Los resultados han motivado el interés de abordar, como trabajo futuro, los siguientes aspectos:

- Evaluar el enfoque propuesto prediciendo otras dimensiones del perfil de autores. En los experimentos mostrados en esta tesis se abordó la predicción de edad y género. Sin embargo, se planea probar el enfoque propuesto clasificando otros rasgos del perfil de los autores

como la personalidad (e.g., extroversión vs introversión). Se cree que la información expuesta en las frases personales revela información del perfil completo de los autores, por ello, características como la personalidad pueden ser también inferidas.

- Estudiar la pertinencia del enfoque propuesto usando textos escritos en otros idiomas. Los resultados mostraron que el método es adecuado tratando con el idioma inglés donde los pronombres deben escribirse. Sin embargo, existe un interés especial de evaluar el método en otros idiomas, particularmente, en aquellos donde los pronombres personales se pueden omitir (*pronoun-dropping languages*) como el idioma español. Inicialmente, y de acuerdo con el diseño del enfoque, se considera que el desempeño, más que verse perjudicado, se enriquecerá del énfasis de aquellas pocas o varias frases con pronombres personales explícitos (en primera persona).

- Probar el enfoque en colecciones carentes de pronombres personales. Ante la carencia de pronombres personales, el enfoque está diseñado para seleccionar características con baja asociación no personal y para pesarlas con la frecuencia normalizada. Sin embargo, es importante estudiar el desempeño general ante tal situación.

- Diseñar estrategias de pesado de instancias. Se planea dirigir el concepto de enfatizar información de tipo personal hacia el pesado o selección de instancias (documentos) en la identificación del perfil del autor, aspecto que a la fecha no ha sido estudiado. En este sentido, las estrategias

a diseñar podrían estimar la calidad de un documento considerando el grado de asociación personal expresado en su contenido.

- Aplicar el enfoque propuesto en la caracterización de tipos de comportamientos sociales de los usuarios. Por ejemplo, la identificación de acoso y la detección de usuarios con depresión.

Referencias

Aggarwal, C. (2015). Data Mining: The Textbook. Springer International Publishing.

Agrawal, M. y Gonçalves, T. (2016). Age and Gender Identification Using Stacking for Classification—Notebook for PAN at CLEF 2016. *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, Évora, Portugal*. Ed. by K. Balog, L. Cappellato, N. Ferro y C. Macdonald. CEUR-WS.org.

Álvarez-Carmona, M., López-Monroy, A., Montes-y-Gómez, M., Villaseñor-Pineda, L. y Escalante, H. (2015). INAOE's participation at PAN'15: Author Profiling task—Notebook for PAN at CLEF 2015. *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France*. Ed. by L. Cappellato, N. Ferro, G. Jones y E. San Juan. CEUR-WS.org.

Álvarez-Carmona, M. A., López-Monroy, A. P., Montes-y-Gómez, M., Villaseñor-Pineda, L. y Meza, I. (2016). Evaluating Topic-Based Representations for Author Profiling in Social Media. *Advances in Artificial*

Intelligence - IBERAMIA 2016: 15th Ibero-American Conference on AI, San José, Costa Rica, November 23-25. Ed. by M. Montes y Gómez, H. J. Escalante, A. Segura y J. d. D. Murillo. Cham: Springer International Publishing, pp. 151–162.

Anguiano-Hernández, E., Villaseñor-Pineda, L., Montes-y-Gómez, M. y Rosso, P. (2010). Summarization as Feature Selection for Document Categorization on Small Datasets. *Advances in Natural Language Processing: 7th International Conference on NLP, IceTAL 2010, Reykjavik, Iceland, August 16-18*. Ed. by H. Loftsson, E. Rögnvaldsson y S. Helgadóttir. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 39–44.

Argamon, S., Dhawle, S., Koppel, M. y Pennebaker, J. W. (2005). Lexical Predictors of Personality Type. *Proceedings of the Joint Annual Meeting of the Interface and the Classification Society of North America*.

Argamon, S., Koppel, M., Fine, J. y Shimoni, A. R. (2003). Gender, Genre, and Writing Style in Formal Written Texts. *Text - Interdisciplinary Journal for the Study of Discourse*, 23(3), pp. 321–346.

Argamon, S., Koppel, M., Pennebaker, J. W. y Schler, J. (2007). Mining the Blogosphere: Age, Gender and the Varieties of Self-expression. *First Monday*, 12(9).

Argamon, S., Koppel, M., Pennebaker, J. W. y Schler, J. (2009). Automatically Profiling the Author of an Anonymous Text. *Commun. ACM*, 52(2), pp. 119–123.

Argamon, S. y Levitan, S. (2005). Measuring the Usefulness of Function Words for Authorship Attribution. *Proceedings of the Joint International*

Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities.

Baeza-Yates, R. A. y Ribeiro-Neto, B. (1999). Modern Information Retrieval. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Baker, L. D. y McCallum, A. K. (1998). Distributional Clustering of Words for Text Classification. *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '98. Melbourne, Australia: ACM, pp. 96–103.

Bayot, R. K. y Gonçalves, T. (2016). Author Profiling Using SVMs and Word Embedding Averages—Notebook for PAN at CLEF 2016. *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, Évora, Portugal*. Ed. by K. Balog, L. Cappellato, N. Ferro y C. Macdonald. CEUR-WS.org.

Bekkerman, R., El-Yaniv, R., Tishby, N. y Winter, Y. (2003). Distributional Word Clusters vs. Words for Text Categorization. *Journal of Machine Learning Research*, 3, pp. 1183–1208.

Bergsma, S., Post, M. y Yarowsky, D. (2012). Stylometric Analysis of Scientific Articles. *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. NAACL HLT '12. Montreal, Canada: Association for Computational Linguistics, pp. 327–337.

Bhatia, N. y Vandana (2010). Survey of Nearest Neighbor Techniques. *International Journal of Computer Science and Information Security (IJCSIS)*, 8(2), pp. 302–305.

Bilan, I. y Zhekova, D. (2016). CAPS: A Cross-genre Author Profiling System. *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, Évora, Portugal*. Ed. by K. Balog, L. Cappellato, N. Ferro y C. Macdonald. CEUR-WS.org.

Blei, D. M., Ng, A. Y. y Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan), pp. 993–1022.

Booker, L. B. (2008). Finding Identity Group "Fingerprints" in Documents. *Computational Forensics: Second International Workshop, IWCF 2008, Washington, DC, USA, August 7-8, Proceedings*. Ed. by S. N. Srihari y K. Franke. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 113–121.

Bouanani, S. E. M. E. y Kassou, I. (2014). Authorship Analysis Studies: A Survey. *International Journal of Computer Applications*, 86(12), pp. 22–29.

Bykh, S. y Meurers, D. (2014). Exploring Syntactic Features for Native Language Identification: A Variationist Perspective on Feature Encoding and Ensemble Optimization. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 1962–1973.

Chang, C. y Lin, C. (2011). LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3), 27:1–27:27.

Cheng, N., Chandramouli, R. y Subbalakshmi, K. P. (2011). Author Gender Identification from Text. *Digital Investigation*, 8(1), pp. 78–88.

Chomboon, K., Chujai, P., Teerarassammee, P., Kerdprasop, K. y Kerdprasop, N. (2015). An Empirical Study of Distance Metrics for k-Nearest Neighbor

Algorithm. *The 3rd International Conference on Industrial Application Engineering 2015 (ICIAE2015)*, pp. 280–285.

Chung, C. y Pennebaker, J. W. (2007). The Psychological Function of Function Words. *Social Communication: Frontiers of Social Psychology*. Psychology Press, pp. 343–359.

Coulthard, M. y Johnson, A. (2007). An Introduction to Forensic Linguistics: Language in Evidence. Routledge.

Cover, T. y Hart, P. (2006). Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*, 13(1), pp. 21–27.

Danenas, P. y Garsva, G. (2011). SVM and XBRL based decision support system for credit risk evaluation. *Proceedings of the 17th International Conference on Information and Software Technologies (IT 2011), Technologija, Kaunas, Lithuania*, pp. 190–198.

Debole, F. y Sebastiani, F. (2004). Supervised Term Weighting for Automated Text Categorization. *Text Mining and its Applications: Results of the NEMIS Launch Conference*. Ed. by S. Sirmakessis. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 81–97.

Demšar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, 7, pp. 1–30.

Dietterich, T. G. (1998). Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, 10(7), pp. 1895–1923.

Ding, S. H. H., Fung, B. C. M., Iqbal, F. y Cheung, W. K. (2016). Learning Stylometric Representations for Authorship Analysis. *CoRR*, abs/1606.01219.

Dixon, P., Weiner, J., Mitchell-Olds, T. y Woodley, R. (1988). Erratum to Bootstrapping the Gini Coefficient of Inequality. *Ecology*, 68, p. 1307.

Duong, D. T., Pham, S. B. y Tan, H. (2016). Using Content-Based Features for Author Profiling of Vietnamese Forum Posts. *Recent Developments in Intelligent Information and Database Systems*. Ed. by D. Król, L. Madeyski y N. T. Nguyen. Cham: Springer International Publishing, pp. 287–296.

Enas, G. G. y Choi, S. C. (1986). Choice of the Smoothing Parameter and Efficiency of k-Nearest Neighbor Classification. *Computers & Mathematics with Applications*, 12(2), pp. 235–244.

Escalante, H. J., García-Limón, M. A., Morales-Reyes, A., Graff, M., Montes-y-Gómez, M., Morales, E. F. y Martínez-Carranza, J. (2015). Term-Weighting Learning via Genetic Programming for Text Classification. *Knowledge-Based Systems*, 83, pp. 176–189.

Escalante, H. J., Solorio, T. y Montes-y-Gómez, M. (2011). Local Histograms of Character N-grams for Authorship Attribution. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*. HLT '11. Portland, Oregon: Association for Computational Linguistics, pp. 288–298.

Estival, D., Gaustad, T., Hutchinson, B., Pham, S. B. y Radford, W. (2007a). Author Profiling for English Emails. *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pp. 263–272.

Estival, D., Gaustad, T., Pham, S. B., Radford, W. y Hutchinson, B. (2007b). TAT: An Author Profiling Tool with Application to Arabic Emails. *Proceedings of the Australasian Language Technology Workshop 2007*. Melbourne, Australia, pp. 21–30.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. y Lin, C.-J. (2008). LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, 9, pp. 1871–1874.

Fast, L. A. y Funder, D. C. (2008). Personality as Manifest in Word Use: Correlations With Self-Report,Acquaintance Report, and Behavior. *Journal of personality and social psychology*, 94(2), p. 334.

Fatima, M., Hasan, K., Anwar, S. y Nawab, R. M. A. (2017). Multilingual author profiling on Facebook. *Information Processing & Management*, 53(4), pp. 886–904.

Feldman, R. y Sanger, J. (2006). Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. New York, NY, USA: Cambridge University Press.

García, S., Fernández, A., Luengo, J. y Herrera, F. (2010). Advanced Nonparametric Tests for Multiple Comparisons in the Design of Experiments in Computational Intelligence and Data Mining: Experimental Analysis of Power. *Information Sciences*, 180(10). Special Issue on Intelligent Distributed Information Systems, pp. 2044–2064.

Gencheva, P., Boyanov, M., Deneva, E., Nakov, P., Kiprov, Y., Koychev, I. y Georgiev, G. (2016). PANcakes Team: A Composite System of Genre-Agnostic Features For Author Profiling—Notebook for PAN at CLEF 2016. *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, Évora, Portugal*. Ed. by K. Balog, L. Cappellato, N. Ferro y C. Macdonald. CEUR-WS.org.

Golbeck, J. (2016). Predicting Personality from Social Media Text. *AIS Transactions on Replication Research*, 2(1), p. 2.

Goswami, S., Sarkar, S. y Rustagi, M. (2009). Stylometric Analysis of Bloggers' Age and Gender. *Proceedings of the Third International ICWSM Conference*. Ed. by E. Adar, M. Hurst, T. Finin, N. S. Glance, N. Nicolov y B. L. Tseng. The AAAI Press, pp. 214–217.

Haddi, E., Liu, X. y Shi, Y. (2013). The Role of Text Pre-processing in Sentiment Analysis. *First International Conference on Information Technology and Quantitative Management*, 17, pp. 26–32.

Halteren, H. van (2004). Linguistic Profiling for Author Recognition and Verification. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. ACL '04. Barcelona, Spain: Association for Computational Linguistics.

Hand, D. J. y Vinciotti, V. (2003). Choosing k for Two-class Nearest Neighbour Classifiers with Unbalanced Classes. *Pattern Recognition Letters*, 24(9), pp. 1555–1562.

Houvardas, J. y Stamatatos, E. (2006). N-Gram Feature Selection for Authorship Identification. *Artificial Intelligence: Methodology, Systems, and Applications: 12th International Conference, AIMSA 2006, Varna, Bulgaria, September 12-15, 2006. Proceedings*. Ed. by J. Euzenat y J. Domingue. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 77–86.

Hu, L.-Y., Huang, M.-W., Ke, S.-W. y Tsai, C.-F. (2016). The Distance Function Effect on k-nearest Neighbor Classification for Medical Datasets. *SpringerPlus*, 5(1), p. 1304.

Joachims, T. (1999). Making large-Scale SVM Learning Practical. *Advances in Kernel Methods - Support Vector Learning*. Ed. by B. Schölkopf, C. Burges y A. Smola. Cambridge, MA: MIT Press. Chap. 11, pp. 169–184.

Joachims, T. (1998). Text Categorization with Suport Vector Machines: Learning with Many Relevant Features. *Proceedings of the 10th European Conference on Machine Learning*. ECML '98. London, UK, UK: Springer-Verlag, pp. 137–142.

Joachims, T. (2001). A Statistical Learning Learning Model of Text Classification for Support Vector Machines. *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '01. New Orleans, Louisiana, USA: ACM, pp. 128–136.

Joachims, T. (2002). Text Classification. *Learning to Classify Text Using Support Vector Machines*. Boston, MA: Springer US, pp. 7–33.

Kacewicz, E., Pennebaker, J. W., Davis, M., Jeon, M. y Graesser, A. C. (2014). Pronoun Use Reflects Standings in Social Hierarchies. *Journal of Language and Social Psychology*, 33(2), pp. 125–143.

Kanji, G. (2006). 100 Statistical Tests. 100 Statistical Tests. SAGE Publications.

Kocher, M. (2015). UniNE at CLEF 2015: Author Profiling. *CLEF (Working Notes)*, pp. 903–911.

Kocher, M. y Savoy, J. (2017). Distance Measures in Author Profiling. *Information Processing & Management*, 53(5), pp. 1103–1119.

Koppel, M., Argamon, S. y Shimoni, A. R. (2002). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4), pp. 401–412.

Koppel, M. y Schler, J. (2004). Authorship Verification As a One-class Classification Problem. *Proceedings of the Twenty-first International Conference on Machine Learning*. ICML '04. Banff, Alberta, Canada: ACM, p. 62.

Koppel, M., Schler, J. y Zigdon, K. (2005). Determining an Author's Native Language by Mining a Text for Errors. *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*. KDD '05. Chicago, Illinois, USA: ACM, pp. 624–628.

Koppel, M. y Winter, Y. (2014). Determining if two Documents Are by the Same Author. *Journal of the Association for Information Science and Technology*, 65(1), pp. 178–187.

Kuralenok, I. y Nekrest'yanov, I. (2000). Automatic Document Classification based on Latent Semantic Analysis. *Programming and Computer Software*, 26(4), pp. 199–206.

Lan, M., Tan, C.-L., Low, H.-B. y Sung, S.-Y. (2005). A Comprehensive Comparative Study on Term Weighting Schemes for Text Xategorization with Support Vector Machines. *Special interest tracks and posters of the 14th international conference on World Wide Web*. ACM, pp. 1032–1033.

Landauer, T. K., Foltz, P. W. y Laham, D. (1998). An Introduction to Latent Semantic Analysis. *Discourse Processes*, 25(2-3), pp. 259–284.

Laufer, B. y Nation, P. (1995). Vocabulary Size and Use: Lexical Richness in L2 Written Production. *Applied Linguistics*, 16(3), pp. 307–322.

Lebanon, G., Mao, Y. y Dillon, J. (2007). The Locally Weighted Bag of Words Framework for Document Representation. *Journal of Machine Learning Research*, 8(Oct), pp. 2405–2441.

Leopold, E. y Kindermann, J. (2002). Text Categorization with Support Vector Machines. How to Represent Texts in Input Space? *Machine Learning*, 46(1), pp. 423–444.

Lewis, D. D. (1990). Representation Quality in Text Classification: An Introduction and Experiment. *Proceedings of the Workshop on Speech and Natural Language*. HLT '90. Hidden Valley, Pennsylvania: Association for Computational Linguistics, pp. 288–295.

Lewis, D. D. (1998). Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval. *Machine Learning: ECML-98: 10th European Conference on Machine Learning Chemnitz*. Ed. by C. Nédellec y C. Rouveirol. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 4–15.

Li, Z., Xiong, Z., Zhang, Y., Liu, C. y Li, K. (2011). Fast Text Categorization using Concise Semantic Analysis. *Pattern Recognition Letters*, 32(3), pp. 441–448.

Li, Z., Li, P., Wei, W., Liu, H., He, J., Liu, T. y Du, X. (2009). AutoPCS: A Phrase-Based Text Categorization System for Similar Texts. *Proceedings in Advances in Data and Web Management: Joint International Conferences, APWeb/WAIM 2009 Suzhou, China, April 2-4, 2009*. Ed. by Q. Li, L. Feng, J. Pei, S. X. Wang, X. Zhou y Q.-M. Zhu. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 369–380.

Liu, Y., Loh, H. T. y Sun, A. (2009). Imbalanced Text Classification: A Term Weighting Approach. *Expert Systems with Applications*, 36(1), pp. 690–701.

López-Monroy, A. P., Montes-y-Gómez, M., Escalante, H. J. y Pineda, L. V. (2014). Using Intra-Profile Information for Author Profiling. *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, 15-18 September, Sheffield, UK*. Ed. by L. Cappellato, N. Ferro, M. Halvey y W. Kraaij. Vol. 1180. CEUR Workshop Proceedings. CEUR-WS.org, pp. 1116–1120.

López-Monroy, A. P., Montes-y-Gómez, M., Escalante, H. J., Villaseñor-Pineda, L. y Stamatatos, E. (2015). Discriminative subprofile-specific representations for author profiling in social media. *Knowledge-Based Systems*, 89, pp. 134–147.

López-Monroy, A. P., Montes-y-Gómez, M., Escalante, H. J., Villaseñor-Pineda, L. y Villatoro-Tello, E. (2013). INAOE's participation at PAN'13: Author Profiling task—Notebook for PAN at CLEF 2013. *CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain*. Ed. by P. Forner, R. Navigli y D. Tufis.

Lorenz, M. O. (1905). Methods of Measuring the Concentration of Wealth. *Publications of the American Statistical Association*, 9(70), pp. 209–219.

Maharjan, S., Shrestha, P., Solorio, T. y Hasan, R. (2014). A Straightforward Author Profiling Approach in MapReduce. *Advances in Artificial Intelligence – IBERAMIA 2014: 14th Ibero-American Conference on AI, Santiago de Chile, Chile, November 24-27, 2014, Proceedings*. Ed. by A. L. Bazzan y K. Pichara. Cham: Springer International Publishing, pp. 95–107.

Markov, I., Gómez-Adorno, H., Sidorov, G. y Gelbukh, A. (2016). Adapting Cross-Genre Author Profiling to Language and Corpus—Notebook for PAN at CLEF 2016. *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, Évora, Portugal*. Ed. by K. Balog, L. Cappellato, N. Ferro y C. Macdonald. CEUR-WS.org.

Marquardt, J., Farnadi, G., Vasudevan, G., Moens, M.-F., Davalos, S., Teredesai, A. y De Cock, M. (2014). Age and Gender Identification in Social Media. *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, 15-18*

September, Sheffield, UK. Ed. by L. Cappellato, N. Ferro, M. Halvey y W. Kraaij. CEUR-WS.org.

McCallum, A. y Nigam, K. (1998). A Comparison of Event Models for Naive Bayes Text Classification. *Learning for Text Categorization: Papers from the 1998 AAAI Workshop*, pp. 41–48.

McCrae, R. R. y John, O. P. (1992). An Introduction to the Five-Factor Model and its Applications. *Journal of personality*, 60(2), pp. 175–215.

Mehl, M. R., Vazire, S., Ramírez-Esparza, N., Slatcher, R. B. y Pennebaker, J. W. (2007). Are Women Really More Talkative Than Men? *Science*, 317(5834), pp. 82–82.

Mihalcea, R. y Hassan, S. (2005). Using the Essence of Texts to Improve Document Classification. *Proceedings of the Conference on Recent Advances in Natural Language Processing*. Borovets, Bulgaria.

Mitchell, T. M. (1997). Machine Learning. 1st ed. New York, NY, USA: McGraw-Hill, Inc.

Moschitti, A. y Basili, R. (2004). Complex Linguistic Features for Text Classification: A Comprehensive Study. *Proceedings of Advances in Information Retrieval: 26th European Conference on IR Research, ECIR 2004, Sunderland, UK, April 5-7, 2004.* Ed. by S. McDonald y J. Tait. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 181–196.

Mosteller, F. y Wallace, D. L. (1984). The Federalist Papers As a Case Study. *Applied Bayesian and Classical Inference: The Case of The Federalist Papers*. New York, NY: Springer New York, pp. 1–15.

Mukherjee, A. y Liu, B. (2010). Improving Gender Classification of Blog Authors. *Proceedings of the 2010 conference on Empirical Methods in*

natural Language Processing. Association for Computational Linguistics, pp. 207–217.

Newman, M. L., Groom, C. J., Handelman, L. D. y Pennebaker, J. W. (2008). Gender Differences in Language Use: An Analysis of 14,000 Text Samples. *Discourse Processes*, 45(3), pp. 211–236.

Newman, M. L., Pennebaker, J. W., Berry, D. S. y Richards, J. M. (2003). Lying Words: Predicting Deception from Linguistic Styles. *Personality and Social Psychology Bulletin*, 29(5), pp. 665–675.

Nguyen, D., Gravel, R., Trieschnigg, R. y Meder, T. (2013). How Old Do You Think I Am?" A Study of Language and Age in Twitter.

Nguyen, D., Smith, N. A. y Rosé, C. P. (2011). Author Age Prediction from Text Using Linear Regression. *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. LaTeCH '11. Portland, Oregon: Association for Computational Linguistics, pp. 115–123.

Nowson, S. y Oberlander, J. (2006). The Identity of Bloggers: Openness and Gender in Personal Weblogs. *Computational Approaches to Analyzing Weblogs: Papers from the 2006 AAAI Spring Symposium*.

Peersman, C., Daelemans, W. y Van Vaerenbergh, L. (2011). Predicting Age and Gender in Online Social Networks. *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*. SMUC '11. Glasgow, Scotland, UK: ACM, pp. 37–44.

Peng, F., Schuurmans, D. y Wang, S. (2003). Language and Task Independent Text Categorization with Simple Language Models. *Proceedings of the 2003 Conference of the North American Chapter of the Association for*

Computational Linguistics on Human Language Technology-Volume 1. Association for Computational Linguistics, pp. 110–117.

Pennacchiotti, M. y Popescu, A.-M. (2011). Democrats, Republicans and Starbucks Afficionados: User Classification in Twitter. *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '11. San Diego, California, USA: ACM, pp. 430–438.

Pennebaker, J. W., Booth, R. J. y Francis, M. E. (2007). Linguistic Inquiry and Word Count: LIWC [Computer software]. *Austin, TX: liwc. net*.

Pennebaker, J. W., Mehl, M. R. y Niederhoffer, K. G. (2003). Psychological Aspects of Natural Language Use: Our words, Our Selves. *Annual review of psychology*, 54(1), pp. 547–577.

Pennebaker, J. W. y Stone, L. D. (2003). Words of Wisdom: Language Use Over the Life Span. *Journal of personality and social psychology*, 85(2), pp. 291–301.

Pennebaker, J. (2011). The Secret Life of Pronouns: What Our Words Say About Us. Bloomsbury USA.

Phan, X.-H., Nguyen, L.-M. y Horiguchi, S. (2008). Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections. *Proceedings of the 17th International Conference on World Wide Web*. WWW '08. Beijing, China: ACM, pp. 91–100.

Posadas-Durán, J. P., Gómez-Adorno, H., Markov, I., Sidorov, G., Batyrshin, I. Z., Gelbukh, A. F. y Pichardo-Lagunas, O. (2015). Syntactic N-grams as Features for the Author Profiling Task: Notebook for PAN at CLEF 2015. *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11*

September, Toulouse, France. Ed. by L. Cappellato, N. Ferro, G. Jones y E. San Juan.

Przybyła, P. y Teisseyre, P. (2015). What do Your Look-alikes Say About You? Exploiting Strong and Weak Similarities for Author Profiling—Notebook for PAN at CLEF 2015. *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France*. Ed. by L. Cappellato, N. Ferro, G. Jones y E. San Juan. CEUR-WS.org.

Qiu, L., Lin, H., Ramsay, J. y Yang, F. (2012). You Are What You Tweet: Personality Expression & Perception on Twitter. *Journal of Research in Personality*, 46(6), pp. 710–718.

Rahmoun, A. y Elberrichi, Z. (2007). Experimenting N-Grams in Text Categorization. *International Arab Journal of Information Technology*, 4(4), pp. 377–385.

Ramnial, H., Panchoo, S. y Pudaruth, S. (2016). Gender Profiling from PhD Theses Using k-Nearest Neighbour and Sequential Minimal Optimisation. *Intelligent Systems Technologies and Applications: Volume 2*. Ed. by S. Berretti, S. M. Thampi y S. Dasgupta. Cham: Springer International Publishing, pp. 369–377.

Rangel, F., Celli, F., Rosso, P., Potthast, M., Stein, B. y Daelemans, W. (2015). Overview of the 3rd Author Profiling Task at PAN 2015. *CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France*. Ed. by L. Cappellato, N. Ferro, G. Jones y E. San Juan. CEUR-WS.org.

Rangel, F. y Rosso, P. (2013). Use of Language and Author Profiling: Identification of Gender and Age. *10th International Workshop on Natural*

Language Processing and Cognitive Cciences NLPCS 2013 CIRM. Marseille, France, October 13–17, pp. 177–186.

Rangel, F. y Rosso, P. (2016). On the Impact of Emotions on Author Profiling. *Information Processing & Management*, 52(1), pp. 73–92.

Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B. y Daelemans, W. (2014). Overview of the 2nd Author Profiling Task at PAN 2014. *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, 15-18 September, Sheffield, UK*. Ed. by L. Cappellato, N. Ferro, M. Halvey y W. Kraaij. CEUR-WS.org.

Rangel, F., Rosso, P., Koppel, M., Stamatatos, E. y Inches, G. (2013). Overview of the Author Profiling Task at PAN 2013. Ed. by P. Forner, R. Navigli y D. Tufis.

Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M. y Stein, B. (2016). Overview of the 4th Author Profiling Task at PAN 2016: Cross-Genre Evaluations. *Working Notes Papers of the CLEF 2016 Evaluation Labs*. Vol. 1609. CEUR Workshop Proceedings. CLEF and CEUR-WS.org.

Rao, D., Yarowsky, D., Shreevats, A. y Gupta, M. (2010). Classifying Latent User Attributes in Twitter. *Proceedings of the 2Nd International Workshop on Search and Mining User-generated Contents*. SMUC '10. Toronto, ON, Canada: ACM, pp. 37–44.

Reddy, T. R., Vardhan, B. V. y Reddy, P. V. (2016). A Survey on Authorship Profiling Techniques. *International Journal of Applied Engineering Research*, 11(5), pp. 3092–3102.

Rodríguez-Fdez, I., Canosa, A., Mucientes, M. y Bugarín, A. (2015). STAC: A Web Platform for the Comparison of Algorithms using Statistical Tests.

Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). Istanbul (Turkey).

Rosso, P., Perez Tellez, F., Pinto, D. y Cardiff, J. (2009). Defining and Evaluating Blog Characteristics. *2013 12th Mexican International Conference on Artificial Intelligence*, 00, pp. 97–102.

Rosso, P., Rangel, F., Potthast, M., Stamatatos, E., Tschuggnall, M. y Stein, B. (2016). Overview of PAN'16—New Challenges for Authorship Analysis: Cross-genre Profiling, Clustering, Diarization, and Obfuscation. *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 7th International Conference of the CLEF Initiative (CLEF 16)*. Ed. by N. Fuhr, P. Quaresma, B. Larsen, T. Gonçalves, K. Balog, C. Macdonald, L. Cappellato y N. Ferro. Berlin Heidelberg New York: Springer.

Rude, S., Gortner, E.-M. y Pennebaker, J. (2004). Language Use of Depressed and Depression-vulnerable College Students. *Cognition & Emotion*, 18(8), pp. 1121–1133.

Sahlgren, M. y Cöster, R. (2004). Using Bag-of-concepts to Improve the Performance of Support Vector Machines in Text Categorization. *Proceedings of the 20th International Conference on Computational Linguistics*. COLING '04. Geneva, Switzerland: Association for Computational Linguistics.

Salton, G. y Buckley, C. (1988). Term-weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, 24(5), pp. 513–523.

Sap, M., Park, G. J., Eichstaedt, J. C., Kern, M. L., Stillwell, D., Kosinski, M., Ungar, L. H. y Schwartz, H. A. (2014). Developing Age and Gender Predictive Lexica over Social Media. *Proceedings of the 2014 Conference*

on Empirical Methods in Natural Language Processing (EMNLP). Ed. by A. Moschitti, B. Pang y W. Daelemans. ACL, pp. 1146–1151.

Schler, J., Koppel, M., Argamon, S. y Pennebaker, J. (2006). Effects of Age and Gender on Blogging. *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, pp. 199–205.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E., et al. (2013). Personality, Gender, and Age in the Language of Social Media: The Open-vocabulary Approach. *PloS one*, 8(9), e73791.

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, 34(1), pp. 1–47.

Segarra, S., Eisen, M. y Ribeiro, A. (2013). Authorship Attribution Using Function Words Adjacency Networks. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pp. 5563–5567.

Shang, W., Huang, H., Zhu, H., Lin, Y., Qu, Y. y Wang, Z. (2007). A Novel Feature Selection Algorithm for Text Categorization. *Expert Systems with Applications*, 33(1), pp. 1–5.

Shankar, S. y Karypis, G. (2000). A feature weight adjustment algorithm for document categorization. *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Boston, MA, USA.

Shen, D., Yang, Q. y Chen, Z. (2007). Noise Reduction Through Summarization for Web-page Classification. *Information Processing and Management: an International Journal*, 43(6), pp. 1735–1747.

Singh, S. R., Murthy, H. A. y Gonsalves, T. A. (2010). Feature Selection for Text Classification Based on Gini Coefficient of Inequality. *The Fourth Workshop on Feature Selection in Data Mining*. Ed. by H. Liu, H. Motoda, R. Setiono y Z. Zhao. Vol. 10, pp. 76–85.

Song, F., Liu, S. y Yang, J. (2005). A Comparative Study on Text Representation Schemes in Text Categorization. *Pattern Analysis and Applications*, 8(1), pp. 199–209.

Spearman, C. (1987). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 100(3/4), pp. 441–471.

Stamatatos, E. (2009). A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*, 60(3), pp. 538–556.

Stein, B., Lipka, N. y Prettenhofer, P. (2011). Intrinsic Plagiarism Analysis. *Language Resources and Evaluation*, 45(1), pp. 63–82.

Surendran, K., Harilal, O. P., Hrudya, P., Poornachandran, P. y Suchetha, N. K. (2017). Stylometry Detection Using Deep Learning. *Computational Intelligence in Data Mining: Proceedings of the International Conference on CIDM, 10-11 December 2016.* Ed. by H. S. Behera y D. P. Mohapatra. Singapore: Springer Singapore, pp. 749–757.

Tausczik, Y. R. y Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of language and social psychology*, 29(1), pp. 24–54.

Tofighi, P., Köse, C. y Rouka, L. (2012). Author's Native Language Identification from Web-Based Texts. *International Journal of Computer and Communication Engineering*, 1(1), p. 47.

Tsoumakas, G. y Katakis, I. (2007). Multi Label Classification: An Overview. *International Journal of Data Warehouse and Mining*, 3(3). Ed. by D. Taniar, pp. 1–13.

Uysal, A. K. (2016). An Improved Global Feature Selection Scheme for Text Classification. *Expert Systems with Applications*, 43, pp. 82–92.

Uysal, A. K. y Gunal, S. (2014). The Impact of Preprocessing on Text Classification. *Information Processing & Management*, 50(1), pp. 104–112.

Villena Román, J. y González Cristóbal, J. C. (2014). DAEDALUS at PAN 2014: Guessing Tweet Author's Gender and Age—Notebook for PAN at CLEF 2014. *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, 15-18 September, Sheffield, UK*. Ed. by L. Cappellato, N. Ferro, M. Halvey y W. Kraaij. CEUR-WS.org.

Vinciarelli, A. y Mohammadi, G. (2014). A Survey of Personality Computing. *IEEE Transactions on Affective Computing*, 5(3), pp. 273–291.

Wang, D., Zhang, H., Liu, R. y Lv, W. (2012). Feature Selection Based on Term Frequency and T-test for Text Categorization. *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. CIKM '12. Maui, Hawaii, USA: ACM, pp. 1482–1486.

Weren, E. R. D., Moreira, V. P. y Palazzo M. de Oliveira, J. (2014a). Exploring Information Retrieval features for Author Profiling—Notebook for PAN at CLEF 2014. *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, 15-18 September, Sheffield, UK*. Ed. by L. Cappellato, N. Ferro, M. Halvey y W. Kraaij. Vol. 1180. CEUR Workshop Proceedings. CEUR-WS.org, pp. 1164–1171.

Weren, E. R., Kauer, A. U., Mizusaki, L., Moreira, V. P., Oliveira, J. P. M. de y Wives, L. K. (2014b). Examining Multiple Features for Author Profiling. *Journal of Information and Data Management*, 5(3), pp. 266–279.

Yan, X. y Yan, L. (2006). Gender Classification of Weblog Authors. *Proceedings of the AAAI Spring Symposium Series on Computation Approaches to Analyzing Weblogs*, pp. 228–230.

Yang, Y. y Liu, X. (1999). A Re-examination of Text Categorization Methods. *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '99. Berkeley, California, USA: ACM, pp. 42–49.

Yang, Y. y Pedersen, J. O. (1997). A Comparative Study on Feature Selection in Text Categorization. *Proceedings of the Fourteenth International Conference on Machine Learning*. ICML '97. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 412–420.

Yarkoni, T. (2010). Personality in 100,000 Words: A Large-scale Analysis of Personality and Word Use among Bloggers. *Journal of Research in Personality*, 44(3), pp. 363–373.

Zhang, W. y Gao, F. (2011). An Improvement to Naive Bayes for Text Classification. *Procedia Engineering*, 15, pp. 2160–2164.

Zhao, Y. y Zobel, J. (2005). Effective and Scalable Authorship Attribution Using Function Words. *Information Retrieval Technology: Second Asia Information Retrieval Symposium, AIRS 2005, Jeju Island, Korea, October 13-15, 2005. Proceedings*. Ed. by G. G. Lee, A. Yamada, H. Meng y S. H. Myaeng. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 174–189.

Zhu, W. y Lin, Y. (2013). Using GINI-index for Feature Weighting in Text Categorization. *Journal of Computational Information Systems*, 9 pp(14), pp. 5819–5826.

Zong, W., Wu, F., Chu, L. y Sculli, D. (2015). A Discriminative and Semantic Feature Selection Method for Text Categorization. *International Journal of Production Economics*, 165, pp. 215–222.