Benemérita Universidad Autónoma de Puebla
Facultad de Ciencias de la Computación

Identificación de perfiles de usuario

Tesis presentada como requisito para para obtener el título de:
Maestro en Ciencias de la Computación

Patricia Maria Espinoza Fong

Asesora:

Dra. Darnes Vilariño Ayala

Co-asesor:

Dr. David Eduardo Pinto Avendaño

Noviembre 2015

Resumen

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

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

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

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

Agradecimientos

Este trabajo se lo dedico a mis padres Edna Patricia Fong Rojo y Agustin Espinoza Gaxiola, a mis hermanos Agustin Espinoza Fong y Luis Angel Espinoza Fong y por último pero no menos importante a mi amor Alvaro Salazar Rios. Gracias por todo, no hay palabras para expresar la gratitud que siento por el apoyo incondicional que recibo de cada uno de ustedes, mejor familia no podría tener.

Agradezco a mi asesora Dra. Darnes Vilariño Ayala por todo su apoyo y guía durante este proceso. Pero sobre todo le agradezco por su dedicación a todo lo que hace y a sus alumnos, ya que nos hace sentir como parte de su familia, **Gracias**.

De igual manera agradezco a:

- Dr. David Eduardo Pinto Avendaño.
- Dra. Mireya Tovar Vidal.
- Dra. María Josefa Somodevilla García.
- Dr. Jose Luis Carballido Carranza.

Los cuales dedicaron su tiempo a la revisión y desarrollo de este trabajo de tesis, agradezco también a los que compartieron sus conocimientos conmigo dentro del salón de clases.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1. Planteamiento de la investigación

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

1.1.1. Problema a resolver

Internet en la actualidad ofrece diversas herramientas para que los usuarios puedan expresarse libremente sin importar la edad, el sexo, el tema que traten o a que se dediquen. La cantidad de conversaciones en línea (foros, salas de chats, redes sociales y blogs, entre otros medios) ha aumentado considerablemente. Dada esta situación es prácticamente imposible analizar manualmente una conversación y detectar el perfil del autor que la ha escrito.

La detección del perfil de un autor que puede ser edad, sexo, lenguaje nativo o tipo de personalidad es un problema que ha ganado importancia en aplicaciones forenses, de seguridad y de mercadotecnia. En particular este problema se está tratando a nivel internacional desde el año 2011. Inicialmente se buscaba detectar la participación de depredadores sexuales en las conversaciones. Hoy en día la comunidad de procesamiento de lenguaje Natural desea estudiar la forma en que se comunican los diferentes grupos de edades y sexo, tratando de detectar los patrones de escritura comunes y diferentes entre estos grupos.

En la presente investigación se desea, dada una conversación detectar la edad y el sexo de la persona que la ha escrito, por lo que se pretende encontrar patrones, características léxicas, sintácticas y semánticas propias de cada grupo de edad y género, para el desarrollo de modelos de aprendizaje.

1.1.2. Objetivos de la investigación

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

- **Objetivo general**
  - Desarrollar modelos para determinar el género y la edad de una persona a partir del estudio de textos cortos.
- **Objetivos específicos**
  - Analizar el trabajo realizado por el equipo de investigación de la facultad para esta tarea en PAN 2013.
  - Estudiar los modelos desarrollados por otros equipos participantes.
  - Desarrollar un primer modelo para participar en el PAN 2014 y que sirva de baseline en la investigación.
  - Desarrollar recursos lingüísticos para el pre procesamiento de los corpus. Estos recursos difieren de acuerdo al idioma.
  - Desarrollar algoritmos que permitan detectar las características léxicas, sintácticas y semánticas que permitan representar el perfil de un determinado autor, para lograr una representación adecuada de los textos.
  - Aplicar los modelos desarrollados en ambos idiomas.

1.1.3. Justificación de la investigación

La gran cantidad de conversaciones que ocurren día a día en la web ha propiciado una nueva forma de búsqueda de victimas por parte de personas maliciosas que intentan, interactuar fundamentalmente con niños o jóvenes. En la web se puede obtener todo tipo de información acerca de una persona sin necesidad de tener contacto físico con ella. Una persona puede crear un perfil falso y establecer todo tipo de conversación con individuos de menor edad sin que esto pueda detectarse.

A pesar de los avances que se han desarrollado en lingüística forense, la captura y seguimiento de depredadores sexuales y personas mal intencionadas en la web es un problema no resuelto, porque no existe ningún mecanismo regulador de la conversación que desarrolla cada individuo.

En particular el grupo de procesamiento de Lenguaje Natural lleva participando en la Conferencia Internacional Uncovering plagiarism, authorship, and social software misuse (PAN), desde hace 2 años. En el año 2012 se participó en la determinación si en conversaciones participaba algún depredador sexual a partir del estudio de conversaciones en salas de chats. En el año 2013 y 2014 se han desarrollado propuestas para la detección del perfil de un autor considerando solamente la edad y el sexo. A pesar de los esfuerzos realizados los resultados obtenidos en todos los años han sido bajos, sobre todo para el idioma Español.

Considerando lo anteriormente expresado el interés de esta tesis surge en la necesidad de desarrollar nuevos modelos de representación de las conversaciones y el descubrimiento de patrones representativos de los grupos de edad y por cada género.

1.1.4. Preguntas de investigación

A continuación se presentan un conjunto de preguntas que se espera resolver una vez concluida la investigación, y que de alguna manera guían los experimentos a desarrollar:

- ¿Qué características son las más propicias para detectar el perfil de un autor?
- ¿Qué características son las más propicias de acuerdo al tipo de corpus?
- ¿Qué tipo de clasificadores son los más adecuados para esta tarea en particular?
- ¿El comportamiento de los modelos es similar para ambos idiomas?

1.2. Aportaciones de la investigación

En esta sección se presentan las aportaciones de la investigación:

- Diseño de métodos para el procesamiento de los corpus, tanto para el idioma Español como para el idioma Inglés.
- Obtención de una metodología para la detección del perfil de un autor (edad, sexo).
- Descripción de las características que identifican a los autores.
- Desarrollo de recursos léxicos de apoyo a la tarea.

1.3. Organización de la tesis

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

- **Capítulo 1**. Introducción. En esta parte se detalla el problema a resolver, los objetivos de la investigación, la justificación, así como también la propuesta de solución al problema y los productos generados por el trabajo. Al mismo tiempo se presentan las aportaciones que genera la construcción de una metodología para la detección del perfil de una persona.
- **Capítulo 2**. Estado del arte. En esta parte se detallan un número importante de trabajos de investigación relacionados con el ámbito de esta investigación.
- **Capítulo 3**. Identificación del perfil de un usuario. Se describen conceptos y términos relacionados con las propiedades más importantes de esta investigación que se utilizarán en el desarrollo de este trabajo.
- **Capítulo 4**. Metodología de desarrollo. En este capítulo se detalla el marco metodológico de la presente investigación, que cubre como se va a resolver la identificación del perfil de un usuario en un documento.
- **Capítulo 5**. Pruebas y resultados. En este capítulo se comentan las pruebas y resultados experimentales de las diferentes metodologías propuestas en el capítulo 4. Se muestran también los algoritmos usados en cada aproximación y por último se hace una comparación de los resultados obtenidos.

- **Capítulo 6**. Conclusiones. Finalmente en este capítulo se muestran las conclusiones alcanzadas en el trabajo de investigación, así como se responden las preguntas de investigación plasmadas en la sección 1.1.4.

Capítulo 2

Estado del arte

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

2.1. Enfoque general

Se realizó un estudio sobre los trabajos desarrollados en esta área, enfatizando sus avances, alcance, enfoques, ventajas y desventajas, así como sus aportaciones científicas, encontrando el siguiente panorama general:

En la propuesta presentada en [1] se desarrollan dos modelos uno para el idioma Español y otro para el idioma Inglés, ambos totalmente diferentes. Para el idioma Inglés se extrajeron características léxicas y sintácticas, sin embargo para el idioma Español, se realizó una representación mediante grafos de las conversaciones y se extrajeron los patrones de cada clase utilizando la herramienta SUBDUE[1]. Se reporta que los resultados para el idioma Inglés superaron considerablemente los resultados obtenidos para el idioma Español.

En la investigación desarrollada en [2] se proponen 2 tipos de características que pueden ser usadas para esta tarea. Características basadas en el contexto y características basadas en el estilo. Las características basadas en el estilo incluyen características léxicas y sintácticas utilizando Pos-tagger como etiquetador. Para las características relacionadas al contexto se extraen las 1000 palabras individuales con mayor frecuencia de un corpus que incluye 19 320 post extraídos de blogs escritos en Inglés. Aplican además Información Mutua para detectar los pares de palabras que son colocaciones. Los

[1] https://ailab.wsu.edu/subdue/

resultados que obtuvieron muestran que las características estilográficas que más ayudan a discriminar el género son las preposiciones para el caso de los hombres y los pronombres para el caso de las mujeres. Y con respecto al contexto, los hombres utilizan palabras relacionadas con la tecnología y las mujeres utilizan más palabras relacionadas a la vida personal y a las relaciones.

El modelo propuesto en [3] se basa en el desarrollo de una variación del algoritmo *Exponential Gradient (EG)*, que permite detectar el género de un autor. Se propone una representación vectorial del conjunto de características que estudian y en cada paso bajo ciertos criterios de eliminación van reduciendo el espacio de representación, quitando aquellas características que aportan poco a la detección del género. Concluyen que las características más representativas son las palabras y las etiquetas de los textos.

En el trabajo desarrollado en [4] se estudia el comportamiento de hombres y mujeres blogueros, y mencionan que las características que mejores resultados ofrecieron son las palabras representativas de cada grupo, los hiperenlaces y palabras comúnmente usadas en los blogs (lol, haha, ur, entre otras).

Los resultados que obtuvieron con estas características fueron del 80 % para el género y del 76 % para la edad. Se llegó a la conclusión de que las mujeres usan más pronombres y los hombres más preposiciones, también mencionan que se encontró que los blogs escritos por adolescentes son en su mayoría mujeres, que las mujeres hablan más sobre su vida privada y familia, mientras que los hombres hablan más sobre tecnología y política.

En otros trabajos precedentes para abordar esta tarea se puede observar, que las características más comúnmente utilizadas son:

- N gramas de palabras, [5],[6],[7] y [8].
- N gramas de caracteres, [5] y [8].
- Longitud de palabras, [9], [6] y [10].
- Longitud de oraciones, [11], [12] y [10].

En [7] y [6] se utiliza la herramienta *Linguistic Inquiry and Word Count (LIWC)*, la cual calcula el grado en que las personas usan diferentes categorías entre un conjunto de documentos, también se puede determinar el grado en el que un texto utiliza emociones positivas o negativas entre otras cosas. Además de las características mencionadas anteriormente, en el trabajo propuesto en [6] también se cuenta la frecuencia de uso de palabras en mayúscula, la frecuencia de uso de intensificadores y la longitud de las oraciones. En [13] las características que se usan son las mencionadas anteriormente y se agregan el uso de signos de puntuación, el uso de emoticones y el uso de las categorías gramaticales POS.

En el trabajo propuesto en [11] se utiliza la frecuencia de las clases a las que pertenecen las palabras. La clasificación de las palabras se realiza con la herramienta *RiTaWordNet* la cual establece la relación de una palabra con su clase mediante sinónimos e hiperónimos. Posteriormente se clasifican las palabras en positivas o negativas usando *SentiWordNet 3.0*, se cuenta los signos de puntuación usados, la frecuencia de las palabras cerradas, frecuencia de uso de pronombres, se reemplazan los emoticones por su palabra equivalente y se cuantifica una lista de palabras foráneas (meee, yesss, thy, u, urs, entre otras).

El modelo propuesto en [9] utiliza algunas de las siguientes características, la frecuencia de uso de palabras escritas en formato *CamelCase* y la frecuencia de uso de etiquetas POS. También comentan que las personas jóvenes utilizan más los pronombres en primera persona y que las personas que no son originarias de los Estados Unidos usan más las abreviaciones "u" y "ur". Mencionan que al igual que en [2] se aplica Información Mutua para detectar los pares de palabras que son colocaciones.

Otro trabajo que es importante destacar es el presentado en [9], donde se mencionan algunas características interesantes como son los determinantes (*a, the, that, these*) y cuantificadores (*one, two, more, some*), que sirven como indicadores para identificar a un hombre y una vez más se menciona que los pronombres (*I, you, she, her, their, myself, yourself, herself*) son indicadores para identificar a una mujer, ya que según los autores las mujeres tienden a personalizar más los textos que escriben, mientras que los hombres los generalizan.

En el trabajo desarrollado por [14] se presenta un metodología para detectar el perfil de un autor, en particular se considera edad y género. Las características que utilizan son: categorías gramaticales, palabras cerradas, sufijos y signos. Los autores logran detectar solamente en un 55 % el género y a lo sumo un 45 % la edad. En este trabajo solo se presentan los resultados para el corpus de blogs en el idioma Inglés y el idioma Español ofrecido en la conferencia PAN 2013.

Entre las técnicas de clasificación más usadas se encuentran: Naive Bayes, que ha sido reportada en los trabajos desarrollados en [5],[10] y [12] y las máquinas de soporte vectorial (SMV) que han sido utilizadas en las investigaciones desarrolladas en [8], [5], [10] y [13].

Las características usadas fundamentalmente han sido léxicas, sintácticas y conteos de las frecuencias de uso de algunos elementos. En la presente investigación se pretende proponer características que de alguna manera permitan detectar el sentido del texto que se está estudiando y con ello analizar si es posible descubrir el perfil del autor.

En cuanto a las aproximaciones basadas en grafos, se estudiaron algunos trabajos para conocer el tipo de diseño que se utiliza al momento de crear los grafos en diferentes tareas. En el trabajo desarrollado por [15], lo que se busca es realizar consultas sobre una base de datos de grafos indexados, para esto, la representación de los grafos se hace por medio de un código o *canonical label* al que llaman *DFS Code*, si dos grafos son iguales entonces comparten el mismo código. Dicho código es generado al realizar una búsqueda en profundidad en el grafo.

De igual forma en [16] proponen un método para representar una imágen de manera formal, la cual consiste en un conjunto de objetos con propiedades y relaciones. Se busca hacer la representación a través de un grafo etiquetado dirigido, el problema que ellos abordan es el de cuales propiedades seleccionar para la construcción del grafo. En esta aproximación los objetos son representados por los nodos, y las relaciones y propiedades son las aristas.

En [17] el objetivo de los autores era diseñar un motor de búsqueda que hiciera uso de la estructura de hiperenlaces de la web para encontrar sitios web de interés. Este motor de búsqueda es capaz de encontrar no solo palabras

clave o de algún tema en particular, si no que puede buscar un hiperenlace con una estructura deseada. En ese grafo cada URL representa un vértice etiquetado como '_page_', las aristas están etiquetadas como '_hyperlink_' y apuntan de un URL padre a URL hijo. También se hace un análisis del texto de cada página, se eliminan signos de puntuación, palabras cerradas, etiquetas HTML y todas las palabras restantes se agregan al grafo como un nodo nuevo etiquetado con la palabra correspondiente y se relacionan con la página correspondiente (nodo '_page_') por medio de una arista etiquetada como '_word_'.

Otro trabajo donde utilizan grafos es en [18], donde el problema a resolver es la correferencia de entidades. Una entidad es un objeto o un conjunto de objetos del mundo real y una mención es una referencia textual a una entidad. El objetivo de este trabajo es identificar a que entidad hace referencia una mención; Para esto ellos utilizan una representación del espacio de correferencia mediante un grafo no dirigido, en donde los nodos representan todas las menciones en el texto y las aristas relacionan a los nodos que se refieren a la misma entidad. Cada arista tiene un peso asignado, el cual representa el grado de confianza de correferencia entre esos nodos.

En los trabajos [19] y [20] también se busca resolver el problema de correferencia. Ambos de igual forma que en el trabajo anterior, crean un grafo donde los nodos son las menciones y las aristas modelan una relación entre esas menciones. Cada arista tiene un peso asignado y en cada trabajo se utiliza un método específico (al que no se hará referencia) para calcular ese peso.

Por último en [14] se busca hacer un análisis del significado de un texto mediante una representación de ese texto en un grafo dirigido, en el cual las palabras del texto se representan por los nodos y las relaciones entre las palabras se representan por las aristas. Un punto interesante de este trabajo es que se crean aristas entre las palabras que están directamente conectadas (una detrás de otra), pero también se conectan palabras que están separadas por un número de palabras definido, para que las palabras que son usadas dentro de un mismo contexto estén conectadas.

Todos estos trabajos nos sirven como referencia para crear un modelo efectivo, pero es importante destacar que no importa el modelo que se esté evaluando, siempre va a ser más simple detectar el género, que la edad,

pues los hombres y las mujeres escriben o se interesan por temas diferentes independientemente de la edad que tienen. Un aspecto importante a estudiar es la técnica de clasificación que se debe usar y su comportamiento frente a los modelos en los que se aplique.

Capítulo 3

Identificación de perfiles de usuario

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

3.1. Características del texto

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

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

3.1.1. Características léxicas

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

El léxico del autor determina con que frecuencia se usan distintas palabras y en qué orden las usa para formar distintas ideas. Para determinar las propiedades léxicas del autor es necesario comprender el dominio que usa, así como el concepto que quiere plasmar, ya que de esto depende el uso del conjunto de palabras a usar, si bien es cierto que aunque los hombres y las mujeres hablen el mismo idioma no es necesariamente una condicional para que usen el mismo conjunto de palabras, por ejemplo el léxico no es el mismo para un blog de moda, que para un blog de política. Por lo anterior se puede ver que el léxico tiene que ver en como escribe una persona y lo que escribe.

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

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

3.1.2. Características sintácticas

Para escribir una o más palabras no sólo es necesario saber que palabras necesitamos para formar nuestra idea, sino que además se debe conocer el conjunto de reglas que gobiernan a las palabras, es decir, necesitamos saber la gramática (el estudio de las reglas y principios que regulan el uso de las lenguas y la organización de las palabras) para coordinar y unir las palabras,

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

Tabla 3.1: Tabla de las características léxicas más usadas.

así como para formar oraciones y expresar conceptos. Ahora bien, en el sentido anterior la sintaxis es el conjunto de pautas que definen las secuencias correctas de los elementos de un lenguaje.

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

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

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

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

3.1.3. Características semánticas

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

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

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

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

Características	**Herramienta requerida**
Dependencias semánticas	CLIPS(Tokenizador,Parser semántico,Pos tagger)

Tabla 3.3: Tabla de las características semánticas más usadas.

3.2. Corpus

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

La decisión del uso de un tipo de corpus se hace en base a las necesidades, en el caso de la identificación de perfiles de usuarios se necesita que el corpus a usar presente abundantes ejemplos de las diferentes clases que queremos identificar y que sean ricos en elementos que distingan a dichas clases. Teniendo lo anterior podemos encontrar la siguiente clasificación general de tipos de corpus [21]:

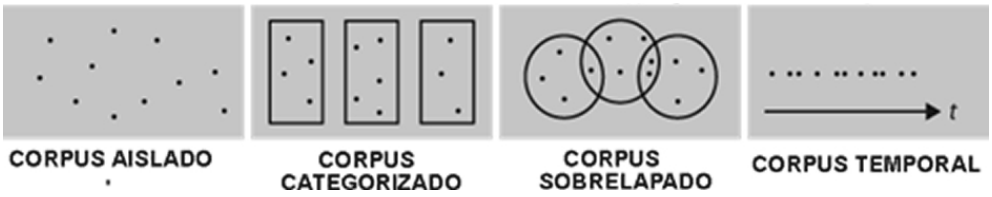

Figura 3.1: Clasificación de los tipos de corpus.

En la Figura 3.1 podemos encontrar 4 tipos fundamentales de corpus presentes en la mayoría de ámbitos:

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

- **Corpus temporal**: En este tipo de corpus existe un dominio claro, el cual se repite a través del tiempo variando los elementos presentes en cada documento.

En corpus que se utiliza para esta tarea es el siguiente:

- **Clef Pan Corpus**:Conjunto de datos en Inglés y en Español, propuesto en el marco del congreso PAN'14, el cual consta de 4 colecciones de textos en Inglés sobrelapados y desbalanceados acerca de blogs, críticas, redes sociales y twitter. Y para el idioma Español consta de 2 colecciones de textos sobrelapados y desbalanceados acerca de blogs y redes sociales.

Para una descripción más detallada de los documentos de cada corpus, ver la sección 5.1.

3.3. Clasificación de textos

La categorización de textos o clasificación de textos es la tarea de etiquetar o asignar a un nuevo documento una clase, basado en un conjunto de documentos previamente clasificados. Este tipo de clasificación es llamado un problema supervisado o semi supervisado [21], esto es, para clasificar una nueva entrada el clasificador entrena con un conjunto de documentos previamente etiquetados (corpus de entrenamiento).

3.3.1. Naïve Bayes

Es un clasificador probabilístico [22] basado en la aplicación del teorema de Bayes, tomando fuertes asunciones de independencia entre sus elementos. En términos simples el clasificador Naïve Bayes asume que la presencia u ausencia de una característica no está relacionada con la presencia u ausencia de otra característica. Para asignar una clase a un texto de origen desconocido, el clasificador calcula la probabilidad a priori sobre cada característica, la cual se calcula con la frecuencia de cada clase en el entrenamiento. La contribución de cada característica es combinada para obtener una probabilidad de estimación (*likelihood*), la clase con la mayor probabilidad de estimación as asignada al documento desconocido.

3.3.2. Máquina de soporte vectorial

Una máquina de soporte vectorial(Support Vector Machine, por sus siglas en Inglés SMO) es un sistema de aprendizaje basado en el uso de un espacio de hipotesis de funciones lineales en un espacio de mayor dimensión inducido por un Kernel, en el cual las hipótesis son entrenadas por un algoritmo tomado de la teoría de optimización, el cual utiliza elementos de la teoría de generalización. Debido a las limitaciones computacionales de las máquinas de aprendizaje lineal estas no pueden ser utilizadas en la mayoría de las aplicaciones del mundo real. La representación por medio del Kernel ofrece una solución alternativa a este problema, proyectando la información a un espacio de características de mayor dimensión, el cual aumenta la capacidad computacional de las máquinas de aprendizaje lineal.

3.3.3. IBK

Este algoritmo tiene la función de mapear las instancias en categorías: dada una instancia extraída de un conjunto, se produce una clasificación la cual es prevista por el atributo clasificador de dicha instancia. Esto es posible mediante 2 métodos principales:

1. Función de similitud: calcula la similitud entre una instancia del conjunto de entrenamiento i y las instancias del conjunto de pruebas.

2. Función de clasificación: esta recibe el resultado de la función de similitud y produce una clasificación para i.

Este algoritmo asume que instancias similares tienen la misma clase, de ahí la tendencia a clasificar instancias de acuerdo a su vecino más cercano. IBK difiere de la mayoría de los algoritmos de clasificación ya que este no construye abstracciones explícitas como árboles de decisión o reglas.

Capítulo 4

Metodología de solución

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

4.1. Primera aproximación: Modelo de conteos

En esta aproximación se desarrolló un modelo supervisado, se escogieron algunas de las características más utilizadas dentro del estado del arte como son, número de palabras de cierta longitud, número de símbolos, etc. Posteriormente estas características han sido extraídas del corpus de entrenamiento considerando cuantas veces aparecen en éste. A continuación se detallan todas las fases que permiten construir este modelo.

4.1.1. Pre procesamiento del corpus

Debido a que el corpus con el que se trabaja es descargado directamente de la página del PAN[1], es necesario hacer varias operaciones antes de trabajar con él, algunas de ellas son:

1. Separar el corpus por autor.

2. Separar el corpus por género.

[1] http://www.uni-weimar.de/medien/webis/events/pan-14/pan14-web/author-profiling.html

3. Sustituir los símbolos HTML que pueda contener el texto, por su equivalente en utf8.

Para el último punto se desarrolló un diccionario de símbolos HTML, el proceso se puede observar en la figura 4.1.

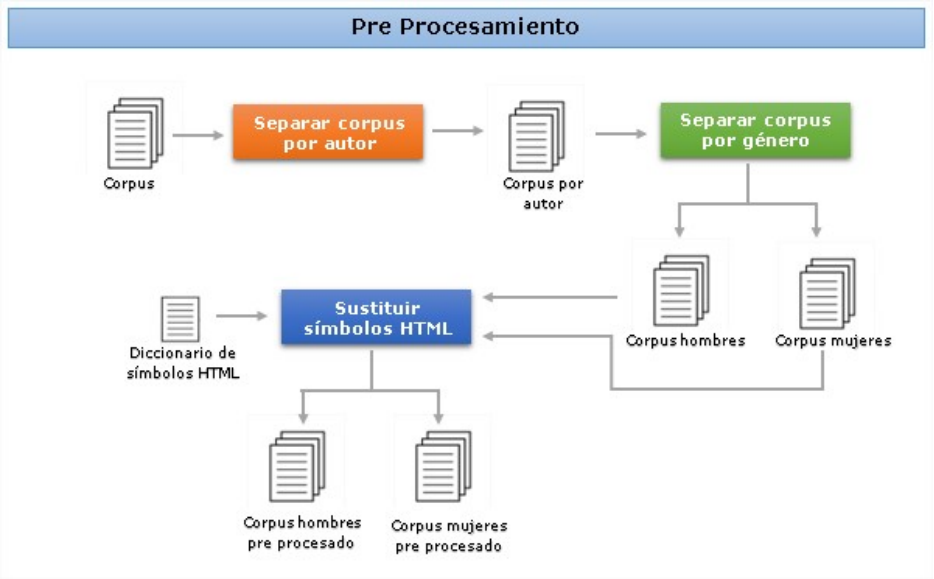

Figura 4.1: Pre procesamiento estándar del corpus.

4.1.2. Características seleccionadas

En el enfoque propuesto se desarrolla un modelo supervisado basado en máquinas de aprendizaje, para el cual se construye un modelo de clasificación usando el siguiente conjunto de 15 características léxicas, obtenidas del corpus proporcionado para esta tarea:

1. Número de slangs.
2. Número de contracciones.
3. Número de prefijos.
4. Número de signos.
5. Conteo de las 100 palabras más utilizadas.

6. Cantidad de palabras mal escritas.
7. Longitud de la oración.
8. Cantidad de números.
9. Número de palabras que empiezan con mayúscula.
10. Número de palabras escritas en mayúscula.
11. Longitud de la palabra más larga.
12. Número de palabras de longitud 1.
13. Número de palabras de longitud 2.
14. Número de palabras de longitud 10.
15. Número de palabras de longitud 15.

Para estos conteos se desarrollaron diferentes recursos léxicos como son: un diccionario de slangs, un diccionario de signos, diccionario de contracciones, diccionario de prefijos, diccionario de palabras más frecuentes y un diccionario que nos permite detectar si la palabra ha sido mal escrita.

El segundo conjunto de características está conformado por las categorías gramaticales. Para desarrollar este conteo, el texto fue etiquetado utilizando la herramienta de Clips llamada pattern.en[2], y se creó un diccionario con las categorías gramaticales que nos proporciona esta herramienta, que en total son 39.

Para el tercer conjunto de características se hizo un análisis del corpus para identificar cuáles son las 100 palabras más frecuentes de cada género, descartando las palabras cerradas y las palabras que se repiten en ambos corpus. Algunas de estas palabras que se extrajeron del corpus de blogs en Inglés se pueden observar en la tabla 4.1.

[2]www.clips.ua.ac.be/pages/pattern-en

Corpus mujeres	Corpus hombres
Art	Banks
Design	Building
Diet	Development
Exercise	Economy
Food	Financial
Gallery	Goverment
Health	Information
Heart	Job
Home	Money
Personal	Payment

Tabla 4.1: Palabras más frecuentes

Por último se agrega el texto de cada autor como una bolsa de palabras.

4.1.3. Representación de las características

Todas las características mencionadas en la sección 4.1.2 permiten construir un vector representativo de cada autor, considerando la frecuencia de aparición de cada una de las características seleccionadas.

Para la fase de entrenamiento, un ejemplo de dicho vector se muestra en la figura 4.2 donde el campo con el valor *Clase* al final del vector es el atributo clasificador del documento, que en el caso del género podría indicar si el documento pertenece a una mujer o a un hombre.

Figura 4.2: Vector de entrenamiento.

Para la fase de prueba se utiliza un vector de características como se muestra en la en la figura 4.3, donde el atributo clasificador se sustituye por un signo de interrogación, ya que se desconoce la clase a la que pertenece.

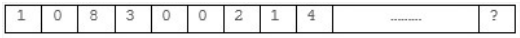

Figura 4.3: Vector de prueba.

4.1.4. Proceso de clasificación

La metodología desarrollada para la clasificación de los documentos mediante el modelo de conteos se puede observar en la figura 4.4, en ella se

muestra el proceso que se sigue y el cual se ha dividido en tres fases:

1. Fase de entrenamiento del modelo género.
2. Fase de entrenamiento de los modelos edadMujer y edadHombre.
3. Fase de prueba de los tres modelos.

Como se puede observar en la figura que se presenta a continuación:

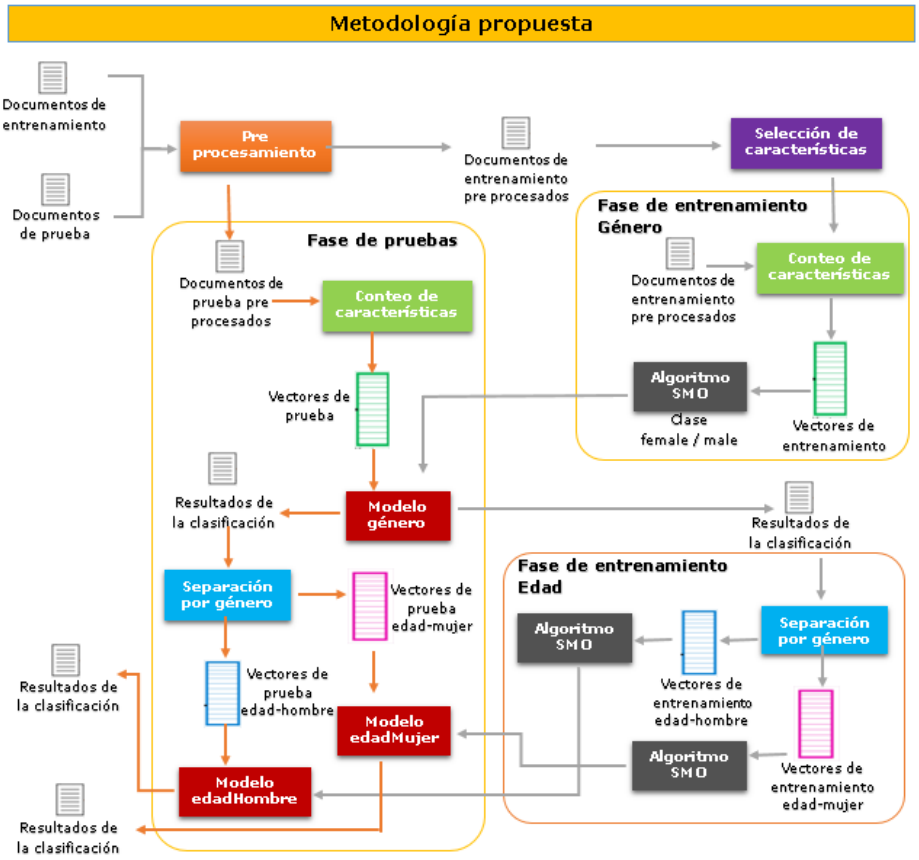

Figura 4.4: Metodología para el modelo de conteos.

Antes de construir o probar los modelos, se realiza el pre procesamiento descrito en la sección 4.1.1 sobre los corpus. Después se hace una selección de las características descritas en la sección 4.1.2.

Una vez seleccionadas las características se realiza el conteo de cada una de ellas y con esto se crea un vector por documento, en donde el atributo clasificador es el género del autor. Con este conjunto de vectores y el algoritmo SMO se crea el *Modelo de clasificación por género.*

En la segunda fase se utiliza el mísmo conjunto de vectores característicos que en la fase anterior, pero el atributo clasificador ahora es el rango de edad del autor. Aquí se crean dos modelos de clasificación diferentes, el *Modelo de clasificación de edadMujer* y el *Modelo de clasificación de edadHombre.* Como ya se sabe de antemano a qué género pertenecen los documentos gracias a la fase anterior, se puede discriminar los documentos para que a cada modelo sólo entren vectores que correspondan a ese género.

Como fase final se utilizan los documentos de prueba pre procesados para contar las características, crear los vectores, probar los modelos recién creados y obtener los resultados de la clasificación.

4.2. Segunda aproximación: Modelo de unigramas

Este modelo se deriva de una aproximación de la probabilidad de una secuencia de palabras, tal y como se menciona en [23], dicho modelo se basa en la propiedad de Markov, la cual dice que la probabilidad del siguiente estado sólo depende del estado actual. En nuestro contexto, la probabilidad de aparición de una palabra depende del conjunto de palabras anteriores a ella.

De acuerdo a este modelo de n-gramas, se puede definir un nuevo modelo de unigramas, calculando la probabilidad condicional que dependa de las n-1 palabras previas en la secuencia, como se indica en la expresión 4.1.

$$p(w) = p(w_1)p(w_2)p(w_3)...p(w_m) \tag{4.1}$$

Cabe resaltar que en este modelo el orden de las palabras no importa pues son independientes unas de otras, entonces la probabilidad de una palabra w_1, está dada por la expresión 4.2.

$$P(w_1) = \frac{numero(w_1)}{palabras} \tag{4.2}$$

Donde:

- **numero(w_1)**: Es el número de veces que aparece w_1 en el corpus.
- **palabras**: Es el número total de palabras en el corpus.

Con la probabilidad de cada palabra que conforma la conversación se construye un vector característico, el cual alimenta a un algoritmo de clasificación, a partir de los datos de entrenamiento.

4.2.1. Pre procesamiento del corpus

Para este modelo, además de aplicarse el pre procesamiento de la sección 4.1.1, al que nos referiremos como pre procesamiento estándar, se realiza el proceso que se observa en la figura 4.5. Se eliminan tanto palabras cerradas como símbolos, ya que no aportan significado al texto. En el próximo paso se sustituyen las contracciones por las palabras equivalentes y por último se sustituyen las palabras por su raíz (lema) correspondiente.

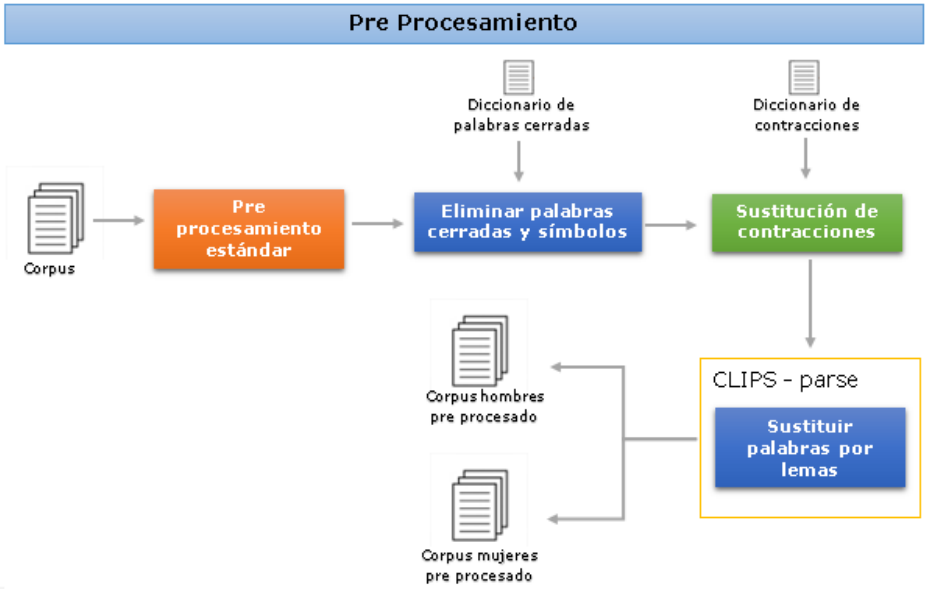

Figura 4.5: Pre procesamiento del corpus.

4.2.2. Representación de las características

Para representar el modelo de unigramas se utiliza un vector del tamaño del vocabulario del corpus, donde cada campo representa la probabilidad de cada palabra del vocabulario.

En la figura 4.6 se muestra el vector creado por este modelo, cuando el campo tiene un valor de cero significa que la palabra no aparece en el texto escrito por ese autor, en caso contrario el valor es la probabilidad que tiene esa palabra en el corpus.

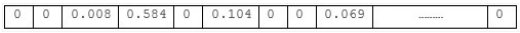

Figura 4.6: Vector de entrenamiento.

4.2.3. Proceso de clasificación

La metodología desarrollada para la clasificación de los documentos mediante el modelo de unigramas se puede observar en la figura 4.7, en ella se muestra el proceso que se emplea y el cual al igual que el proceso de clasificación de la sección 4.1.4 se ha dividido en tres fases.

Se realiza el pre procesamiento descrito en la sección 4.2.1 sobre los corpus. Después se hace el cálculo de la probabilidad de cada palabra del vocabulario y se crean los vectores de cada documento del corpus con las probabilidades correspondientes, como se explicó en la sección anterior.

Posteriormente se hace una selección de las características descritas en la sección 4.1.2 para hacer una combinación entre el modelo de unigramas y el modelo de conteos.

Una vez seleccionadas las características se realiza el conteo de cada una de ellas y con esto se crea un vector por documento, este vector se combina con el vector de probabilidades de ese mismo documento, y así para cada documento del corpus, por último se agrega al final del vector el atributo clasificador, que en este caso es el género del autor. Los vectores obtenidos son las características de los documentos del corpus, con el que se puede obtener un *Modelo de clasificación por género*, en particular se propone la

máquina de soporte vectorial (SMO).

En la segunda fase se utiliza el mísmo conjunto de vectores característicos que en la fase anterior, pero el atributo clasificador ahora es el rango de edad del autor. Aquí se crean dos modelos de clasificación diferentes, el *Modelo de clasificación de edadMujer* y el *Modelo de clasificación de edadHombre*.

Como fase final se utilizan los vectores de probabilidades de prueba, los cuales se unen con los vectores de prueba creados al contar las características del corpus pre procesado, se prueban los modelos recién creados y se obtienen los resultados de la clasificación.

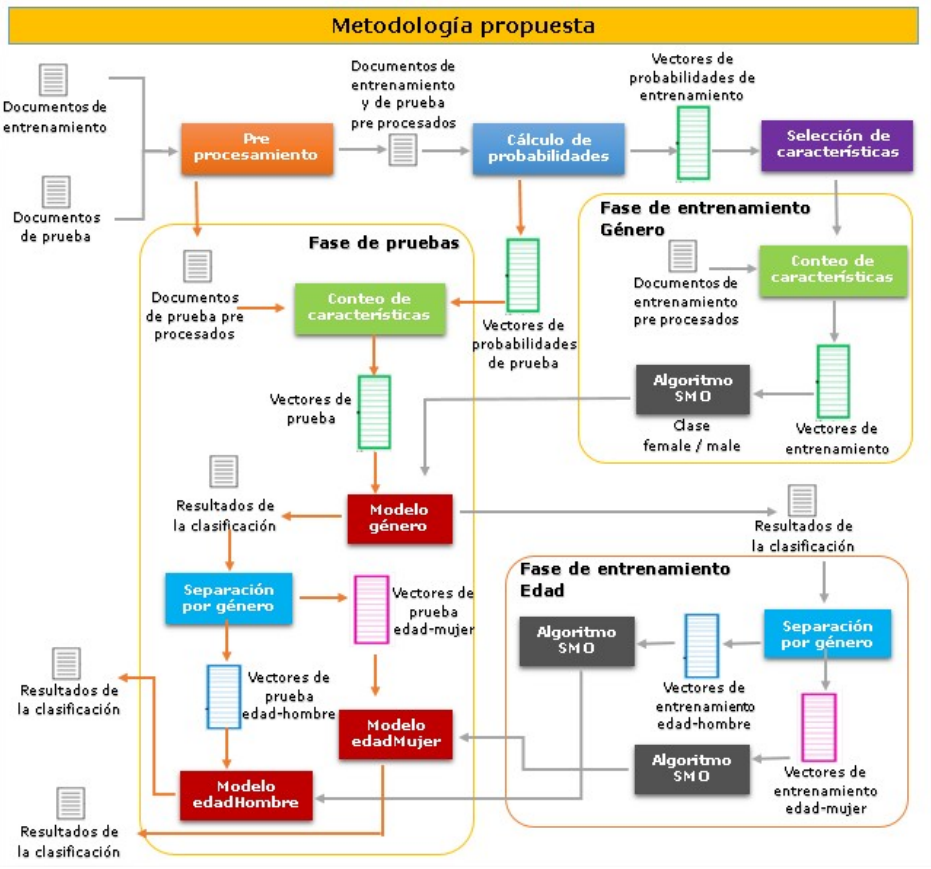

Figura 4.7: Metodología del modelo de unigramas.

4.3. Tercera aproximación: Clips

Se utilizó la herramienta pattern.vector [3] de Clips, ésta es un módulo que contiene herramientas de aprendizaje automático, entre las que se incluyen: funciones para conteo de palabras, modelo de bolsa de palabras, análisis semántico latente y algoritmos de clasificación y de clustering (Naïve Bayes, k-NN, Perceptron, SVM).

A continuación se describe el proceso que se desarrolló para la creación de este modelo.

4.3.1. Pre procesamiento del corpus

El proceso se puede observar en la figura 4.8.

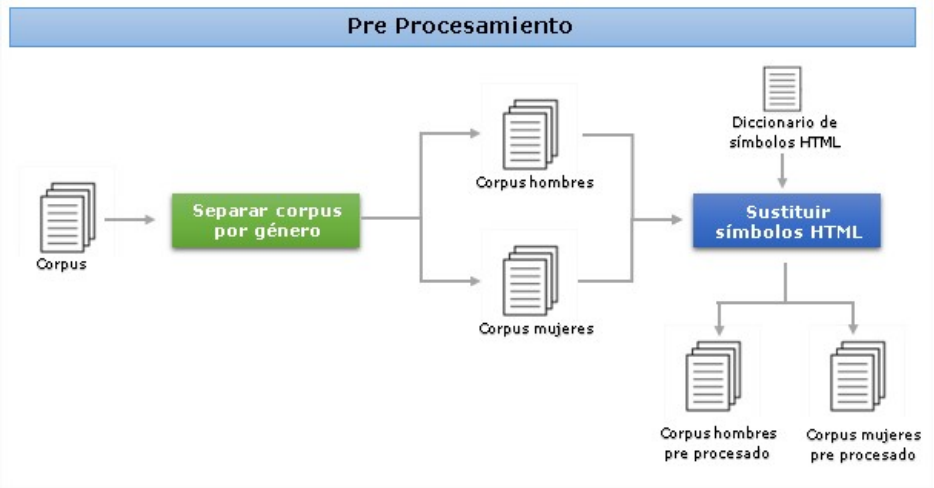

Figura 4.8: Pre procesamiento del corpus.

El pre procesamiento del corpus para este modelo se hizo de la siguiente manera:

1. Separar el corpus por género.
2. Sustituir los símbolos HTML que pueda contener el texto, por su equivalente en utf8.

[3] www.clips.ua.ac.be/pages/pattern-vector

Para el último punto se utilizó el mismo diccionario de símbolos HTML que en la sección 4.1. Como resultado se obtuvo un corpus de mujeres y otro de hombres, donde cada línea del corpus representa un muestra positiva para cada clase observada.

4.3.2. Creación del modelo

La forma en la que se representan los corpus en este modelo es a través de **Documentos** (cada muestra positiva es un documento) los cuales son una representación del texto mediante un vector, en el cual cada posición representa una palabra. Cada palabra tiene un peso asignado, el cual puede estar dado por:

1. Frecuencia bruta del término (F): *f(t,d)*= Número de veces que aparece el término t en el documento d.

2. Frecuencia booleana del término (FB): *f(t,d)*= 1 si t ocurre en d, y 0 si no.

3. Frecuencia normalizada del término (TF): está dada por la frecuencia bruta entre la frecuencia máxima de algún término en el documento.

$$tf(t,d) = \frac{f(t,d)}{max\{f(w,d) : w \in d\}} \tag{4.3}$$

4. Frecuencia normalizada por frecuencia inversa del documento(TF-IDF)

$$tfidf(t,d,D) = tf(t,d)xidf(t,D) \tag{4.4}$$

IDF se obtiene dividiendo el número total de documentos por el número de documentos que contienen el término, y se toma el logaritmo de ese cociente.

$$idf(t,D) = log\frac{|D|}{|\{d \in D : t \in d\}|} \tag{4.5}$$

Una vez que se elige el peso que se va a utilizar(que puede ser cualquiera de los mencionados anteriormente) y se crean los documentos, se genera un **Modelo** en donde se guarda cada vector/documento con la clase a la que pertenece.

4.3.3. Proceso de clasificación

La metodología desarrollada para la clasificación de los documentos mediante las herramientas que ofrece clips se puede observar en la figura 4.9, en ella se muestra el proceso que se sigue y el cual se ha dividido en tres fases.

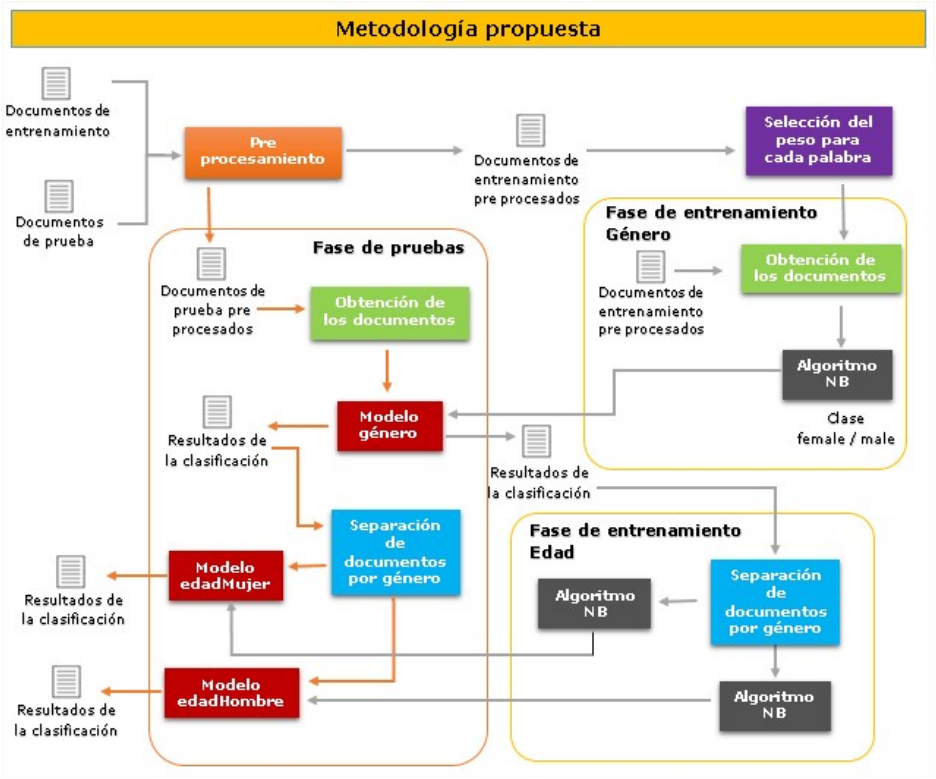

Figura 4.9: Metodología para el modelo de clips.

Para esta metodología a diferencia de las anteriores los vectores de entrenamiento son creados por la herramienta Clips y se utilizó el algoritmo de Naïve Bayes, ya que el de Máquinas de soporte vectorial (SMO) consume más recursos y no fue posible realizar los experimentos por restricciones de Hardware.

4.4. Cuarta aproximación: Gephi

Para esta aproximación se desarrolla un grafo de co-ocurrencia; el tamaño de la ventana considerado y el diseño del grafo es el presentado en el trabajo [14] . En dicho grafo las palabras o conceptos son los nodos y las aristas representan las relaciones entre estas palabras. Para la creación y extracción de la información del grafo se han utilizado las herramientas NetworkX[4] (creación del grafo) y Gephi[5] (análisis del grafo).

A continuación se presenta el proceso realizado para este modelo.

4.4.1. Pre procesamiento del corpus

En la siguiente figura se puede observar el proceso que se realiza para preparar el texto antes de la creación del grafo.

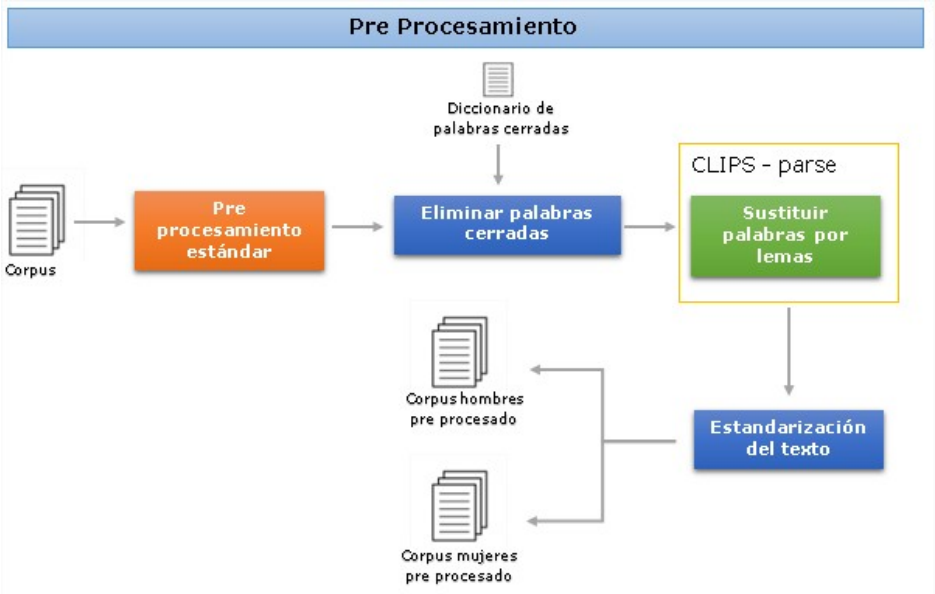

Figura 4.10: Pre procesamiento del texto para la creación del grafo.

[4]https://networkx.github.io/
[5]http://gephi.github.io/

Como primer paso se realiza el pre procesamiento estándar de la sección 4.1.1, posteriormente se remueven del corpus las **palabras cerradas**(artículos, conjunciones, verbos auxiliares, etc) , ya que son las que se utilizan con más frecuencia, pero en realidad no aportan significado o no cambian el contexto del texto. La detección de las palabras cerradas se hace a partir de un diccionario en Inglés y otro en Español.

En el tercer paso se sustituyen las palabras restantes en el texto por su correspondiente lema, esto se hace con el objetivo de simplificar y hacer más eficientes los procesos siguientes, ya que se reduce la complejidad de la red resultante, disminuyendo el tamaño del vocabulario. Para realizar este proceso se utilizó la función parse[6] que viene dentro de la librería de Clips utilizada en aproximaciones anteriores.

Como último paso se eliminan los signos de puntuación, los números y se lleva todo el texto a minúsculas (lo que evita que una misma palabra sea considerada como dos palabras diferentes).

Un fragmento del texto resultante se puede observar a continuación:

```
currently see wave idea datum center throw traditional model
datum center management air accelerate demand process datum
storage capacity globally come together environmental demand
create area.
```

4.4.2. Creación del grafo

Despúes de realizar el pre procesamiento de los corpus, el siguiente paso es usar el texto resultante para crear un grafo de co-ocurrencia. Este tipo de grafos se ha convertido en una de las formas más simples y efectivas de representar las relaciones entre las palabras, ya que su implementación es muy fácil de realizar.

Se dice que dos palabras co-ocurren si entre ellas se encuentra un número fijo de palabras, a esto se le llama ventana. En este caso se utilizaron dos

[6] http://www.clips.ua.ac.be/pages/pattern-en

tipos de ventanas: una para relacionar los términos que estan uno junto al otro (ventana de 0), y otra para relacionar palabras dentro de una ventana igual a 3. El objetivo de la segunda ventana es el de reforzar la relación entre palabras que ocurren en contextos similares.

Un grafo de co-ocurrencia dirigido G es un par ordenado G=(V,E), donde:

- V : Conjunto de vértices o nodos los cuales representan las palabras del texto.

- E : Conjunto de pares ordenados de elementos de V que representan la relación entre estos nodos.

En la figura 4.11 se puede observar el grafo para la siguiente oración: *"currently see wave idea datum center throw traditional model datum center management air accelerate demand process datum storage capacity globally come together environmental demand create area"*; Se muestran las relaciones que se crean entre las palabras no secuenciales.

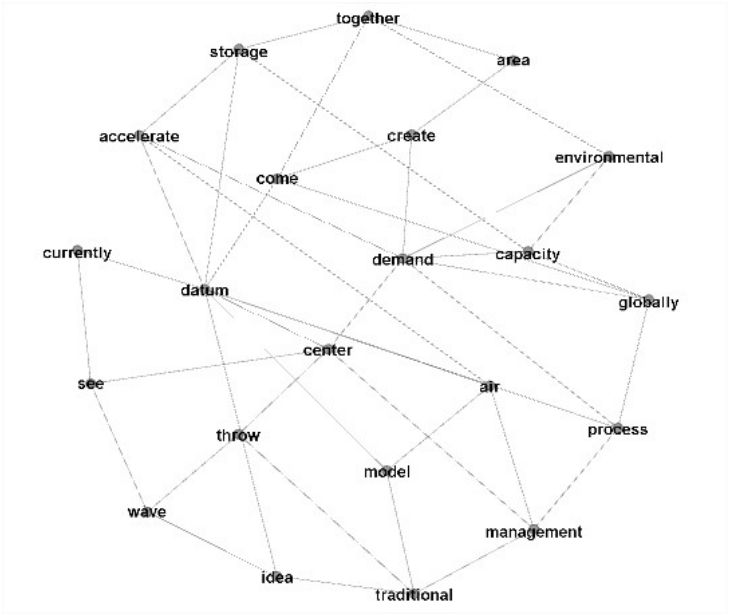

Figura 4.11: Grafo de co-ocurrencia.

El proceso para la creación del grafo se puede observar en la figura 4.12. Se creó un grafo por género {female, male}, se separó el corpus por grupos de edad y se creó un grafo por cada grupo de edad, este proceso se realizó por cada corpus en Inglés y en Español. Al final se obtuvo un total de **72** grafos, los cuales se guardan en un formato xml, para posteriormente crear una representación visual del grafo por medio de Gephi y calcular las medidas de centralidad deseadas.

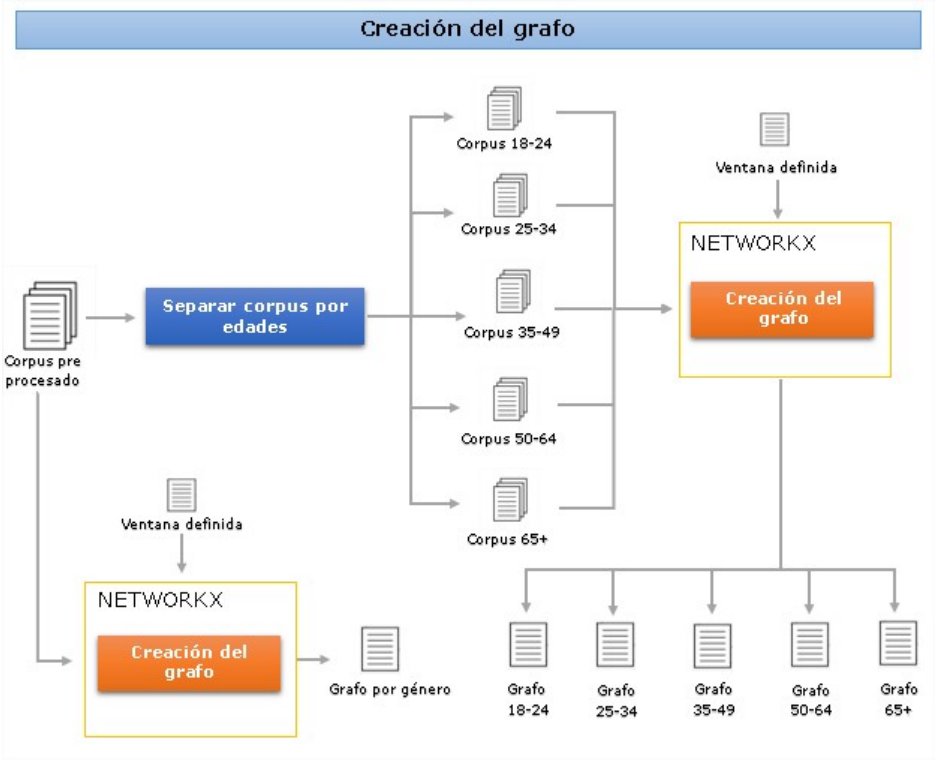

Figura 4.12: Creación del grafo.

4.4.3. Extracción de las carácteristicas del grafo

El desarrollo de grafos de co-ocurrencia permite extraer las palabras relevantes dentro del texto, por medio de medidas de centralidad y de modularidad, disponibles dentro de la herramienta de análisis de grafos Gephi. Estas medidas se explican a continuación:

- **Interconectividad (Betweenness centrality)**: es un indicador de la centralidad de un nodo dentro de la red. Es igual al número de veces que se pasa por ese nodo para llegar a otros nodos en el grafo o dicho de otra forma, es el número de veces que un nodo aparece al calcular el camino más corto de los otros nodos en la red. Los nodos con una interconectividad alta se pueden decir que son los que tienen mayor influencia dentro de la red, ya que son capaces de representar el contexto en el que se encuentra una cierta palabra.

En la figura 4.13 se tiene un ejemplo de esta medida, utilizando el mismo grafo de la sección anterior, pero ahora el tamaño de los nodos está dado por el grado de interconectividad, fácilmente se puede observar que los más grandes son los más interconectados ya que conectan los dos extremos del grafo.

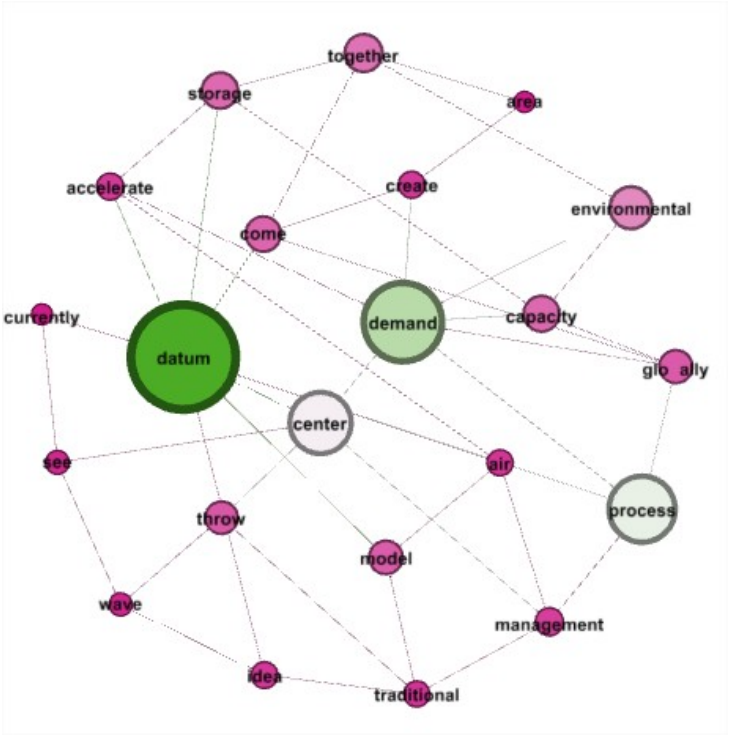

Figura 4.13: Ejemplo de interconectividad.

- **Modularidad (Modularity)**: Mide la fuerza con la que se divide una

red en módulos (grupos, clusters o comunidades). Los grafos con gran modularidad tienen conexiones densas entre los nodos que se encuentran en el mismo módulo y conexiones escasas entre nodos de otros módulos. Esta medida nos ayuda a encontrar palabras que se relacionan en torno a un tema.

Para el mismo ejemplo, en la figura 4.14 se muestran por color los clusters en los que se agrupan los nodos del grafo, el tamaño de los nodos está dado por la medida anterior.

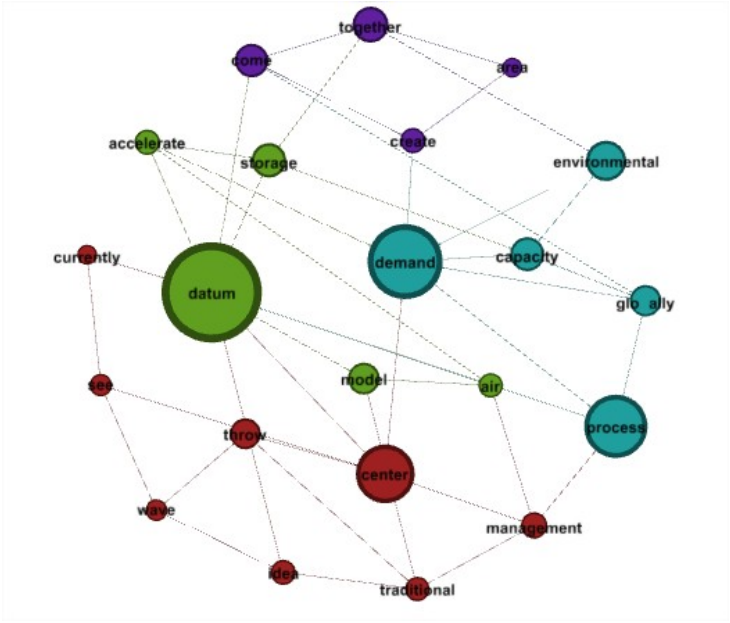

Figura 4.14: Ejemplo de modularidad.

El proceso para el análisis del grafo se puede observar en la figura 4.15. Se recibe el grafo en formato XML y se calcula el grado de interconectividad entre los nodos, esta herramienta permite observar visualmente los nodos con mayor interconectividad del grafo, ya que se puede filtrar por tamaño y color.

La segunda medida que se calcula es la modularidad, para que agrupe los nodos por comunidades y se puedan distinguir cada comunidad con un color. Al final lo que interesa es obtener una lista de palabras, en donde cada palabra tenga 2 medidas, el grado de interconectividad y la comunidad a la que pertenece.

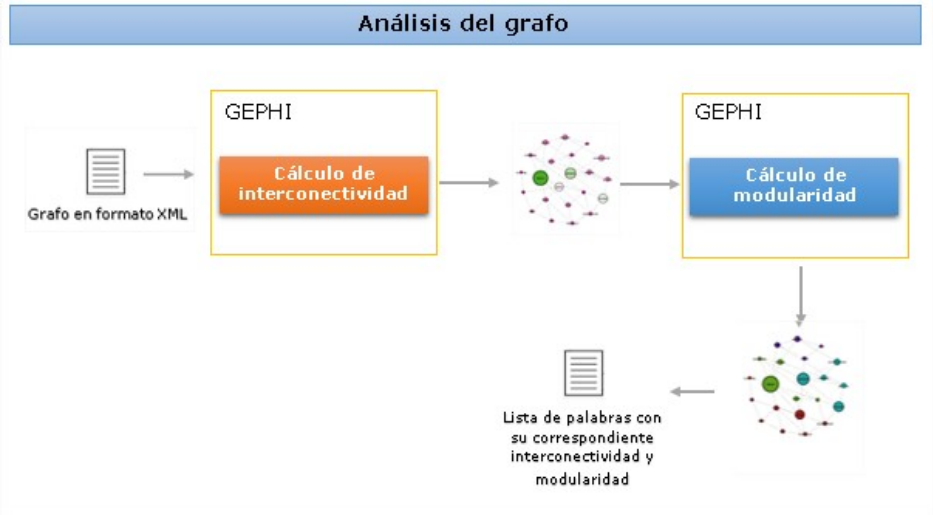

Figura 4.15: Análisis del grafo.

4.4.4. Proceso de clasificación

Se desarrolló un modelo supervisado el cual se puede observar en la figura 4.16. Como primer paso se tiene el pre procesamiento que se realiza para preparar los corpus, posteriormente la creación y análisis del grafo. Después se seleccionan las características o palabras relevantes para ese corpus y esa clase, y se realiza un conteo de las veces que aparece cada palabra en cada documento. También se utilizan todas las comunidades resultantes del análisis y cada vez que se cuenta una palabra, se incrementa el valor de la comunidad o comunidades a las que pertenece.

Se genera un vector por cada documento, donde la longitud de éste es igual al número de palabras elegidas más el número de comunidades . Cada posición del vector corresponde al número de veces que aparece esa palabra en el documento y en el caso de las comunidades, corresponde al número de palabras que pertenecen a esa comunidad en el documento. El atributo clasificador corresponde al género del autor. Ya que se tienen listos estos vectores se utiliza el algoritmo SMO para crear el *Modelo de clasificación por género.*

Posteriormente se separan por género los vectores y se les asigna el atributo clasificador correspondiente al rango de la edad del autor. Aquí se crean

dos modelos de clasificación diferentes, el *Modelo de clasificación de edadMujer* y el *Modelo de clasificación de edadHombre*. Para que a cada modelo solo entren vectores que correspondan a ese género.

En la fase de pruebas se realiza el mismo proceso para crear los vectores con los documentos de prueba y se evalúan los modelos construídos.

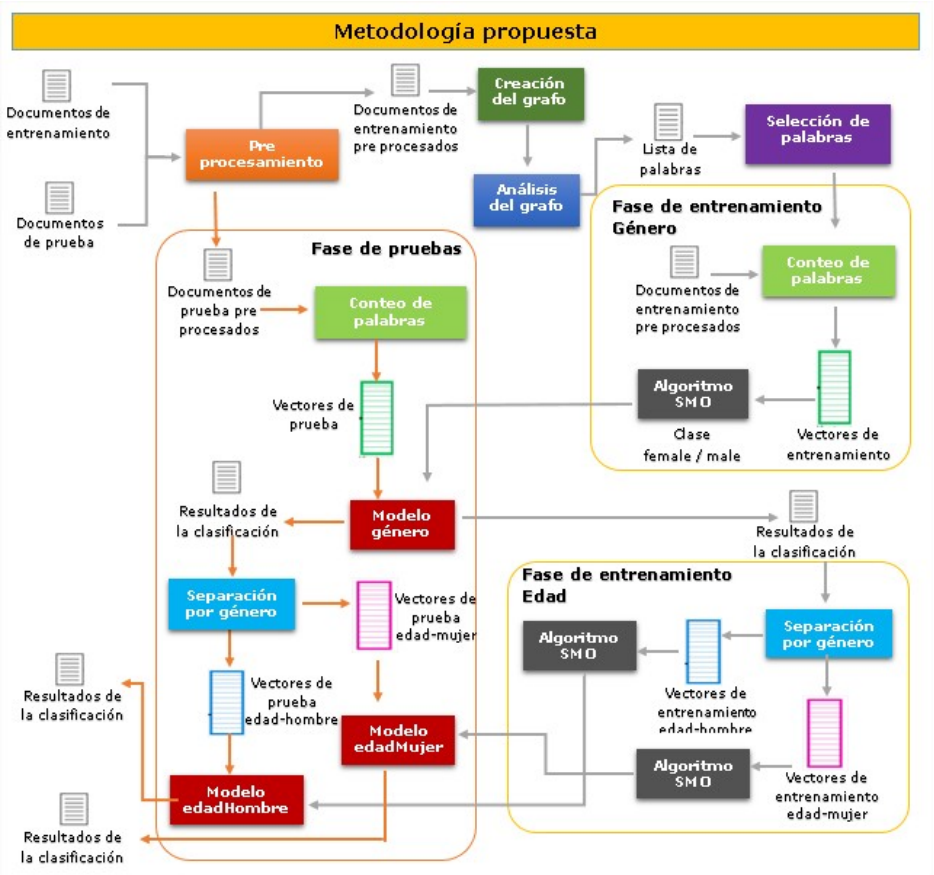

Figura 4.16: Metodología para el modelo creado a partir de Gephi.

Capítulo 5

Pruebas y resultados

En esta sección se presentan y discuten los resultados obtenidos en la tarea de identificación de perfiles de usuario, se describen las pruebas realizadas sobre cada aproximación propuesta, y se comparan los mejores resultados.

5.1. Conjunto de datos

Se trabajó con un conjunto de datos en Inglés obtenidos del sitio web del PAN 2014, cuya descripción se puede ver en la tabla 5.1.

Tipo de corpus	autores	Número de documentos
Blogs	147	2,273
Reviews	4,160	5,452
Social media	7,032	133,319
Twitter	301	197,772

Tabla 5.1: Descripción de los corpus en Inglés usados para esta tarea.

La descripción del conjunto de datos en Español, se puede ver en la tabla 5.2.

Tipo de corpus	autores	Número de documentos
Blogs	88	1,668
Social media	1,272	22,097

Tabla 5.2: Descripción de los corpus en Español usados para esta tarea.

5.1.1. Distribución del conjunto de datos

En esta sección se presentan las tablas 5.3 y 5.4 en ellas se muestra el número de instancias con las que se cuenta en los diferentes corpus. La columna Autores representa el número de autores, las columnas 18 a 24, 25 a 34, 35 a 49, 50 a 64 y 65 o más representan los rangos de edad de estos autores.

Género	Autores	18 a 24	25 a 34	35 a 49	50 a 64	65 o más
Hombre(blog)	74	3	30	27	12	2
Mujer(blog)	73	3	30	27	11	2
Hombre(review)	2,080	180	500	500	500	400
Mujer(review)	2,080	180	500	500	500	400
Hombre(socialmedia)	3,529	693	945	1035	851	5
Mujer(socialmedia)	3,503	699	944	1025	828	7
Hombre(twitter)	149	9	44	63	29	4
Mujer(twitter)	152	10	43	65	30	4

Tabla 5.3: Número de instancias por corpus en Inglés y por clase.

Género	Autores	18 a 24	25 a 34	35 a 49	50 a 64	65 o más
Hombre(blog)	44	2	13	21	6	2
Mujer(blog)	44	2	13	21	6	2
Hombre(socialmedia)	636	165	213	162	80	16
Mujer(socialmedia)	636	165	213	162	80	16

Tabla 5.4: Número de instancias por corpus en Español y por clase.

5.2. Primera aproximación

En la primera aproximación (véase el Capítulo 4.1) se utilizan los 3 algoritmos de clasificación (véase el Capítulo 3.3) para determinar el género y rango de edad de un autor.

Todos los experimentos se hicieron por autor aplicando validación cruzada de 10 pliegues. En las siguientes tablas se muestran los resultados obtenidos utilizando los algoritmos de clasificación IBk, SMO y Naïve Bayes sobre el conjunto de características escogidas, las cuales fueron descritas en la sección 4.1.2. Se muestra en **negritas** los mejores resultados de cada experimento y en **rojo** los mejores resultados para ese corpus.

5.2.1. Descripción de los experimentos

El objetivo de estos experimentos, es probar las características reportadas en trabajos actuales (véase sección 2) como las mejores y más utilizadas sobre el corpus proporcionado para esta tarea. También se busca encontrar el

algoritmo de clasificación con mejor comportamiento para que sea utilizado en aproximaciones posteriores.

A continuación se muestra una descripción de las características que incluye cada experimento:

Experimento 1: Total 15 características.

1. Número de slangs.
2. Número de contracciones.
3. Número de prefijos.
4. Número de signos.
5. Conteo de las 100 palabras más utilizadas.
6. Número de palabras mal escritas.
7. Longitud de la oración.
8. Cantidad de números.
9. Número de palabras que empiezan con mayúscula.
10. Número de palabras escritas en mayúscula.
11. Longitud de la palabra más larga.
12. Número de palabras de longitud 1.
13. Número de palabras de longitud 2.
14. Número de palabras de longitud 10.
15. Número de palabras de longitud 15.

Experimento 2: Características del **Experimento 1** combinadas con las 39 categorías gramaticales. Total 54 características.

Experimento 3: Características del **Experimento 2**, pero se removieron del corpus las palabras cerradas, los signos de puntuación y los números.

Con el objetivo de que no se cause ruido al momento de contar las categorías gramaticales (el número de signos y la cantidad de números se cuentan antes de ser removidos). Total 54 características.

Experimento 4: Características del **Experimento 2** combinadas con las 100 palabras más frecuentes de los hombres y las 100 palabras más frecuentes de las mujeres. Total 254 características.

Experimento 5: Características del **Experimento 4** combinadas con una bolsa de palabras del texto de cada documento. Total 254 características más el tamaño del vocabulario del corpus.

5.2.2. Resultados obtenidos para el corpus en Inglés

El mejor clasificador para la mayoría de los experimentos de la tabla 5.5 es el de máquinas de soporte vectorial (SMO), para este corpus las mejores características son las del Experimento 5.

Blogs

Características	Tipo de clasificación	Naïve Bayes	SMO	IBk
Experimento 1	Por género	61.22	**62.58**	49.65
	Por edad (mujeres)	35.61	**42.46**	30.13
	Por edad (hombres)	**40.54**	39.18	36.48
Experimento 2	Por género	58.5	**61.9**	45.57
	Por edad (mujeres)	30.13	**39.72**	24.65
	Por edad (hombres)	32.43	**43.24**	40.54
Experimento 3	Por género	58.5	**62.58**	49.65
	Por edad (mujeres)	28.76	**43.83**	24.65
	Por edad (hombres)	29.72	**47.29**	43.24
Experimento 4	Por género	60.54	**67.34**	59.18
	Por edad (mujeres)	35.61	**38.35**	38.35
	Por edad (hombres)	31.08	**44.59**	37.83
Experimento 5	Por género	61.9	**68.7**	60.54
	Por edad (mujeres)	**38.35**	36.98	38.35
	Por edad (hombres)	45.94	**54.05**	47.29

Tabla 5.5: Resultados de la primera aproximación para el corpus de blogs en Inglés.

Al igual que en el corpus de blogs, el de reviews arrojó resultados similares (se muestran en la tabla 5.6 , SMO como mejor clasificador con el mejor resultado superior al 60 % en el género, pero resultados muy bajos para la detección de la edad.

Reviews

Características	Tipo de clasificación	Naïve Bayes	SMO	IBk
Experimento 1	Por género	51.82	**53.89**	51.32
	Por edad (mujeres)	22.01	**27.3**	21.63
	Por edad (hombres)	26.15	**28.31**	23.6
Experimento 2	Por género	50.79	**57.88**	51.82
	Por edad (mujeres)	21.39	**27.93**	22.78
	Por edad (hombres)	21.73	**29.61**	24.08
Experimento 3	Por género	51.29	**56.17**	52.06
	Por edad (mujeres)	22.59	**27.83**	24.37
	Por edad (hombres)	22.93	**28.05**	24.13
Experimento 4	Por género	52.59	**59.44**	52.33
	Por edad (mujeres)	22.64	**27.45**	23.75
	Por edad (hombres)	22.4	**28.65**	23.26
Experimento 5	Por género	60.5	**65.69**	52.54
	Por edad (mujeres)	27.21	**30.28**	22.93
	Por edad (hombres)	24.8	27.83	23.07

Tabla 5.6: Resultados de la primera aproximación para el corpus de reviews en Inglés.

Los resultados para el corpus de socialmedia se muestran en la tabla 5.7, sorpresivamente el algoritmo IBk superó a los resultados obtenidos al aplicar SMO. En el campo que se muestra con –, significa que no se pudieron obtener los resultados por de hardware. Con respecto a la edad, se observa que los resultados no superan el 40 %.

Socialmedia

Características	Tipo de clasificación	Naïve Bayes	SMO	IBk
Experimento 1	Por género	**51.46**	51.05	51.18
	Por edad (hombres)	30.4	**36.66**	30.8
Experimento 2	Por género	50.99	51.39	**51.52**
	Por edad (mujeres)	25.77	**37.59**	30.28
	Por edad (hombres)	29.01	**36.78**	31.02
Experimento 3	Por género	51.08	50.49	**52.11**
	Por edad (mujeres)	27.14	**37.65**	30.65
	Por edad (hombres)	29.13	**36.61**	30.91
Experimento 4	Por género	51.4	51.67	**52.38**
	Por edad (mujeres)	25.17	**37.91**	32.57
	Por edad (hombres)	27.51	**36.21**	31.68
Experimento 5	Por género	51.89	–	**52.06**
	Por edad (mujeres)	27.97	30.77	**31.85**
	Por edad (hombres)	**32.47**	30.46	31.7

Tabla 5.7: Resultados de la primera aproximación para el corpus de socialmedia en Inglés.

Para el corpus de twitter (tabla 5.8), el mejor clasificador fue sin duda SMO, se obtuvo de los mejores resultados para la edad, alcanzando casi el 50 % y se obtuvo el mejor resultado para el género, superando el 80 %.

Twitter

Características	Tipo de clasificación	Naïve Bayes	SMO	IBk
Experimento 1	Por género	50.84	**58.05**	54.66
	Por edad (mujeres)	33.09	**40.28**	30.21
	Por edad (hombres)	28.86	**47.42**	28.86
Experimento 2	Por género	51.27	**62.71**	53.81
	Por edad (mujeres)	27.33	**48.29**	35.25
	Por edad (hombres)	27.83	**46.39**	31.95
Experimento 3	Por género	50	**60.16**	54.23
	Por edad (mujeres)	22.3	**46.04**	36.69
	Por edad (hombres)	31.95	**49.48**	38.14
Experimento 4	Por género	56.77	**68.64**	55.93
	Por edad (mujeres)	29.49	**38.84**	31.65
	Por edad (hombres)	28.86	**43.29**	35.05
Experimento 5	Por género	71.18	**83.05**	65.67
	Por edad (mujeres)	32.37	**46.04**	31.65
	Por edad (hombres)	35.05	**49.48**	44.32

Tabla 5.8: Resultados de la primera aproximación para el corpus de twitter en Inglés.

En resumen se puede decir que el mejor resultado para el modelo de género se alcanzó con el corpus de twitter con un **%83.05** con el Experimento 5. Para el modelo de edadMujer se alcanzó un **%48.29** con el corpus de twitter utilizando el Experimento 2. Con respecto al modelo edadHombre se alcanzó un **%54.05** con el corpus de blogs en el Experimento 5.

5.2.3. Resultados obtenidos para el corpus en Español

El mejor clasificador para la mayoría de los experimentos de la tabla 5.9 es el de máquinas de soporte vectorial (SMO), para este corpus las mejores características son las del Experimento 5.

Blogs

Características	Tipo de clasificación	Clasificador		
		Naïve Bayes	SMO	IBk
Experimento 1	Por género	**76.13**	72.72	70.45
	Por edad (mujeres)	25	**47.72**	31.81
	Por edad (hombres)	56.81	**63.63**	38.63
Experimento 2	Por género	71.59	**72.72**	69.31
	Por edad (mujeres)	15.9	**43.18**	27.27
	Por edad (hombres)	56.81	**65.9**	40.9
Experimento 4	Por género	67	**72.72**	65.9
	Por edad (mujeres)	22.72	**45.45**	18.18
	Por edad (hombres)	65.9	**61.36**	27.27
Experimento 5	Por género	71.59	**79.54**	72.72
	Por edad (mujeres)	**47.72**	43.18	36.36
	Por edad (hombres)	65.9	**68.18**	47.72

Tabla 5.9: Resultados de la primera aproximación para el corpus de blogs en Español.

De igual forma que para el corpus de socialmedia el mejor clasificador es SMO, con los mejores resultados para todos los experimentos y de nuevo las mejores características son las del Experimento 5.

Socialmedia

Características	Tipo de clasificación	Clasificador		
		Naïve Bayes	SMO	IBk
Experimento 1	Por género	57.94	**59.19**	54.24
	Por edad (mujeres)	29	**33.17**	28.3
	Por edad (hombres)	27.35	**35.22**	27.35
Experimento 2	Por género	56	**59.82**	53.93
	Por edad (mujeres)	24.68	**32.54**	28.14
	Por edad (hombres)	30.97	**35.53**	26.25
Experimento 4	Por género	55.81	**60.61**	54.16
	Por edad (mujeres)	14.62	**33.8**	30.18
	Por edad (hombres)	29.55	**34.43**	30
Experimento 5	Por género	57.38	**59.98**	57.38
	Por edad (mujeres)	36.32	**38.36**	33.33
	Por edad (hombres)	32.38	**40.72**	31.28

Tabla 5.10: Resultados de la primera aproximación para el corpus de socialmedia en Español.

En resumen para el idioma Español, los modelos obtuvieron los mejores resultados para el corpus de blogs con un **%79.54** y un **%68.18** para el género y la edad de los hombres con el Experimento 5 respectivamente. Y un **%47.72** para la edad de las mujeres con el Experimento 1.

5.2.4. Resumen de los mejores resultados

En la siguiente tabla se muestra un resumen con los mejores resultados obtenidos en los diferentes experimentos para cada corpus, se muestra en **negritas** los mejores resultados por idioma.

Características	Tipo de clasificación	Clasificador	Presición
INGLÉS			
Blogs			
Experimento 5	Por género	SMO	68.7
Experimento 3	Por edad (mujeres)	SMO	43.83
Experimento 5	Por edad (hombres)	SMO	**54.05**
Reviews			
Experimento 5	Por género	SMO	65.69
Experimento 5	Por edad (mujeres)	SMO	30.28
Experimento 2	Por edad (hombres)	SMO	29.61
Socialmedia			
Experimento 4	Por género	IBk	52.38
Experimento 4	Por edad (mujeres)	SMO	37.91
Experimento 2	Por edad (hombres)	SMO	36.78
Twitter			
Experimento 5	Por género	SMO	**83.05**
Experimento 2	Por edad (mujeres)	SMO	**48.29**
Experimento 3	Por edad (hombres)	SMO	49.48
ESPAÑOL			
Blogs			
Experimento 5	Por género	SMO	**79.54**
Experimento 1	Por edad (mujeres)	SMO	**47.72**
Experimento 5	Por edad (hombres)	SMO	**68.18**
Socialmedia			
Experimento 4	Por género	SMO	60.61
Experimento 5	Por edad (mujeres)	SMO	38.36
Experimento 5	Por edad (hombres)	SMO	40.72

Tabla 5.11: Resumen de la primera aproximación para ambos idiomas.

El corpus de blogs en Español se comportó mejor que el corpus de blogs en Inglés, superando casi en 10 % los resultados de género y de edad de hom-

bres. Lo importante a destacar es que el mismo conjunto de características son las que mejores resultados dieron para ambos idiomas (Experimento 5).

Para el corpus de socialmedia en Español, los resultados de igual forma superan a los resultados para el corpus de socialmedia en Inglés, pero en este caso las diferencias de presición no superan el 8 %. En cuanto a las características que mejores resultados nos arrojaron para el género son las 254 para ambos idiomas (Experimentos 4 y 5).

5.3. Segunda aproximación

A continuación se muestran los resultados de la segunda aproximación. Los experimentos se hicieron por autor aplicando validación cruzada de 10 pliegues. Se utilizaron los algoritmos de clasificación IBk, SMO y Naïve Bayes sobre el conjunto de características descritas en la sección 4.1.2, además de las probabilidades calculadas como se explica en la sección4.2.

5.3.1. Descripción de los experimentos

El objetivo de estos experimentos, es analizar el comportamiento de esta aproximación sobre el conjunto de datos. Se utiliza el modelo de unigramas y se verifica si dicha combinación incrementa el porcentaje de precision.

A continuación se muestra una descripción de las características que incluye cada experimento:

Experimento 1: Se combinan las carácteristicas del Experimento 4 de la sección 5.2.1 con la probabilidad de unigramas de cada corpus. Total 254 características más el número de unigramas de cada corpus.

Experimento 2: Características del **Experimento 1** combinadas con una bolsa de palabras del texto de cada documento. Total 254 características más el tamaño del vocabulario del corpus más el número unigramas de cada corpus.

Experimento 3: Características del **Experimento 2**, para este caso se utilizó un corpus etiquetado al momento de calcular las probabilidades de

unigramas, hace un total 254 características más el tamaño del vocabulario del corpus más el número unigramas de cada corpus.

Dichos experimentos solo pudieron realizarse sobre los corpus de blogs y reviews en Inglés, y sobre el de blogs en Español, debido a que el tamaño del vocabulario y el número de muestras con las que se contaba era muy grande.

5.3.2. Resultados obtenidos para el corpus en Inglés

En la tabla 5.12 se muestran los resultados de los experimentos para el corpus de blogs, se obtuvo casi un 90 % para la detección del género. El algoritmo de Máquinas de soporte vectorial es el que ofrece los mejores resultados para esta clasificación.

Blogs				
Características	**Tipo de clasificación**	**Clasificador**		
		Naïve Bayes	**SMO**	**IBk**
Experimento 1	Por género	62.58	**85.71**	53.06
	Por edad (mujeres)	38.35	**43.83**	43.83
	Por edad (hombres)	41.89	**47.29**	41.89
Experimento 2	Por género	61.9	**84.35**	55.1
	Por edad (mujeres)	**39.72**	39.72	42.46
	Por edad (hombres)	44.59	**52.7**	40.54
Experimento 3	Por género	61.9	**87.07**	49.65
	Por edad (mujeres)	39.72	**43.83**	42.46
	Por edad (hombres)	**48.64**	45.49	40.54

Tabla 5.12: Resultados de la segunda aproximación para el corpus de blogs en Inglés.

En la tabla 5.13 se muestran los resultados para el corpus de reviews, SMO tuvo el mejor desempeño y se obtuvo casi un 100 % para la detección del género con todos los experimentos. Sin embargo, los resultados para los modelos de edad fueron bajos, ya que no lograron superar el 30 %.

Reviews

Características	Tipo de clasificación	Clasificador		
		Naïve Bayes	SMO	IBk
Experimento 1	Por género	94.27	**99.85**	94.1
	Por edad (mujeres)	27.88	**30.76**	21.49
	Por edad (hombres)	25.1	**26.74**	23.03
Experimento 2	Por género	91.15	**99.87**	79.08
	Por edad (mujeres)	27.83	**30.76**	21.68
	Por edad (hombres)	26.26	**28.13**	24.05
Experimento 3	Por género	91.05	**99.85**	74.46
	Por edad (mujeres)	27.88	**29.9**	22.35
	Por edad (hombres)	25.92	**28.86**	24.19

Tabla 5.13: Resultados de la segunda aproximación para el corpus de reviews en Inglés.

5.3.3. Resultados obtenidos para el corpus en Español

A continuación se muestran los resultados para el idioma Español, se observa que el algoritmo que mejor se comporta es nuevamente SMO y el Experimento con los mejores resultados fue el 3.

Blogs

Características	Tipo de clasificación	Clasificador		
		Naïve Bayes	SMO	IBk
Experimento 1	Por género	72.72	**72.72**	53.4
	Por edad (mujeres)	27.27	**36.36**	34
	Por edad (hombres)	56.81	**59.09**	15.9
Experimento 2	Por género	72.72	**72.72**	53.4
	Por edad (mujeres)	34	**36.36**	34
	Por edad (hombres)	59	**61.36**	15.9
Experimento 3	Por género	72.72	**76.13**	55.68
	Por edad (mujeres)	38.63	**38.63**	29.54
	Por edad (hombres)	65.9	**68.18**	31.81

Tabla 5.14: Resultados de la segunda aproximación para el corpus de blogs en Español.

5.3.4. Resumen de los mejores resultados

En la siguiente tabla se muestra un resumen con los mejores resultados de los experimentos de cada corpus, se muestra en **negritas** los mejores resultados por idioma.

Características	Tipo de clasificación	Clasificador	Presición
INGLÉS			
Blogs			
Experimento 3	Por género	SMO	87.07
Experimento 3	Por edad (mujeres)	SMO	**43.83**
Experimento 2	Por edad (hombres)	SMO	**52.7**
Reviews			
Experimento 2	Por género	SMO	**99.87**
Experimento 2	Por edad (mujeres)	SMO	30.76
Experimento 3	Por edad (hombres)	SMO	28.86
ESPAÑOL			
Blogs			
Experimento 3	Por género	SMO	**76.13**
Experimento 3	Por edad (mujeres)	SMO	**38.63**
Experimento 3	Por edad (hombres)	SMO	**68.18**

Tabla 5.15: Resumen de la segunda aproximación para ambos idiomas.

Si se comparan los resultados de la tabla 5.15 con los de la primera aproximación (tabla 5.11) se puede observar que para la detección del género en los corpus de blogs y reviews, hubo un incremento en la precision de entre 20 % y 30 %.

Sin embargo, para la detección de la edad, las probabilidades de los unigramas no son características representativas, ya que no se nota un cambio favorable en la precision obtenida.

5.4. Tercera aproximación

En las siguientes tablas se muestran los resultados de la tercera aproximación. Estos experimentos fueron realizados con la herramienta pattern de Clips (véase el Capítulo 4.3) usando la implementación del algoritmo de Naïve Bayes.

5.4.1. Descripción de los experimentos

Como ya se describió anteriormente hay diferentes pesos (para una descripción detallada de cada medida, véase la sección 4.3.2) que se pueden utilizar para medir la importancia de cada palabra en un corpus, utilizando esos pesos, fue que se realizaron los siguientes experimentos:

Experimento 1: Se utilizó la frecuencia **F** de cada palabra.

Experimento 2: Se utilizó la frecuencia booleana **F(B)** de cada palabra.

Experimento 3: Se utilizó la frecuencia normalizada del término **TF** de cada palabra.

Experimento 4: Se utilizó la frecuencia normalizada por frecuencia inversa del documento **TF-IDF** de cada palabra.

Se utilizó el algoritmo de Naïve Bayes para todos los experimentos.

5.4.2. Resultados obtenidos para el corpus en Inglés

En la tabla 5.16 se puede observar los resultados para el corpus de blogs en Inglés. Debido a los bajísimos resultados en la clasificación por autor, se tomó la desición de realizar todos lo experimentos solo por documento.

Blogs

Características	**Tipo de clasificación**	Por autor	Por documento
Experimento 1	Por género	26.27	75.93
	Por edad (mujeres)	16.09	52.82
	Por edad (hombres)	16.2	**58.93**
Experimento 2	Por género	29.21	76.39
	Por edad (mujeres)	12.38	**52.98**
	Por edad (hombres)	12.41	56.22
Experimento 3	Por género	32.54	76.33
	Por edad (mujeres)	13.79	50.75
	Por edad (hombres)	11.21	58.2
Experimento 4	Por género	27.73	**76.64**
	Por edad (mujeres)	15.77	54.33
	Por edad (hombres)	17.14	58.89

Tabla 5.16: Resultados de la tercera aproximación para el corpus de blogs en Inglés.

Los resultados para el corpus de reviews (tabla 5.17) fueron mas bajos que los del corpus de blogs tanto en género como en edad.

Reviews

Características	**Tipo de clasificación**	Por documento
Experimento 1	Por género	58.57
	Por edad (mujeres)	27.66
	Por edad (hombres)	25.92
Experimento 2	Por género	58.34
	Por edad (mujeres)	27.77
	Por edad (hombres)	27.63
Experimento 3	Por género	58.69
	Por edad (mujeres)	**28.14**
	Por edad (hombres)	27.55
Experimento 4	Por género	**59.19**
	Por edad (mujeres)	27.25
	Por edad (hombres)	**27.91**

Tabla 5.17: Resultados de la tercera aproximación para el corpus de reviews en Inglés.

Para el corpus de twitter los resultados se muestran en la tabla 5.18 y son mejores que los obtenidos para el corpus de blogs y en general para esta aproximación.

Twitter

Características	**Tipo de clasificación**	Por documento
Experimento 1	Por género	**85.12**
	Por edad (mujeres)	71.41
	Por edad (hombres)	66.52
Experimento 2	Por género	84.97
	Por edad (mujeres)	71.49
	Por edad (hombres)	**66.75**
Experimento 3	Por género	85.02
	Por edad (mujeres)	71.37
	Por edad (hombres)	66.76
Experimento 4	Por género	84.98
	Por edad (mujeres)	**71.61**
	Por edad (hombres)	66.62

Tabla 5.18: Resultados de la tercera aproximación para el corpus de twitter en Inglés.

Para el corpus de socialmedia se obtuvieron resultados muy parecidos a los del corpus de reviews. Estos resultados se pueden observar en la tabla 5.19.

Socialmedia

Características	**Tipo de clasificación**	Por documento
Experimento 1	Por género	**52.89**
	Por edad (mujeres)	33.49
	Por edad (hombres)	38.57
Experimento 2	Por género	55.64
	Por edad (mujeres)	34
	Por edad (hombres)	**38.71**
Experimento 3	Por género	52.88
	Por edad (mujeres)	**34.25**
	Por edad (hombres)	38.66
Experimento 4	Por género	52.88
	Por edad (mujeres)	33.39
	Por edad (hombres)	38.44

Tabla 5.19: Resultados de la tercera aproximación para el corpus de socialmedia en Inglés.

5.4.3. Resultados obtenidos para el corpus en Español

Para el idioma Español los resultados del corpus de blogs se muestan en la tabla 5.20.

Blogs

Características	**Tipo de clasificación**	Por documento
Experimento 1	Por género	71.13
	Por edad (mujeres)	52.67
	Por edad (hombres)	43.69
Experimento 2	Por género	70.98
	Por edad (mujeres)	53.61
	Por edad (hombres)	**51.2**
Experimento 3	Por género	**71.2**
	Por edad (mujeres)	**57.51**
	Por edad (hombres)	49.65
Experimento 4	Por género	70
	Por edad (mujeres)	54.94
	Por edad (hombres)	48.72

Tabla 5.20: Resultados de la tercera aproximación para el corpus de blogs en Español.

Se puede observar que tanto para el corpus de blogs, como para el de socialmedia, los resultados fueron muy similares a los del idioma Inglés.

Socialmedia		
Características	**Tipo de clasificación**	Por documento
Experimento 1	Por género	53.43
	Por edad (mujeres)	25.61
	Por edad (hombres)	26.04
Experimento 2	Por género	**53.56**
	Por edad (mujeres)	**25.72**
	Por edad (hombres)	25.89
Experimento 3	Por género	53.46
	Por edad (mujeres)	25.43
	Por edad (hombres)	26.02
Experimento 4	Por género	53.54
	Por edad (mujeres)	25.54
	Por edad (hombres)	**26.14**

Tabla 5.21: Resultados de la tercera aproximación para el corpus de socialmedia en Español.

5.4.4. Resumen de los mejores resultados

En la tabla 5.22 se muestra un resumen con los mejores resultados de los experimentos de cada corpus, indicando en **negritas** los mejores resultados por idioma.

Si se comparan los resultados de esta aproximación con los de la primera aproximación (tabla 5.11) se puede notar que se superaron los resultados para los corpus en Inglés de blogs y de twitter, tanto para el género como para la edad.

En cuanto a los demás corpus incluyendo los del idioma Español, los resultados quedaron por arriba o por debajo de los resultados de la primera aproximación por pequeñas diferencias en los porcentajes.

5.5. Cuarta aproximación

A continuación se muestra una descripción de los experimentos y por último los resultados de la cuarta aproximación. La clasificación se realizó con el algoritmo máquinas de soporte vectorial (SMO) implementado en weka.

Características	Tipo de clasificación	Presición
INGLÉS		
Blogs		
Experimento 4	Por género	76.64
Experimento 2	Por edad (mujeres)	52.98
Experimento 1	Por edad (hombres)	58.93
Reviews		
Experimento 4	Por género	59.19
Experimento 3	Por edad (mujeres)	28.14
Experimento 4	Por edad (hombres)	27.91
Socialmedia		
Experimento 1	Por género	52.89
Experimento 3	Por edad (mujeres)	34.25
Experimento 2	Por edad (hombres)	38.71
Twitter		
Experimento 1	Por género	**85.12**
Experimento 4	Por edad (mujeres)	**71.61**
Experimento 2	Por edad (hombres)	**66.75**
ESPAÑOL		
Blogs		
Experimento 3	Por género	**71.2**
Experimento 3	Por edad (mujeres)	**57.51**
Experimento 2	Por edad (hombres)	**51.2**
Socialmedia		
Experimento 2	Por género	53.56
Experimento 2	Por edad (mujeres)	25.72
Experimento 4	Por edad (hombres)	26.14

Tabla 5.22: Resumen de la tercera aproximación para ambos idiomas.

5.5.1. Descripción de los experimentos

Para estos experimentos se tomaron varios conjuntos de palabras, para analizar el comportamiento del clasificador, cabe destacar que se realizó el mismo proceso descrito en el capítulo 4.4 para cada clase {female, male} de cada corpus y cada experimento se probó por documento y por autor. Con estos conjuntos de palabras de crearon los modelos para clasificar los documentos por género y por edad, los experimentos se explican en detalle a continuación:

- **Experimento 1**: Se escogieron las 1000 palabras con mayor interconectividad (véase el Capítulo 4.4.3) de cada clase {female, male}.
- **Experimento 2**: Se tomaron todas las palabras del vocabulario de cada clase, excluyendo las que tienen una interconectividad igual a cero.
- **Experimento 3**: Se excluyeron las que tienen una interconectividad igual a cero. Se dividió el total de palabras entre 2 y se tomó mil palabras arriba de la mitad y mil palabras abajo de la mitad, haciendo un total de 2000 palabras por clase.
- **Experimento 4**: Se excluyeron las que tienen una interconectividad igual a cero. Se calculó el promedio de la interconectividad de cada palabra y se tomó mil palabras arriba del promedio y mil palabras abajo del promedio, un total de 2000 palabras por clase.

Los Experimentos 3 y 4 se realizaron con la hipótesis de que las palabras con mediana interconectividad serían más representativas de su clase, ya que hubo menos intersección de las palabras entre las clases, a comparación de los experimentos anteriores.

Por último se realizaron 2 experimentos más, pero ahora especificamente para crear un modelo para calcular la edad de los autores de los documentos. Para esto se crearon 10 grafos adicionales por cada corpus. Como se tienen dos clases para el género {female, male} y 5 clases para la edad {18-24, 25-34, 35-49, 50-64, 65+}, se creó un grafo por cada clase género-edad (female-18-24, female-25-34, etc). Obteniendo como resultado 5 conjuntos de palabras con su respectiva interconectividad por cada género, para entrenar

cada modelo edadHombre y edadMujer (véase 4.4.4) se utilizaron las instancias correspondientes al género del modelo que se entrenó.

- **Experimento 5**: Se escogieron las 1000 palabras con mayor interconectividad de cada clase (female-18-24, female-25-34, etc), haciendo un total de 5000 palabras para cada modelo.
- **Experimento 6**: Se escogieron las 1000 palabras con mayor interconectividad de cada clase como en el experimento anterior, pero se observó que las clases que más se confunden entre ellas son: 25-34, 35-49 y 50-64. Debido a esto se decidió tomar las siguientes mil palabras con mayor interconectividad de estas clases en particular, 1000 palabras para las clases 18-24 y 65 y 2000 palabras para las clases mencionadas anteriormente, haciendo un total de 8000 palabras para cada modelo.

A continuación se muestran los resultados de los experimentos para cada corpus. Debido a que los Experimentos 5 y 6 se diseñaron para calcular la edad, no aplican las pruebas sobre el corpus por género, esto se indica con N/A.

5.5.2. Resultados obtenidos para el corpus en Inglés

En la tabla 5.23, se puede observar que de igual forma que en la tercera aproximación, el modelo se comporta mejor cuando el corpus está organizado por documentos.

Blogs

Tipo de clasificación	Exp 1	Exp 2	Exp 3	Exp 4	Exp 5	Exp 6
Por Autores						
Por género	**66.66**	63.94	55.78	62.58	N/A	N/A
Por edad (mujeres)	28.72	36.96	**42.46**	31.5	32.24	39.72
Por edad (hombres)	44.59	47.29	44.59	**47.29**	41.89	44.59
Por Documentos						
Por género	74.6	**80.76**	70.99	79.09	N/A	N/A
Por edad (mujeres)	58.84	66.89	56.38	60.01	65	**67.58**
Por edad (hombres)	65.86	72.4	61.4	66.26	70.81	**73.36**

Tabla 5.23: Resultados de la cuarta aproximación para el corpus de blogs en Inglés.

Para el corpus de reviews, organizar el corpus por documentos no hace gran diferencia, ya que el número de autores es muy elevado y cada autor

presenta pocas instancias. En este caso por autor tuvo mejor precision en el género y edad de las mujeres.

Reviews

Tipo de clasificación	Exp 1	Exp 2	Exp 3	Exp 4	Exp 5	Exp 6
Por Autores						
Por género	**66.82**	63.36	59.3	66.63	N/A	N/A
Por edad (mujeres)	31.63	30.43	26.77	33.22	**33.89**	32.98
Por edad (hombres)	28.65	29	26	30	30.86	**31.53**
Por Documentos						
Por género	**66.3**	64	60.76	65.97	N/A	N/A
Por edad (mujeres)	**33.05**	31.53	27.8	31	32.97	31.87
Por edad (hombres)	30.85	30.78	27.27	30.53	30.67	**31.63**

Tabla 5.24: Resultados de la cuarta aproximación para el corpus de reviews en Inglés.

Para el corpus de socialmedia, se muestran los resultados en la tabla 5.25. En los campos de la tabla que se muestran en –, significa que no se pudieron ejecutar debido a restricciones de Hardware. Los mejores resultados se obtienen separando el corpus por documentos.

Socialmedia

Tipo de clasificación	Exp 1	Exp 2	Exp 3	Exp 4	Exp 5	Exp 6
Por Autores						
Por género	51.44	**53.2**	47.5	46.65	N/A	N/A
Por edad (mujeres)	30.44	31.7	24.44	24.58	32.96	**33.1**
Por edad (hombres)	33.61	31.2	31.63	30.92	**35.46**	33.75
Por Documentos						
Por género	**62.39**	–	60.9	60.74	N/A	N/A
Por edad (mujeres)	52.69	–	32.02	32.22	55.47	**57.67**
Por edad (hombres)	52.27	–	32.99	32.96	55.74	**56.85**

Tabla 5.25: Resultados de la cuarta aproximación para el corpus de socialmedia en Inglés.

La ventaja de organizar el corpus por autor es que cuando el corpus es muy grande, el número de instancias que se crean en el clasificador es menor al que se crea cuando el corpus está organizado por documentos. Y se pueden ejecutar experimentos, como en este caso para el corpus de twitter (tabla 5.26) muchos de los experimentos por documentos no pudieron ejecutarse, de nuevo por restricciones de hardware.

Twitter

Tipo de clasificación	Exp 1	Exp 2	Exp 3	Exp 4	Exp 5	Exp 6
Por Autores						
Por género	70.76	65.66	60.51	**72.1**	N/A	N/A
Por edad (mujeres)	46	47.1	40.57	39.13	45.65	**51.44**
Por edad (hombres)	38.14	**49.47**	47.36	32.63	41.05	42.1
Por Documentos						
Por género	–	–	–	–	N/A	N/A
Por edad (mujeres)	**61.47**	–	–	–	–	–
Por edad (hombres)	62.24	–	48	57.88	**70.61**	–

Tabla 5.26: Resultados de la cuarta aproximación para el corpus de twitter en Inglés.

Para el corpus de Twitter, todos los resultados por documento superaron a los resultados por autor.

5.5.3. Resultados obtenidos para el corpus en Español

Se muestran en la tabla 5.27 los resultados obtenidos para el corpus de blogs en Español, para los tres modelos se supero en 74 % y para el género casi se alcanzó el 85 %.

Blogs

Tipo de clasificación	Exp 1	Exp 2	Exp 3	Exp 4	Exp 5	Exp 6
Por Autores						
Por género	70.45	**77.27**	70.45	73.86	N/A	N/A
Por edad (mujeres)	34	31.81	47.72	31.81	40.9	**47.72**
Por edad (hombres)	61.36	61.36	54.54	61.36	59	**63.63**
Por Documentos						
Por género	83.7	**84.79**	76.56	78.57	N/A	N/A
Por edad (mujeres)	65.65	73.8	65.17	66.29	73.8	**74.92**
Por edad (hombres)	76.42	**84.24**	66.77	75.45	80.7	83.6

Tabla 5.27: Resultados de la cuarta aproximación para el corpus de blogs en Español.

Socialmedia

Tipo de clasificación	Exp 1	Exp 2	Exp 3	Exp 4	Exp 5	Exp 6
Por Autores						
Por género	**63.67**	–	54.63	55.5	N/A	N/A
Por edad (mujeres)	38.05	40.72	28.93	36.63	**47.64**	46
Por edad (hombres)	32.27	**40.4**	34.74	34.43	35.53	34.27
Por Documentos						
Por género	**59.67**	–	54.75	56.94	N/A	N/A
Por edad (mujeres)	37.74	–	34.71	36.88	40.2	**40.38**
Por edad (hombres)	39.46	–	34.41	35.45	41.24	**41.36**

Tabla 5.28: Resultados de la cuarta aproximación para el corpus de socialmedia en Español.

Contrario a los resultados del corpus de blogs, el de socialmedia no alcanzó el 70 % para ningún modelo. Incluso no se pudo ejecutar el Experimento 2, el cual nos dio los mejores resultados para el caso del corpus de blogs.

5.5.4. Resumen de los mejores resultados

Se puede observar en la tabla 5.29 que los mejores resultados se obtienen con el corpus de blogs, pero en general el corpus de blogs en Español se obtuvieron los mejores resultados tanto para el género como para la edad. Otro detalle importante a resaltar es que para la edad, el Experimento con mejor desempeño para casi todos los corpus fue el número 6 y para el caso del género fueron el 2 y el 1.

Características	Tipo de clasificación	Tipo de organización	Presición
INGLÉS			
Blogs			
Experimento 2	Por género	Por documento	**80.76**
Experimento 6	Por edad (mujeres)	Por documento	**67.58**
Experimento 6	Por edad (hombres)	Por documento	**73.36**
Reviews			
Experimento 1	Por género	Por autor	66.82
Experimento 5	Por edad (mujeres)	Por autor	33.89
Experimento 6	Por edad (hombres)	Por documento	31.63
Socialmedia			
Experimento 1	Por género	Por documento	62.39
Experimento 6	Por edad (mujeres)	Por documento	57.67
Experimento 6	Por edad (hombres)	Por documento	56.85
Twitter			
Experimento 4	Por género	Por autor	72.1
Experimento 1	Por edad (mujeres)	Por documento	61.47
Experimento 5	Por edad (hombres)	Por documento	70.61
ESPAÑOL			
Blogs			
Experimento 2	Por género	Por documento	**84.79**
Experimento 6	Por edad (mujeres)	Por Documento	**74.92**
Experimento 2	Por edad (hombres)	Por Documento	**84.24**
Socialmedia			
Experimento 1	Por género	Por autor	63.67
Experimento 5	Por edad (mujeres)	Por autor	47.64
Experimento 6	Por edad (hombres)	Por documento	41.36

Tabla 5.29: Resumen de la primera aproximación para ambos idiomas.

Capítulo 6

Conclusiones

En este capítulo se presentan las conclusiones finales acerca de los distintos métodos para determinar el perfil de un autor, las propuestas de trabajo a futuro y como punto final las respuestas a las preguntas de investigación.

6.1. Conclusiones finales y trabajo a futuro

En la presente investigación, se cumplieron todos los objetivos generales y específicos definidos en la sección 1.1.2, los cuales se enfocaron en resolver la tarea de identificar el perfil del autor de un documento dado. Dicho perfil esta compuesto por el género y la edad de la persona.

Para cumplir con los objetivos primeramente se realizó un investigación de los trabajos realizados en el área o áreas que se relacionan de alguna manera con esta tarea, y se encontró en dichos trabajos que el método más utilizado es la extracción y representación de las características de los textos mediante vectores. Por lo que en esta investigación se quiso probar la eficacia de dicho método de representación combinado con cuatro diferentes técnicas para la extracción de características de los textos.

En cuanto a los resultados de los corpus en Inglés se obtuvo un 99 % con el modelo basado en unigramas, el de edadMujer y edadHombre superan el 70 % con el modelo de clips y gephi respectivamente. Se pudo observar que en general el de blogs genera los mejores resultados en casi todos los modelos, excepto en el de Clips, en el cual ya se había notado que entre mas instancias o documentos se tienen, mejor se comporta el modelo y este es el caso para el corpus de twitter.

Para los corpus en español, se observa que en este caso el corpus de blogs tiene los mejores en todos los modelos, pero recordemos que solo se contó

con el de blogs y el de socialmedia.

En general, el modelo que mejor resultados obtuvo, fue sin duda el de Gephi, el cual estuvo por arriba del 84 % para detectar el género y la edad de los hombres y por arriba del 74 % para la edad de las mujeres.

Desafortunadamente los corpus proporcionados para esta tarea están muy des balanceados, es decir, la distribución de las instancias que se tienen es muy dispareja y eso dificulta el proceso de entrenamiento de los modelos, y causa que clases como la edad 18-25 y 65+ (que son las que menos instancias tienen para todos los corpus), se confundan fácilmente con otras clases, ya que no se tiene suficiente evidencia como para extraer características relevantes. También el tamaño de los corpus varia mucho, ya que corpus como el de socialmedia es difícil de manejar si no se cuenta con el Hardware necesario, porque que el número y tamaño de los documentos es muy grande.

En las siguientes figuras, se muestra una comparación de los resultados obtenidos en esta investigación, con los resultados reportados dentro de la bibliografia.

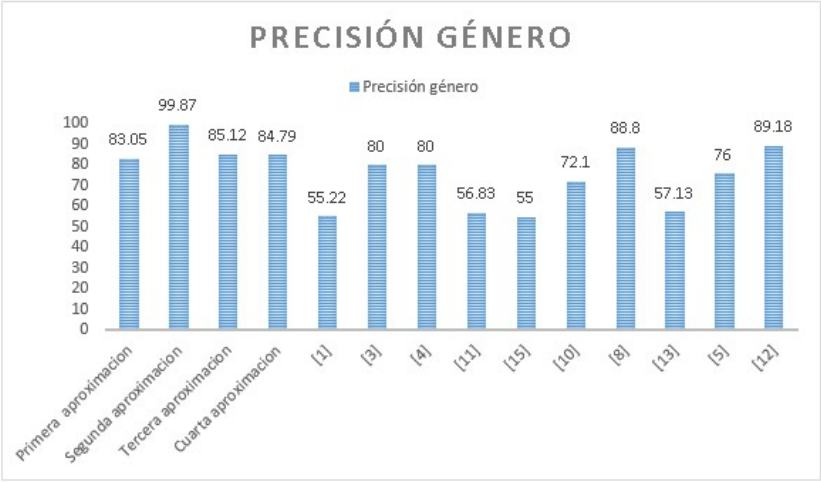

Figura 6.1: Comparación de resultados de género.

En la figura 6.1, se puede observar que los resultados obtenidos en las 4 aproximaciones de esta investigación superan a la mayoría de los resultados

reportados en los trabajos analizados. Recordemos que el resultado obtenido en la segunda aproximación fue con el corpus de twitter, el cual tiene un número de instancias muy grandes y por ende el vocabulario también es muy grande, para esa aproximación se calculó la probabilidad de las palabras y siendo el vocabulario tan grande, el número de palabras que no aparecieron el en vocabulario (y por lo tanto no se calculo su probabilidad) fue muy pequeño. Esto influye directamente en el resultado, entre menos palabras se tengan con probabilidad cero (o sin probabilidad calculada), más evidencia tiene el clasificador para entrenar el modelo.

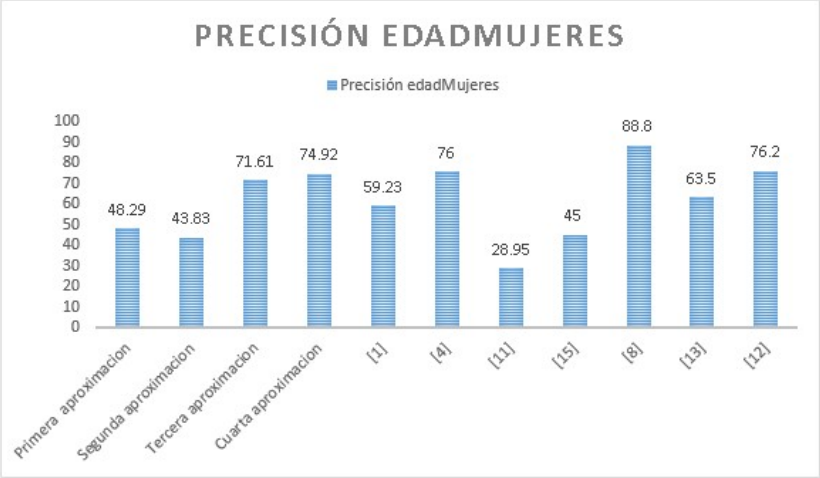

Figura 6.2: Comparación de resultados de edad de mujeres.

También se compararon los resultados de los modelos de edad de mujeres en la figura 6.2 y podemos concluir que aunque la primera y segunda aproximación se encuentran por encima del resultado más bajo que se reporta (28 %), los resultados son bajos, ya que el promedio supera al 60 %. Se considera que los resultados de la tercera y cuarta aproximación son buenos, ya que las pruebas se hicieron con un corpus desbalanceado a diferencia del 88 % obtenido en la bibliografía, que se reporta que se obtuvo con un corpus balanceado.

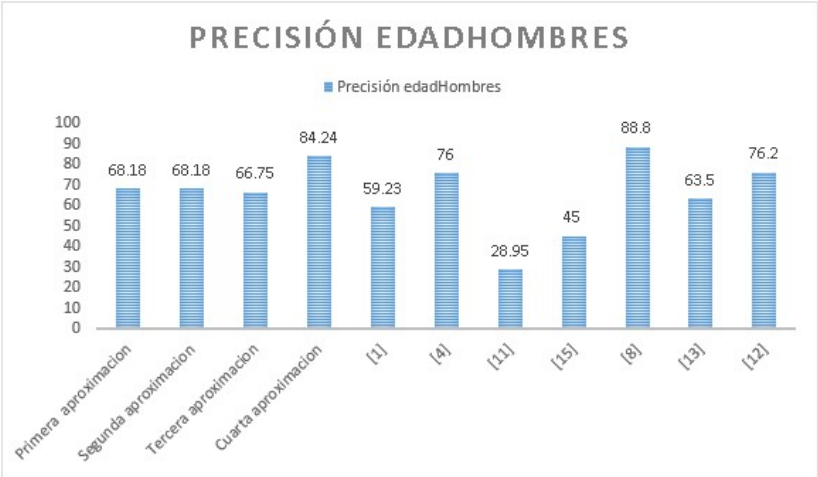

Figura 6.3: Comparación de resultados de edad de hombres.

Por último en la figura 6.3 se muestran los resultados para los modelos de edad de hombres, los cuales se encuentran por encima del 65 %. Se puede observar que el resultado de la cuarta aproximación supera el 80 % y se encuentra muy cerca del 88 % reportado sobre corpus balanceados. Para los modelos de edad, se puede concluir que se obtuvieron esos resultados gracias al análisis de las palabras que se hizo en el Experimento 6 de la sección 5.5.1, en el cual se buscó hacer un balance entre las 5 diferentes clases, por medio del número palabras representativas de cada una, elevando el número de palabras si la clase contaba con pocos documentos o muestras.

En general la metodología basada en grafos brinda un resultado estable que supera el 80 % en los modelos de género y edad de hombres y el 70 % para el modelo de edad de mujeres, esto es debido a que se crearon y analizaron los grafos por corpus y de cada corpus por grupo de edades, esto nos arroja un análisis mas detallado por categorías el cual da muy buenos resultados.

6.2. Trabajo a futuro

Como trabajo futuro se podrían considerar los siguientes puntos:

- La extracción de nuevas características que no hayan sido consideradas dentro de esta investigación.

- Probar las metodologías creadas en esta investigación con corpus diferentes en Inglés y Español.
- Probar las metodologías creadas en esta investigación con corpus en lenguajes diferentes al Inglés y Español.
- Probar las metodologías creadas en esta investigación con los corpus en Español que faltaron (Twitter y Reviews).
- Ejecutar los experimentos que no se pudieron realizar por restricciones de Hardware.
- Poder identificar otros aspectos del perfil de un usuario, como pueden ser el lenguaje nativo o diferentes aspectos de la personalidad.

6.3. Respuestas a las preguntas de investigación

A continuación se tienen las preguntas de investigación de este trabajo, con sus respectivas respuestas:

1. ¿Qué características son las más propicias para detectar el perfil de un autor?

En general utilizar las palabras como características, ya sea por probabilidades como en el modelo de unigramas, como por conteos de las más relevantes como en el modelo de gephi, es lo más adecuado para detectar tanto la edad como el género de un autor. Pero ya vimos que esto va a depender del las características del corpus con el que se este trabajando, para esta investigación y para el corpus de blogs el mejor modelo fue el de gephi.

2.¿Qué características son las más propicias de acuerdo al tipo de corpus?

Se puede observar que para todos los corpus, ya sea en Inglés o en español, las características más adecuadas parecen ser constantes entre ellos, excepto por el corpus de twitter, que es el que se observa que tiene un comportamiento diferente, y parece que las características que funcionan bien para los demás, no siempre funcionan para este corpus. Se cree que esto es debido al tamaño de los documentos y a la variedad de temas que se abordan. Pero de nuevo las

palabras como características siguen siendo propicias para todos los corpus, lo que se tiene que encontrar es qué representación funciona para cada corpus.

3.¿Qué tipo de clasificadores son los más adecuados para esta tarea en particular?

El algoritmo de máquinas de soporte vectorial (SMO) fue el que mejor se comporto para esta tarea, con los modelos aquí desarrollados.

4. ¿El comportamiento de los modelos es similar para ambos idiomas?

Si.

Apéndice A.

En este capítulo se muestran las herramientas utilizadas en esta investigación para la identificación del perfil de un autor.

6.4. Herramientas utilizadas y desarrolladas

En la Tabla 6.1 se presentan las distintas herramientas y paquetes utilizados en cada una de las aproximaciones realizadas para resolver el problema de atribución de autoría.

		Aproximación 1 y 2	
Proceso	**Herramienta**	**Paquete**	**Implementación**
Etiquetado	Clips	parse	Usado para obtener la etiqueta gramatical de las palabras y etiquetar corpus.
Bolsa de palabras	Weka	weka.filters.unsupervised.attribute.StringToWordVector	Usadoa para crear bolsa de palabras.
Clasificación	Weka	weka.classifiers.bayes.NaiveBayes weka.classifiers.functions.SMO weka.classifiers.lazy.IBk	Usado para clasificar
		Aproximación 2	
Proceso	**Herramienta**	**Implementación**	
Calcular probabilidades	Matlab	Implementado para crear el modelo de unigramas.	
		Aproximación 3	
Proceso	**Herramienta**	**Paquete**	**Implementación**
Clasificación	Clips	pattern.vector.NB	Usado para clasificar (algoritmo Naïve Bayes).
Bigramas	Clips	pattern.text.en.ngrams	Usado para calcular bigramas de una oración.
		Aproximación 4	
Proceso	**Herramienta**	**Paquete**	**Implementación**
Creación	NetworkX	nx.DiGraph	Usado para la creación del grafo.
Análisis	Gephi	Metrics	Usado para el análisis del grafo.

Tabla 6.1: Herramientas usadas en las aproximaciónes presentadas.

Bibliografía

[1] Aleman, Y., Loya, N., Vilariño, D.: Two methodologies applied to the author. PAN 2013 (2014)

[2] Argamon, S., Koppel, M., J., P., Schler, J.: Automatically profiling the author of an anonymous text. Communications of the ACM - Inspiring Women in Computing **52**(2) (2009) 119–123

[3] Koppel, M., Argamon, S., A., S.: Automatically categorizing written texts by author gender. Literary and Linguistic Computing **17**(4) (2002) 401–412

[4] Schler, J., Koppel, M., S., A., J., P.: Effects of age and gender on blogging. In: Proceedings of the AAAI Spring Symposium on Computational. (2006)

[5] Burger, J.D., Henderson, J., Kim, G., G., Z.: Discriminating gender on twitter. In: EMNLP '11 Proceedings of the Conference on Empirical Methods in Natural Language Processing. (2011) 1301–1309

[6] Nguyen, D., Gravel, R., Trieschnigg, D., Meder, T.: "how old do you think i am?"; a study of language and age in twitter. In: Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media. (2013)

[7] Nguyen, D., Smith, N., Rosé, C.: Author age prediction from text using linear regression. In: LaTeCH '11 Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. (2011) 115–123

[8] Peersman, C., Daelemans, W., Vaerenbergh, L.V.: Predicting age and gender in online social networks. In: SMUC '11 Proceedings of the 3rd international workshop on Search and mining user-generated contents. (2011) 37–44

[9] Estival, D., Gaustad, T., Pham, S.B., Radford, W., Hutchinson, B.: Author profiling for english emails. In: Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics. (2007) 263–272

[10] Zhang, C., Zhang, P.: Predicting gender from blog posts. http://people.cs.umass.edu/ pyzhang/course/genderClassify.pdf (2010)

[11] Gopal, P., Banerjee, S., Das, D.: Automatic author profiling based on linguistic and stylistic features. In: Proceedings of the 9th PAN at CLEF Conference. (2013)

[12] Goswami, S., Sarkar, S., Rustagi, M., Meder, T.: Stylometric analysis of bloggers age and gender. In: Proceedings of the Third International ICWSM Conference. (2013)

[13] Rangel, F., Rosso, P.: Use of language and author profiling: Identification of gender and age. http://users.dsic.upv.es/ prosso/resources/RangelRosso_NLPCS13.pdf (2009)

[14] Paranyushkin, D.: Identifying the pathways for meaning circulation using text network analysis. Nodus Labs (2011)

[15] Yan, X., Yu, P.S., Han, J.: Graph indexing: A frequent structure-based approach. In: SIGMOD '04 Proceedings of the 2004 ACM SIGMOD international conference on Management of data. (2004) 335–346

[16] Krahmer, E., Verleg, A., Erk, S.: Graph-based generation of referring. In: Computational Linguistics archive. (2003) 53–72

[17] Cook, D., Manocha, N., Holder, L.B.: Using a graph-based data mining system to perform web search|. International Journal of Pattern Recognition and Artificial Intelligence **17**(705) (2003)

[18] Cristina, N., Gabriel, N.: Bestcut: a graph algorithm for coreference resolution. In: EMNLP '06 Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. (2006) 275–283

[19] Jie, C., Michael, S.: End-to-end coreference resolution via hypergraph partitioning. In: COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics. (2010) 143–151

[20] Vincent, N.: Graph-cut-based anaphoricity determination for coreference resolution. In: NAACL '09 Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. (2009) 575–583

[21] Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly, Beijing (2009)

[22] Dai, W., Xue, G.R., Yang, Q., Yu, Y.: Transferring naive bayes classifiers for text classification. In: AAAI. 540–545

[23] Banchs, R.E.: Text mining with MATLAB. Springer, Springer New York Heidelberg Dordrecht London (2013)