Identificación de Usos Medicinales de Plantas utilizando Información Sintáctica y Semántica

Por:

Oscar Pérez Sánchez

Tesis sometida como requerimiento parcial para obtener el grado de

Maestro en Ciencias, en el área de Ciencias Computacionales

En el

Instituto Nacional de Astrofísica, Óptica y Electrónica

Diciembre, 2017

Tonantzintla, Puebla

Supervisores:

Dr. Manuel Montes y Gómez, INAOE

Dr. Luis Villaseñor Pineda, INAOE

A mi Familia

Gracias por todo el apoyo recibido.

A mis profesores

Gracias por sus enseñanzas.

Índice general

Índice de figuras

Índice de tablas

Agradecimientos

Agradezco a mi familia por todo el apoyo que me dieron para poder continuar con mis estudios.

A mis profesores que me guiaron estos 2 años de estudio.

A mis compañeros y amigos que conocí a lo largo de este proceso.

A CONACyT por a verme otorgado una beca para continuar con estos estudios.

Resumen

En México y en todo el mundo se han utilizado a las plantas para combatir enfermedades y malestares. En estos días es posible encontrar esta información gracias al crecimiento de Internet. Al encontrarse la información en forma de texto ¿es posible identificar automáticamente oraciones que describan un uso medicinal mediante técnicas de procesamiento del lenguaje natural (PLN)?. El principal reto a superar es encontrar la forma de relacionar las oraciones que describan un uso medicinal de aquellas que no lo hacen. Utilizando técnicas de PLN se sugiere explotar información sintáctica y semántica extraída de la fuente original para descubrir relaciones que no se detectan de manera superficial.

Ya que el propósito es identificar un uso medicinal que puede ser encontrado en oraciones que componen al texto, la tarea puede ser vista como una tarea de clasificación de textos cortos. La principal característica de esta tarea consiste en trabajar con pequeñas porciones de texto, en este caso oraciones que no superan las 30 palabras. El principal enfoque consiste en enriquecer la poca información disponible con información que permita descubrir relaciones entre las oraciones que no pueden ser detectadas con la información original.

Por el motivo anterior en esta tesis se aborda la tarea de identificación automática de usos medicinales de plantas utilizando información sintáctica y semántica. Se propone un método que obtenga información sintáctica y semántica de las oraciones para poder relacionar aquellas oraciones que describan un uso medicinal

de aquellas que no lo hacen.

La información sintáctica comúnmente es utilizada en tareas de estilo tal como la identificación de un autor por sus documentos escritos, En esta tesis se considera que la estructura de una oración que describe un uso medicinal puede aportar información que permita diferenciarla de oraciones donde no lo hacen, para ello se utilizaron trigramas de etiquetas de parte del habla para identificar patrones utilizados en las oraciones que describen usos, especialmente los medicinales.

Por otro lado la información semántica se ha utilizado en tareas relacionadas a categorías, como diferenciar entre noticias de deportes o finanzas. En este caso las oraciones de interés pertenecen al dominio medico, por lo cual el enriquecer las oraciones con palabras relacionadas a este dominio puede ser útil para esta clasificación. Para obtener la información semántica se hizo uso del recurso semántico BabelNet con el cual se busca relacionar las palabras de las oraciones mediante la generalización a su hiperónimo directo.

Se realizaron experimentos con cada una de las representaciones por separado y mediante combinaciones entre éstas. Los resultados obtenidos indican que el anexar información de tipo semántico aporta información útil, que combinada con la información léxica obtiene resultados superiores que cada tipo de información por separado.

Se realizaron otros experimentos, el primero se realizó con la idea de observar la cantidad mínima de oraciones que pueden componer al conjunto de entrenamiento. Para este experimento los resultados obtenidos indican que los conjuntos de entrenamiento pueden ser reducidos hasta utilizar solo el 6 % (alrededor de 120 oraciones) utilizando solo información semántica y un 12 % (alrededor de 250 oraciones) para la representación que consiste en la combinación de información léxica y semántica.

El último experimento consistió en aplicar el método propuesto para la clasificación de oraciones que pertenezcan a una clase diferente a la medicinal, teniendo

como objetivo la generalización del método. Para ello se utilizaron las oraciones que componen la clase negativa llamada "No Medicinal" la cual está compuesta por oraciones de 3 tipos diferentes: Otros usos (usos diferentes al medicinal), Localización (información sobre el lugar u origen de la planta) y Descripción (información en general acerca de una planta).

Los resultados obtenidos demuestran que se puede utilizar este método para diferentes contextos o dominios y que no depende directamente de la temática que se esté abordando.

Abstract

In Mexico and all over the world, plants have been used to treat diseases and discomforts. In these days it is possible to find information related to medicinal plants thanks to the growth of the internet. By finding this information in the form of text, we may ask whether it is possible to automatically identify sentences that describe a medicinal use using natural language processing techniques (NLP)?. The main issue is to find the way to relate the sentences that describe a medicinal use of those that do not. Using NLP techniques will exploit the syntactic and semantic information extracted from the original source to discover relations that are not detected superficially.

Since the purpose is to identify a medicinal use that can be found in the sentences that compose the text, the task can be seen as a task of short texts classification. The main characteristic of these tasks is work with small portions of text, in this case sentences that do not exceed 30 words. The main approach is to enrich the few information available with information that allows to discover relations between sentences that can not be detected with the original data.

For the previous reason, this thesis addresses the task of automatic identification of medicinal uses of plants using syntactic and semantic information. We propose a method that obtains syntactic and semantic information of the sentences to relate those that describe a medicinal use.

The syntactic information is commonly used in stylistic tasks such as the identification of an author by his written documents, for this task it is considered that the structure of a sentence that describes a medicinal use can provide information that allows to differentiate it from another sentences. part of the speech trigrams were used to identify patterns used in sentences that describe uses, especially medicinal ones.

On the other hand the semantic information has been used in tasks related to categories, like differentiating between sports or political news. In this case the sentences of interest belong to the medical domain, so enriching the sentences with words related to this domain may be useful for this classification. In order to obtain the semantic information, the BabelNet semantic resource was used, with this, we want to relate words of the sentences by generalization to their direct hyperonym.

Experiments were performed with each of the representations separately and by combinations of these. The results obtained indicate that the addition of semantic information provides useful information, which combined with lexical information achieves higher results than each type of information separately.

Other experiments were conducted, the first was done with the idea of observing the minimum number of sentences that can compose the training set. For this experiment the obtained results indicate that the training set can be reduced to only use 6 % (about 120 sentences) using only semantic information and 12 % (about 250 sentences) for the representation that consists of the combination of lexical and semantic information.

The last experiment consisted in applying the proposed method for the classification of sentences belonging to a different class than the medicinal one, aiming at the generalization of the method. For this, the sentences that compose the negative class called "No Medicinal"were used, which is composed of sentences of 3 different types: Other uses (non-medicinal uses), Location (information about the place or

origin of the plant) And Description (general information about a plant).

The results obtained demonstrate that this method can be used for different contexts and that does not depend directly of the domain that is being addressed.

Capítulo 1

Introducción

Con el avance de la tecnología se ha incrementado la cantidad de información que se tiene disponible en todos los dominios de ciencia y tecnología; el dominio botánico no es la excepción, cada vez es más frecuente encontrar información acerca de plantas tales como: sus características, lugar de origen, historia, usos, etc. [Thessen et al., 2012].

Gracias a las diferentes propiedades y características de las plantas se pueden utilizar de diferentes maneras, ya sea en el ámbito medicinal, industrial, culinario, cosmético, etc. En el ámbito medicinal se han utilizado remedios medicinales de plantas desde hace mucho tiempo y en todo el mundo. Esta información se ha conservado en libros y mediante el traspaso de conocimiento entre generaciones. Ahora esta información se encuentra disponible en Internet a través de diversos sitios web dedicados a la recopilación de información de plantas o botánica en general.

Este trabajo se centra en la clasificación de oraciones donde se exprese el uso medicinal de una planta, para ello se hará uso de técnicas de Procesamiento de Lenguaje Natural (PLN) ya que esta área se dedica a desarrollar y utilizar métodos para el procesamiento de información oral y escrita. La importancia de desarrollar

un método de clasificación para este dominio reside en identificar usos potenciales y propuestas de nuevos medicamentos basados en plantas.

1.1. Problemática

En México y muchas partes del mundo se han utilizado remedios y medicamentos a partir de plantas por mucho tiempo. Este conocimiento ha sido transferido hasta la actualidad por medio de libros y a través de generaciones. Se ha comprobado científicamente las propiedades medicinales de algunas plantas y se sigue investigando el de otras. Por otro lado, socialmente este conocimiento se ha aceptado mediante la experiencia, con el paso del tiempo las personas han probado diferentes tratamientos domésticos, algunos de éstos utilizando plantas como fuente. Por medio de la experiencia muchas personas han aliviado dolencias, malestares y enfermedades.

Con el crecimiento de Internet este conocimiento puede ser adquirido por todo el mundo gracias a sitios web que se especializan en información de plantas, en especial las que tienen un uso medicinal. La mayoría de esta información se encuentra en forma textual, por lo que utilizar técnicas y métodos del área de PLN es la opción adecuada ya que en esta área se estudian diferentes métodos para la clasificación de información textual .

En este trabajo se plantean las siguientes preguntas: ¿Mediante técnicas de PLN se puede clasificar oraciones de uso medicinal de aquellas que no lo son?, ¿la información de tipo sintáctica y semántica es relevante para esta tarea?, ¿qué tipo de información sintáctica y semántica puede ser utilizada?

1.2. Metodología Propuesta

La solución propuesta se basa en enriquecer la representación de las oraciones con información ya sea de tipo sintáctico y/o semántico o la combinación de ambas. Se espera que con la ayuda de este tipo de información se pueda distinguir las oraciones que hablan de un uso medicinal de aquellas que no lo hacen. La propuesta se divide de la siguiente manera:

- Creación del conjunto de datos. Al trabajar en el idioma español y al no haber recursos disponibles para esta tarea se debe de construir la colección de oraciones. Para ello se obtendrán oraciones mediante la consulta de varios sitios web dedicados a la recopilación de información relacionada a plantas medicinales, sus usos y a plantas en general.
- Identificación de la información sintáctica. Mediante la representación de las oraciones por su categoría sintáctica, se busca generalizar combinaciones de palabras que sean comunes para la descripción de usos medicinales de plantas.
- Identificación de la información semántica. Con el uso de recursos semánticos se obtendrán palabras que estén relacionadas con la colección de datos. Estas relaciones pueden ser: sinónimos, hiperónimos o hipónimos.
- Clasificación de las oraciones. La clasificación se realizará utilizando cada tipo de información por separado y la combinación de éstas. Se espera obtener mejores resultados mediante la combinación de los tipos de información.

1.3. Motivación

Al poder clasificar automáticamente oraciones que describan el uso de plantas medicinales podemos reunir evidencia del empleo de plantas para el tratamiento de

enfermedades o dolencias. Esta información puede ser útil para estudios posteriores en los cuales se busque conocer el empleo más común que se le da a una planta en particular, esto puede generar oportunidades de comercializar productos que estén relacionados con el uso de plantas medicinales.

Otra utilidad consiste en tener conocimiento previo al realizar estudios de laboratorio, teniendo evidencia del uso de ciertas plantas para aliviar enfermedades o dolencias.

Esta información también puede ser de gran ayuda para realizar catálogos de plantas medicinales. Se puede utilizar esta información para poder llevar un control sobre los usos más comunes que se les da a las plantas en las diferentes regiones del país.

1.4. Objetivos

1.4.1. Objetivo general

El objetivo de este trabajo es identificar oraciones donde se especifique el uso de plantas medicinales mediante la propuesta e implementación de un método de clasificación que utilice información léxica, sintáctica y semántica.

1.4.2. Objetivos específicos

Los objetivos específicos de este trabajo son los siguientes:

- Creación del conjunto de datos. Mediante la recolección de oraciones que mencionen a alguna planta,se utilizaran sitios web especializados en este dominio.
- Evaluación de la información sintáctica aplicada a esta tarea.

- Evaluación de la información semántica aplicada a esta tarea.
- Clasificación de las oraciones considerando la información sintáctica, semántica y la combinación de ambas.

1.5. Organización de la tesis

La tesis está organizada de la siguiente manera.

En el capítulo 2 se describen los conceptos que son relevantes a esta investigación y que son necesarios para comprender la tarea y la solución propuesta.

En el capítulo 3 se presentan los trabajos relacionados con esta investigación y a los conceptos y técnicas utilizadas.

En el capítulo 4 se describe detalladamente la metodología utilizada en este trabajo.

En el capítulo 5 se plantean los experimentos realizados utilizando la información léxica, sintáctica y semántica además de presentar los resultados obtenidos.

En el capítulo 6 se presentan las conclusiones y las pautas a seguir para el trabajo a futuro.

Capítulo 2

Marco Teórico

En este capítulo se introducen los conceptos necesarios para comprender este trabajo de investigación. Inicialmente se describe el proceso de clasificación de texto utilizado. Posteriormente se presentarán conceptos relacionados a las diferentes representaciones utilizadas para manejar la información de tipo textual.

2.1. Clasificación de texto

La clasificación de texto es el proceso de separar documentos en categorías predefinidas con anterioridad. Para realizar esto los documentos de texto son representados mediante características que suelen ser subconjuntos de palabras que contienen la información más importante acerca del contenido del documento.

La clasificación de documentos tiene muchas aplicaciones hoy en día, tales como: filtrado de e-mail, clasificación de noticias, atribución de autoría, detección de plagio, etc.

Para realizar este procedimiento de clasificación se debe seguir cierto proceso el cual se describe a continuación:

- Creación o adquisición del conjunto de datos. Como primer paso se debe analizar el tipo de información con la que se va a trabajar, esta información debe representar las diferentes categorías o clases a las cuales se asignaran para su procesamiento.

- Realizar algún tipo de pre-procesamiento. Esto puede ser opcional si los datos se encuentran con el formato más adecuado para su clasificación y dependen de la tarea a realizar. Para los datos de tipo textual los pre-procesamientos usuales son:

    - Conversión a minúsculas o mayúsculas. Los datos originales pueden estar escritos con una combinación de mayúsculas y minúsculas, lo que puede causar errores al comparar palabras. Por lo que se recomienda que todos los datos se encuentren en minúsculas o mayúsculas para evitar estos errores.

    - Eliminación de signos de puntuación. Dependiendo la tarea a realizar, los signos de puntuación pueden ser eliminados o no de los datos.

    - Substitución o eliminación de información no deseada. Cuando la información proviene de Internet, ésta puede venir acompañada por etiquetas HTML o metadatos. Este tipo de información puede no ser útil en ese estado por lo que se debe eliminar o substituir por un atributo que la generalice.

    - Lematizado o truncamiento de las palabras. El lematizado consiste en representar a las palabras por su raíz, por lo tanto, se debe eliminar todo tipo de conjugación para poder abarcar variantes de la conjugación con un solo atributo. El identificar la raíz puede ser un proceso complicado por lo que se puede aplicar un truncamiento que consiste en eliminar cierta cantidad de caracteres de las palabras buscando generalizarlas.

    - Eliminación de Palabras vacías. Una palabra vacía es aquella que no apor-

ta información categórica al aparecer con una alta frecuencia en todos los documentos. Estas palabras pertenecen a las siguientes categorías de palabras: artículos, pronombres, preposiciones, etc.

- Construcción de la representación de la información. En su representación original el texto puede ser difícil de manejar y limita las operaciones que se pueden realizar con el. Debido a esto se debe de realizar una transformación a una representación que sea mas adecuada para su procesamiento. Una de estas representaciones es el llamado modelo de espacio vectorial el cual se tratara en la sección 2.1.1.

- Métodos de clasificación. Dependiendo de la representación de la información, dimensión de los atributos y naturaleza de la información se pueden utilizar distintos métodos de clasificación para obtener los mejores resultados posibles.

Este proceso se realiza en dos fases, la de entrenamiento y la de prueba. En la primera fase como su nombre lo indica se entrena al método de clasificación con la mayor parte de los datos, para que el clasificador puede caracterizar las distintas clases provistas mediante etiquetas asignadas a los datos. Lo que se desea en esta fase, es que el clasificador pueda identificar que atributos son importantes para cada clase. Una vez que el modelo está listo, se inicia la segunda fase donde se le proporciona la información de prueba. Esta información debe de ser nueva para el clasificador, es decir que no haya sido proporcionada en la parte de entrenamiento. Con esta nueva información se comprueba la efectividad del modelo ante nuevos datos. Como resultado nos devuelve la información asignada a una de las clases proporcionadas en el entrenamiento, para conocer la efectividad del clasificador se proporciona la asignación correcta de las clases del conjunto de prueba para poder comparar los resultados devueltos por el clasificador como se verá más adelante.

En la clasificación de texto una de las representaciones más usadas es el modelo de espacio vectorial. Es de las primeras representaciones en utilizarse y hasta el día de

hoy es una de las más empleadas por obtener resultados satisfactorios en la mayoría de las tareas de PLN.

2.1.1. Modelo de Espacio Vectorial

El modelo de espacio vectorial es un modelo algebraico para representar documentos de texto como vectores de términos donde cada dimensión corresponde a un término en particular. Esto se puede visualizar como una matriz la cual es llamada matriz de termino-documento como se muestra en la figura 2.1.

En primer lugar se debe obtener el diccionario de la colección de documentos, el cual se construye mediante la lista de palabras únicas en toda la colección. Cada una de las palabras representa una columna en la matriz, mientras que cada documento de la colección es representado como una fila.

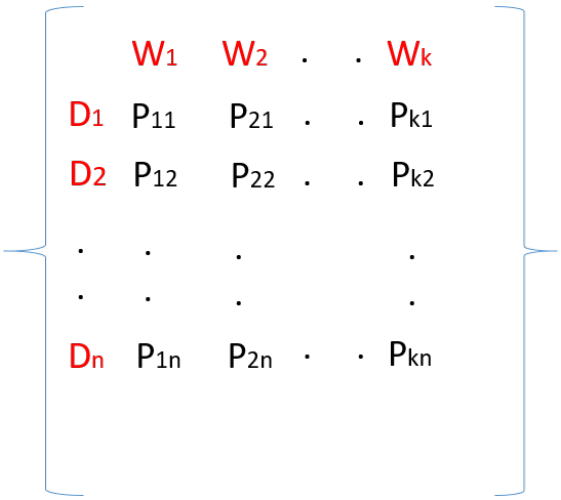

Figura 2.1: Representación de los documentos de una colección en el modelo vectorial.

El valor de P_{11} indica el valor de la palabra W_1 en el documento D_1, el valor de P_{21} indica el valor de la palabra W_2 para el mismo documento y así para todas las palabras hasta W_k. De esta forma se evalúan todas las palabras de la colección, se encuentren o no en el documento D_1.

Al tomar cada una de las filas de la matriz por separado se forma un vector por cada documento, estos vectores pueden ser evaluados con diferentes medidas de distancia. En la figura 2.2 se muestran 3 vectores evaluados mediante la medida de similitud del coseno la cual consiste en calcular el angulo del coseno entre los vectores, si el ángulo es corto los vectores son similares mientras que si el ángulo es grande indica que los vectores son diferentes.

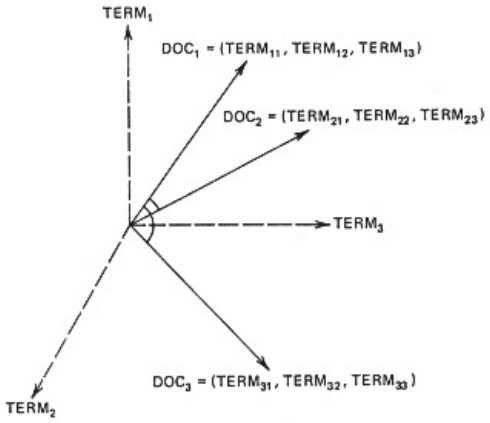

Figura 2.2: Representación gráfica del modelo de espacio vectorial.

Pesado de Términos

El pesado de términos (pt) para el modelo vectorial se basa en la frecuencia de los términos en el documento y la frecuencia de los términos en la colección de documentos. Los pesados más usados son:

- Binario. Donde $pt(t, d) = 1$ si el término (t) está en el documento (d) y 0 si no lo está.
- Frecuencia del término. $pt(t, d) = f(t, d)$ se contabiliza la frecuencia del término en el documento y ese valor es el asignado.
- El pesado TF/IDF. Consiste en dividir la frecuencia del termino en el docu-

mento ($TF = f(t,d)$) con la frecuencia inversa del termino en la colección ($IDF = \frac{f(t,d)}{|C|}$) ($|C|$ es el número de veces que el término aparece en toda la colección). Mediante este pesado se castigan aquellos términos que son muy comunes en todos los documentos y se eligen términos que distingan a los documentos entre sí.

Una vez obtenida la representación de los datos, se pueden utilizar diferentes algoritmos de clasificación buscando separar los documentos mediante la comparación de sus vectores y agrupando los vectores que son similares. De esta manera los documentos serán asignados a la categoría correcta a la que pertenecen.

Para afrontar estas deficiencias del modelo, se deben de analizar las oraciones en busca de otro tipo de información. De esta manera se busca enriquecer la información original para poder realizar una clasificación más acertada.

2.1.2. Métodos de Clasificación

En la literatura se pueden encontrar múltiples algoritmos de clasificación para abordar tareas relacionadas a PLN. Dependiendo de la representación y cantidad de los atributos de los documentos, el desempeño puede ser variable entre algoritmos. A continuación, se introducen algunos de ellos.

Vecinos más Cercanos (KNN)

El clasificador de vecinos más cercanos ha sido utilizado comúnmente en tareas de clasificación textual[Sebastiani, 2002] debido a su efectividad. En este clasificador, para decidir si el documento d_i pertenece a la clase C_l, se calcula la similitud $Sim(d_i, d_j)$ o la disimilitud $Diss(d_i, d_j)$ para todos los documentos d_j en el conjunto de entrenamiento.

Los k vecinos (documentos) más similares son seleccionados. La proporción de vecinos con la misma clase puede tomarse como un estimador para la probabilidad de la clase. De esta manera la clase con la más alta proporción es asignada al documento d_i.

El algoritmo tiene dos parámetros (k y la medida de similitud) los cuales decidirán el desempeño del clasificador y son determinados empíricamente. Sin embargo, el valor óptimo de k puede ser determinado mediante validación cruzada con un conjunto de entrenamiento adicional [Hotho et al., 2005]. En la figura 2.3 se muestra un ejemplo en el cual se utilizan 3 vecinos más cercanos para clasificar un elemento nuevo, el cual se clasifica como blanco al tener una cantidad mayor de vecinos más cercanos de ese color.

La mayor desventaja de este clasificador es el esfuerzo computacional durante la clasificación, ya que la medida de similitud debe ser calculada por cada uno de los documentos de prueba a todos los documentos del conjunto de entrenamiento.

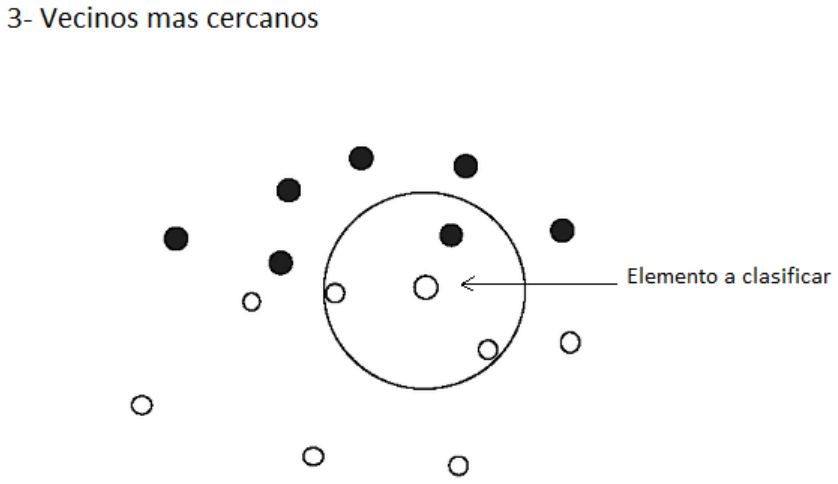

Figura 2.3: Representación de KNN con k = 3

Naïve Bayes

El clasificador Naïve Bayes es el más simple de los clasificadores probabilísticos usado para la clasificación de documentos[Rigutini and Maggini, 2004]. El clasificador estima la probabilidad de un documento d_i de pertenecer a la $Clase_k$.

$$P(C_k|d_i) \tag{2.1}$$

La salida del clasificador es la probabilidad de que el documento pertenezca a cada clase y es un vector de $|C|$ elementos. Para la clasificación se elige la clase con la probabilidad mas alta.

$$Clase = MAX(C_1, C_2, ..., C_{|C|}) \tag{2.2}$$

La probabilidad puede ser estimada utilizando una fórmula de Bayes simple y $P(C_k|d_i)$ puede ser reescrita como:

$$P(C_k|d_i) = P(d_i|C_k) * \frac{P(C_k)}{P(d_i)} \tag{2.3}$$

El clasificador estima $P(d_i|C_k)$, $P(C_k)$, donde $P(d_i|C_k)$ es la probabilidad del documento d_i de pertenecer a la clase k. $P(C_k)$ es la probabilidad previa de la clase C_k y $P(d_i)$ la probabilidad del documento de entrenamiento d_i. $P(d_i)$ es constante, por lo que en el contexto de clasificación textual, usando la representación de bolsa de palabras (Bow) se puede calcular $P(d_i|C_k)$ de la siguiente manera:

$$P(d_i|C_k) = P(Bow(d_i)|C_k) = P(W_{1,i}, W_{2,i}, ..., W_{|V|,i}|C_k)P(C_k) \tag{2.4}$$

Pero la suposición del clasificador es que la palabra j^{th} en el documento i^{th} no esta correlacionada con las demás palabras.

$$P(d_i|C_k) = P(W_{1,i}, W_{2,i}, ..., W_{|V|,i}|C_k) = \prod_{j}^{|V|} P(W_{j,i}|C_k)P(C_k) \tag{2.5}$$

Reduciendo el problema a estimar la probabilidad de la palabra W_{ji} con respecto a la clase C_k. Como se muestra en la siguiente formula.

$$P(W_{ij}|C_k) = \frac{nW_{i,j} + 1}{|D| + |U|} \tag{2.6}$$

Donde $nW_{i,j}$ indica el número de veces que aparece la palabra $W_{i,j}$ en los documentos de la clase C_k, $|D|$ es el número de palabras únicas en la clase C_k y $|u|$ es el total de palabras únicas en toda la colección.

Máquinas de Soporte Vectorial (SVM)

La máquina de soporte vectorial es un algoritmo de clasificación supervisado que ha sido extensivamente utilizado para clasificación de texto dado a sus resultados satisfactorios[Joachims, 1998]. Un documento d_j es representado por vector $t_{d1}, t_{d2}, ..., t_{dj}$ pesado por la frecuencia de los términos. El algoritmo puede separar dos clases: una clase positiva $L1$ (indicada por $y = +1$) y una clase negativa $L2$ (indicada por $y = -1$).

En el espacio de vectores de entrada un hiperplano puede ser definido ajustando $y = 0$ en la siguiente ecuación lineal:

$$y = f(\overrightarrow{t_d}) = b_0 + \Sigma_{j=1}^{N} b_j t_{d_j} \tag{2.7}$$

El algoritmo determina un hiperplano el cual está localizado entre los ejemplos positivos y negativos del conjunto de entrenamiento. El parámetro b_j es adaptado de tal forma que la distancia ξ llamada "margen" sea la más cercana a los ejemplos positivos y negativos. Los documentos que tengan una distancia igual a ξ son llamados "vectores de soporte" y determinan la localización del hiperplano. Por lo general solo una fracción de los documentos serán vectores de soporte como se muestra en la figura 2.4 solo 3 documentos se consideran vectores de soporte, 2 para la clase 1 y 1 para la clase 2.

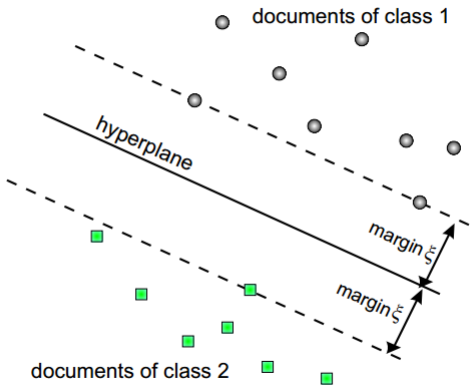

Figura 2.4: Hiperplano con la máxima distancia (margen) de los documentos de la clase positiva y negativa construido por SVM.

Un documento nuevo con un vector de términos $\overrightarrow{t_d}$ es clasificado como $L1$ si el valor $f\overrightarrow{t_d} > 0$ y como $L2$ si $f\overrightarrow{t_d} < 0$.
En caso de que los vectores de los documentos de dos clases no sean linealmente separables, el hiperplano es colocado de tal forma que la menor cantidad de documentos sean colocados del lado equivocado. Las ventajas de este clasificador son las siguientes:

- El algoritmo SVM es independiente de la dimensión de los atributos.
- Para problemas donde el espacio de características es muy disperso el algoritmo SVM es de los más apropiados.
- La mayoría de los problemas de categorización de textos son linealmente separables.

2.1.3. Medidas de evaluación

Para la evaluación de los resultados en las tareas de clasificación de texto se pueden utilizar diferentes medidas de evaluación. Ya que éstas juegan un rol muy importante para discriminar y obtener un clasificador óptimo.
Para tareas de clasificación se tienen los siguientes términos:

- Verdaderos Positivos (VP). Resultados positivos identificados correctamente.
- Falsos Positivos (FP). Resultados negativos identificados como positivos.
- Verdaderos Negativos (VN). Resultados negativos identificados correctamente.
- Falsos Negativos (FN). Resultados positivos identificados como negativos.

Con estos términos se pueden definir las siguientes métricas de evaluación para los resultados obtenidos por el clasificador.

Exactitud

La exactitud es una medida global, ya que se refiere a la capacidad del clasificador para categorizar correctamente los documentos. El valor de exactitud ésta definido entre los valores de 0 y 1. Se define de la siguiente manera:

$$Exatitud = \frac{VP + VN}{VP + FP + VN + FN} \tag{2.8}$$

Precisión

La precisión indica la especificidad del clasificador y puede ser vista como la probabilidad de un elemento que el clasificador marco como positivo en realidad lo sea. Está definida de la siguiente manera:

$$precision = \frac{VP}{VP + FP} \tag{2.9}$$

Una precisión alta indica una cantidad menor de falsos positivos. Por lo que los resultados obtenidos serán correctos.

Recuerdo

El recuerdo indica la completitud del clasificador y puede ser visto como la probabilidad de que un documento positivo sea identificado correctamente por el clasificador. Está definido de la siguiente manera:

$$Recuerdo = \frac{VP}{VP + FN} \tag{2.10}$$

Un recuerdo alto indica una cantidad menor de falsos positivos. Los resultados devueltos abarcaran a la mayoría de resultados que corresponden a las diferentes clases predefinidas.

Las dos medidas anteriores (precisión y recuerdo) están relacionadas, por lo general si se desea incrementar alguno de estos valores el otro se verá afectado reduciéndose. Por lo que se debe de realizar un análisis para conocer cual valor es más importante para la tarea que se esté realizando. De este modo se pueden hacer ajustes para obtener resultados más altos para alguno de los valores en específico.

Medida F1

La precisión y el recuerdo se pueden combinar para producir una sola medida conocida como medida F1. La cual es la media armónica ponderada de la precisión y del recuerdo multiplicado por una constante 2. El valor de la medida F1 se encuentra entre los valores de $0 y 1$. Esta medida está representada en la siguiente formula:

$$MedidaF1 = 2 * \frac{precision * recuerdo}{precision + recuerdo} \tag{2.11}$$

2.2. Características Sintácticas

La sintaxis es el conjunto de reglas que se utilizan para la construcción de oraciones, estas reglas pueden ser diferentes en cada idioma. La sintaxis se encarga de decidir si una oración es gramaticalmente correcta, ésta utiliza una gramática muy extensa formada por todas las reglas del lenguaje en cuestión.

2.2.1. Partes de la oración

Las partes de la oración son las categorías en las cuales son agrupadas todas las palabras de un idioma. Estas categorías son definidas dependiendo del lenguaje, algunas palabras pueden pertenecer a varias categorías dependiendo de la semántica de la oración.

En el idioma español las palabras pertenecen a 9 categorías fundamentales (sustantivos, pronombres, adjetivos, artículos, verbos, adverbios, preposiciones, conectores e interjecciones). Las primeras cinco son variables, es decir, al usarlas cambian su terminación dependiendo del género y el número al que se estén refiriendo (artículos, sustantivos, pronombres y adjetivos). Para los verbos la terminación depende de la persona, el número, el tiempo y el modo. Las últimas cuatro son invariables lo que significa que nunca cambian su forma en ningún momento en cualquier oración. En la figura 2.5 se muestran las categorías con algunos ejemplos.

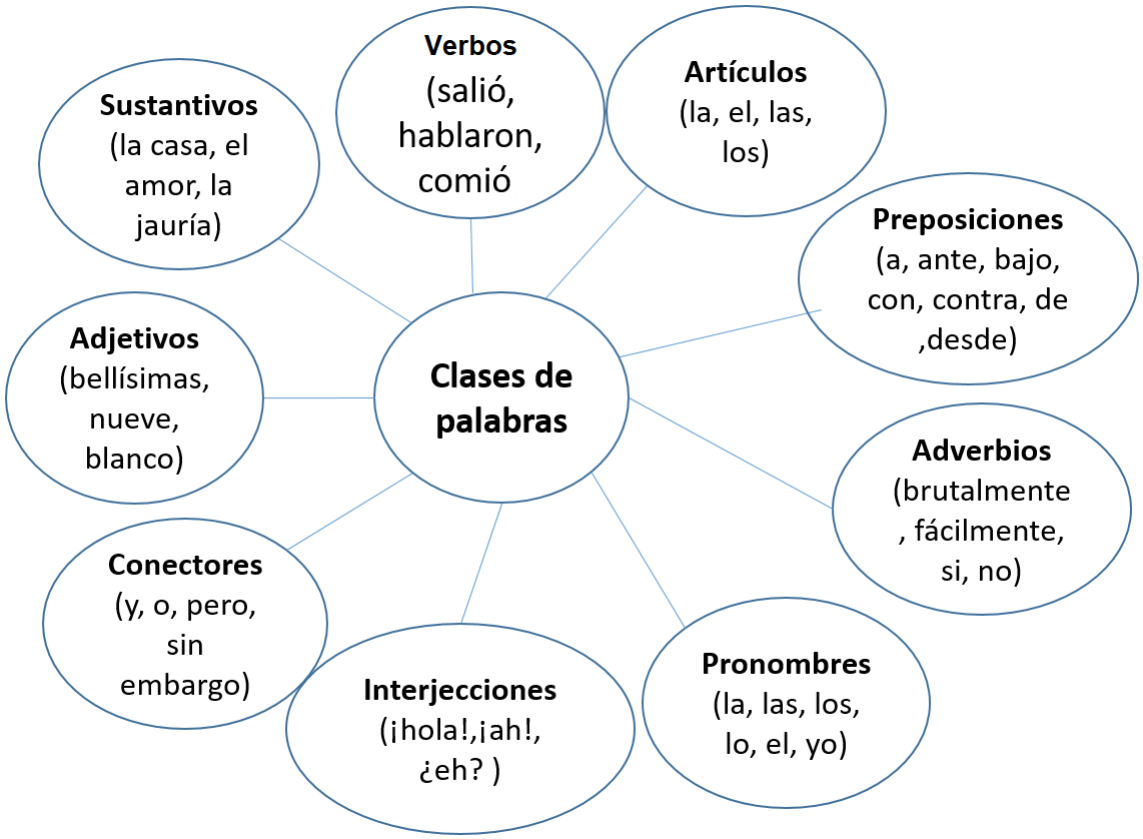

Figura 2.5: Categorías de las palabras en el idioma Español

2.2.2. Etiquetado de partes del habla

El etiquetado de partes del habla o gramatical es el proceso de asignar a cada una de las palabras de un texto su categoría gramatical. Este proceso puede ser realizado de acuerdo con la definición de la palabra o el contexto en que aparece. Se realiza mediante el empleo de algoritmos que realizan el etiquetado mediante etiquetas descriptivas predefinidas.

Existen dos propuestas generales para abordar este proceso, utilizando aproximaciones lingüísticas o aproximaciones de aprendizaje automático. La primera esta basada en la creación de un conjunto de reglas establecidas por expertos o aprendidas de forma semi-automatica. La segunda esta basada en aprendizaje basadas en

corpus las cuales utilizan textos anotados con información lingüística para establecer los modelos estadísticos.

En este trabajo se optara por la segunda opción ya que se utilizara una herramienta que esta construida mediante modelos creados a partir de aprendizaje automático.

2.2.3. N-gramas

Un $N-grama$ es una secuencia de N elementos de una secuencia dada. Se ha utilizado en estudios de procesamiento de lenguaje natural (PNL), secuenciado de genes y en el estudio de la secuencia de aminoácidos.
En el estudio de PNL se pueden construir $N-gramas$ sobre la base de distintos tipos de elementos, como, por ejemplo:

- fonemas
- sílabas
- letras
- palabras

Como se muestra en el siguiente ejemplo, una oración puede dividirse en n-gramas de la siguiente manera:

El eclipse de sol duro solamente un par de minutos.
unigramas: El, eclipse, de, sol, duro, solamente, un, par, de, minutos.
bigramas: El eclipse, eclipse de, de sol, sol duro, duro solamente, solamente un, un par, par de, de minutos.
trigramas: El eclipse de, eclipse de sol, de sol duro, sol duro solamente, duro solamente un, un par de, par de minutos.

Esta técnica es ampliamente utilizada en algoritmos de aprendizaje automático para la extracción de datos a partir de cadenas de texto, también se han utilizado para la caracterización de perfiles y en la clasificación temática.

2.3. Características Semánticas

La semántica es el estudio de los aspectos del significado, sentido o interpretación de signos lingüísticos, tales como símbolos, palabras, expresiones o representaciones formales. Mientras que la sintaxis solo estudia las reglas de construcción de expresiones, en otras palabras, estudia la construcción correcta de oraciones según el lenguaje en que se esté escribiendo o hablando.

La semántica aparte de estudiar el significado de las palabras, también estudia sus relaciones. Este tipo de relaciones pueden ser alguna de las siguientes:

- Hiperonimia e Hiponimia: un hiperónimo es una palabra cuyo significado abarca al de otras que se conocen como hipónimos. Ejemplo: Mueble es hiperónimo de silla o mesa.
- Antonimia: dos palabras son antónimos cuando su significado es contrario. Ejemplo: alto y bajo, negro y blanco.
- Monosemia: cuando una palabra tiene un solo significado.
- Polisemia: Las palabras polisémicas son aquellas que tienen diferentes significados.
- Sinonimia: dos palabras son sinónimas si tienen significados muy parecidos, pero están escritas de diferente manera.

El anexar información semántica a la información léxica ha servido para agregarle un contexto o significado a los documentos u oraciones. Con ésta información se

pueden clasificar los documentos por categorías, por ejemplo: noticias por temática (política, deportes, cultura, etc.), distinguir entre libros de diferentes tipos (misterio, comedia, educativos).

Existen diversas maneras de utilizar la información semántica, una de ellas es a través de recursos semánticos.

2.3.1. Recursos semánticos

Una red semántica es una forma de representación de conocimiento lingüístico en la que los conceptos y sus interrelaciones se presentan mediante un grafo. Si no existen ciclos estas redes pueden ser visualizadas como árboles.

Las redes semánticas están conformadas por:

- Nodos: estos son representaciones de palabras o conceptos.
- Enlaces o aristas: éstas expresan las relaciones semánticas que tienen entre si las palabras.
- Etiquetas de aristas: que indican la relación en particular que tienen los nodos.

Como se ha mencionado existen varios tipos de relaciones semánticas. Dado un conjunto de conceptos, estos comienzan a relacionarse con todos aquellos que tienen alguna relación semántica de las indicadas anteriormente.

Existen muchas redes semánticas en Internet, principalmente para el idioma inglés y para diferentes dominios, sin duda la mas utilizada es WordNet ya que es de propósito general y es de un uso sencillo.

WordNet

WordNet es la red semántica en idioma inglés que más se ha utilizado en tareas de PLN, creada en 1985 en la Universidad de Princeton[Fellbaum, 1998]. Esta red está compuesta por synsets que son grupos de palabras que tienen una relación de sinonimia, además de proveer una pequeña descripción de las palabras y registros de varias relaciones semánticas entre ese conjunto de sinónimos y otros synsets.

Como se muestra en la figura 2.6 al buscar una palabra en WordNet se obtiene una descripción de la palabra buscada, ademas de listar las relaciones encontradas que pueden ser consultadas.

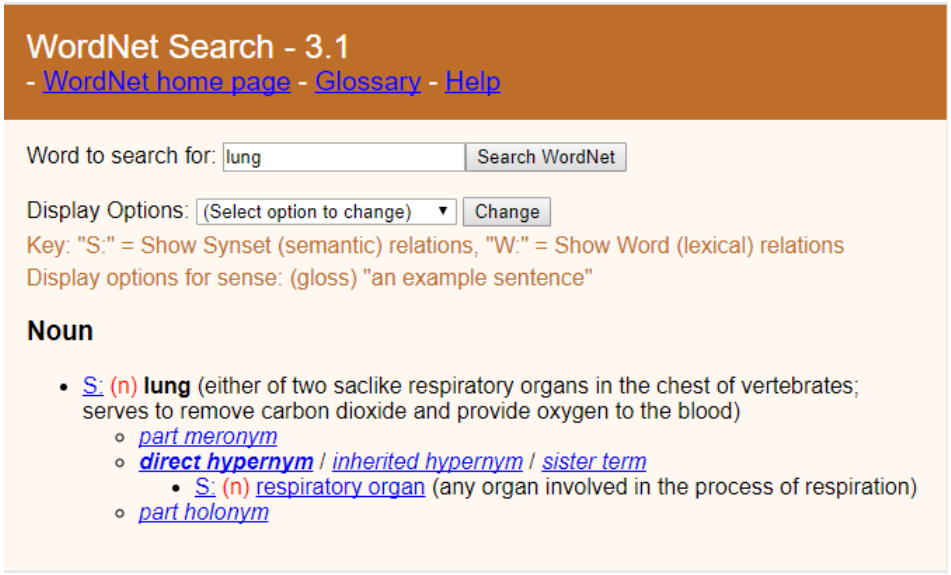

Figura 2.6: Búsqueda de relaciones para la palabra lung.

BabelNet

BabelNet nace con la integración de WordNet y Wikipedia creando una red semántica multilingüe que provee conceptos y entidades lexicalizadas en muchos idiomas y conectadas a través de vastas relaciones semánticas. Similar a WordNet en

BabelNet se agrupan a las palabras de distintos idiomas en conjuntos de sinónimos llamados BabelSynsets, por cada uno de estos grupos se proveen definiciones en varios idiomas obtenidos tanto de WordNet como de Wikipedia.

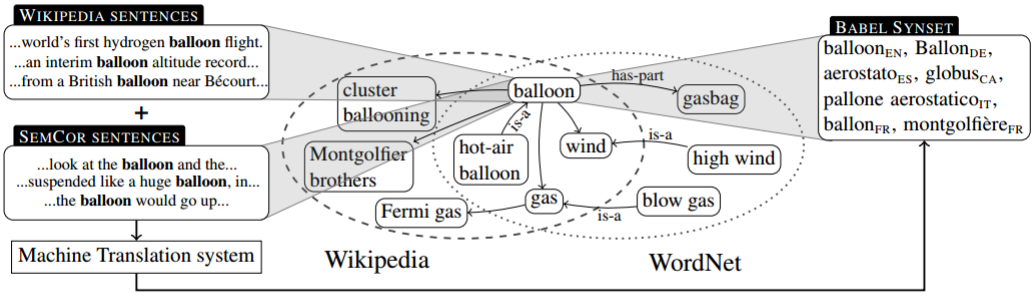

Figura 2.7: BabelNet integra información de WordNet y Wikipedia

La metodología de BabelNet mostrada en la figura 2.7 consiste de tres partes:

- Combinación de conceptos. De manera automática se busca integrar los conceptos de Wikipedia y WordNet, de esta forma se fusionan conceptos iguales además de evitar conceptos duplicados. Mediante este proceso se enriquece BabelNet con información de ambas fuentes.
- Información Multilingüe. Se recopila toda la información multilingüe de los conceptos obtenidos en el primer paso, para ello se utilizan las traducciones generadas por humanos provistas por Wikipedia.
- Establecer relaciones entre Synsets. Esto se realiza mediante la recolección de todas las relaciones encontradas en WordNet. Además de obtener las relaciones entre paginas o conceptos de Wikipedia, además de realizar las relaciones de los conceptos en los lenguajes de interés en Wikipedia.

BabelNet actualmente cubre seis idiomas: inglés, catalán, francés, alemán, italiano y español. Contiene alrededor de 3 millones de conceptos y más de 26 millones

de relaciones (disponibles para todos los idiomas registrados en BabelNet). Todas las relaciones en BabelNet son de tipo semántico, la mayor parte proceden de Wikipedia debido a que WordNet está diseñado principalmente para el idioma inglés.

BabelNet puede ser consultado mediante su sitio web *www.babelnet.org* para conceptos particulares o mediante el API provista en *babelnet.org/guide* para consultas orientadas a la investigación.

Capítulo 3

Trabajo relacionado

Para resolver tareas de Procesamiento de Lenguaje Natural (PLN) se ha optado por agregar información de tipo sintáctica y/o semántica a la información de tipo léxica, para poder solventar las debilidades que esta representación tiene. Dependiendo de las tareas que se estén abordando la información sintáctica o la semántica pueden ayudar a mejorar los resultados obtenidos.

A continuación, se presentan trabajos relacionados con las técnicas utilizadas para realizar este trabajo en tareas pertenecientes al área de PLN, además de trabajos relacionados con la problemática abordada en esta investigación.

En [Harish et al., 2010] se hace una revisión sobre las diferentes representaciones con las que se ha trabajado con documentos de texto. Tales como Bolsa de palabras o modelo vectorial, n-gramas, análisis de semántica latente, lenguaje de red universal o representaciones basadas en conocimiento. Cada una de las representaciones tienen sus ventajas y desventajas. Ya sea en cuestiones de tiempo de procesamiento, carga de memoria, perdida de información o dificultad de implementación.

Ahora bien la tarea a resolver en este trabajo puede ser vista como una tarea de

textos cortos al tratarse de oraciones que no exceden de las 30 palabras por oración, debido a esto, se presentan trabajos relacionados a la tarea de textos cortos, además de trabajos relacionados al uso de información sintáctica y semántica también se presentan trabajos relacionados a plantas medicinales resueltos mediante técnicas de PLN.

3.1. Clasificación de Textos Cortos

Los textos cortos han sido usados en muchos campos tales como: mensajes SMS, mensajes instantáneos, títulos de noticias, comentarios de blogs, comentarios de noticias, etc. Su principal característica es que la longitud del texto es muy corta, no más de 200 caracteres [Song et al., 2014]. Generalmente la característica principal de los textos cortos es:

- Escasez de información. Un texto corto solo contiene pocas a una docena de palabras, es decir pocos atributos. Por esta razón no proveen suficientes co-ocurrencias de palabras o no comparten un contexto para una buena medida de similitud por lo que es difícil el extraer características del lenguaje válidas.

3.1.1. Basados en Recursos Semánticos

El problema de la clasificación de textos cortos recae en elegir una representación razonable, la forma de escoger los atributos correctos, la reducción de dimensiones y ruido. Todo esto para incrementar la exactitud de los resultados obtenidos en la clasificación.

Se han realizado diferentes enfoques para la solución de este problema, centrándose en el enriquecimiento de la información base. Para ello se han utilizado diferentes recursos externos que sirvan para obtener esta información uno de ellos es Wikipe-

dia como en [Li et al., 2017] donde se busca relacionar conceptos obtenidos de las oraciones base, con conceptos encontrados en Wikipedia. Se utilizan diferentes formas de agregar información como: agregar el concepto directamente de Wikipedia, agregar el valor de relación entre el concepto original y el encontrado en Wikipedia, por último también se pueden agregar todas aquellas palabras que se encuentren en la página del concepto de Wikipedia a la oración original.

Otro trabajo que utiliza Wikipedia como fuente de información semántica es [Takeda et al., 2017] donde se construyen árboles con pesado. Estos son construidos mediante la categorización de los artículos contenidos en Wikipedia en categorías establecidas con anterioridad. Posteriormente para realizar la clasificación se realiza la construcción del árbol con las categorías encontradas en los datos de prueba y construir el árbol correspondiente con estas categorías, para obtener un valor de similitud se busca encontrar el árbol que contenga las categorías del conjunto de pruebas, además se obtiene una mejor similitud si partes de los árboles son similares, es decir, si comparten nodos y subnodos.

Otro enfoque utilizado con recursos externos puede verse en [Wang et al., 2014] donde se utiliza la representación de "bolsa de conceptos"(BOC) en sustitución de la comúnmente utilizada "bolsa de palabras"(BOW) donde se crean modelos de conceptos relacionados a cada una de las clases a categorizar, mediante la conversión de entidades extraídas del texto a conceptos que pueden ser agrupados en estos modelos. Posteriormente la consulta es "conceptualizada" para poder compararla con los modelos generados y así poder asignarle una categoría.

Además de utilizar bases de conocimiento, también se han utilizado herramientas para el enriquecimiento de textos cortos, como se muestra en [Batool et al., 2013] se utilizan herramientas para resolver la tarea de análisis de sentimiento en Twitter. De los Twitts se obtienen palabras clave y su sentimiento relacionado, posteriormente mediante una herramienta que utiliza Wordnet como fuente de conocimiento

se obtienen palabras relacionadas (sinónimos) que serán agregadas a las palabras originales para su clasificación. Mediante el uso de estas herramientas se espera que la clasificación tenga un mejor desempeño.

Otro tipo de recurso semántico que se puede utilizar para agregar información es el uso de diccionarios, en [jin Tang et al., 2013] se hace uso de un diccionario semántico creado manualmente mediante la inclusión de "palabras efectivas" provenientes del repositorio de HowNet y otras librerías orientadas al campo financiero. El valor de pesado de las palabras está relacionado con su pertenencia a cada una de las categorías que contiene el diccionario. Para la parte de clasificación de las palabras que contiene cada elemento del conjunto de prueba, se busca en el diccionario y se le asigna el valor de peso que tenga asignado en el diccionario. En este mismo proceso se agregan palabras al diccionario si es necesario evaluando la palabra con las categorías presentes en el diccionario. Así se va enriqueciendo el diccionario para posteriores usos.

3.1.2. Basados en Motores de Búsqueda

La idea de este enfoque es la de incluir información obtenida a través de un buscador a los datos de entrenamiento. En [Meng et al., 2013] se utiliza este tipo de enfoque, el cual se basa en realizar una consulta a el navegador con cada uno de los datos de entrenamiento, los resultados devueltos a manera de enlaces y resúmenes son almacenados, los resúmenes son combinados junto con la consulta original. Esto se realiza para cada uno de los elementos del conjunto de entrenamiento.

Mediante esta expansión los elementos para entrenar crecen significativamente en tamaño además que las palabras agregadas están relacionadas directamente con los datos originales.

Otro trabajo donde se utiliza la expansión vía motores de búsqueda es en

[Wei et al., 2010] donde se utiliza para clasificar información de "intención de comercio en línea", la idea de este trabajo es clasificar consultas que estén relacionadas a alguna forma de comercio, para ello se realiza una clasificación de consultas agregando información del contenido de páginas relevantes obtenidas a través de un navegador, de ésta manera se reporta que se obtiene un 10 % de mejora en exactitud con respecto a la información inicial.

3.1.3. Basados en Corpus

Al utilizar esta técnica se tiene como idea enriquecer los datos de entrenamiento con información similar que se encuentre dentro de otros conjuntos de datos ya sea relacionados a la tematica o de proposito general. En [Islam et al., 2012] la idea es analizar la similitud de un par de palabras ($p1$ y $p2$) basada en los $tri-gramas$ que comiencen con la primera palabra y terminen con la segunda ($p1-px-p2$) donde px es cualquier palabra y viceversa ($p2-px-p1$). Para ello utilizarán información provista por *Google n-Grams* [Michel et al., 2011]. Para llevarlo a cabo se obtiene la frecuencia de los $tri-gramas$ que satisfacen ambas combinaciones, mediante esta estadística lo que se busca encontrar es en que grado las palabras de una oración están relacionadas.

En [Zhang and Wu, 2015] se utiliza un modelo basado en $n-gramas$ para extender las características de los textos cortos. El enfoque consiste en obtener del conjunto de entrenamiento conjuntos de $bi-gramas$ o $tri-gramas$ obtenidos de manera probabilista, es decir que la probabilidad de que una palabra preceda a otro pase cierto umbral establecido. estos $n-gramas$ son almacenados en una librería. Posteriormente cuando se evalúa el conjunto de prueba se buscan las palabras en cada uno de los $n-gramas$ agregando las palabras faltantes del $n-grama$ al elemento original. Una vez realizado esto se puede hacer uso de un clasificador para realizar la categorización.

Otro enfoque basado en corpus es utilizar información que se puede obtener directamente de éste, en [Shrestha, 2011] donde además de la información provista por la información inicial, se agrega información relacionada a los términos en el corpus como: la importancia de cada término en la colección (idf), la co-ocurrencia de términos y la distribución sobre todas las oraciones en la colección. Mediante la medida de similitud de coseno se puede obtener un valor de similitud entre las oraciones.

3.2. Trabajos relacionados a plantas medicinales

El estudio de las interacciones de la sociedad con la naturaleza, puede ser abordado con diferentes herramientas y desde diferentes perspectivas. Hoy en día se han realizado trabajos donde se utilizan técnicas computacionales para resolver estudios relacionados a la etnobotánica.

A continuación, se presentan algunos de estos trabajos.

La bioprospección puede comprenderse como una nueva forma de usar la biodiversidad a través de la búsqueda o exploración sistemática de fuentes biológicas con potencial de explotación económica mediante el desarrollo de nuevos productos o componentes.

En [Barguil et al., 2016] se realizó un sistema para la recuperación de información acerca de plantas partiendo de documentos científicos para poder ayudar en la toma de decisiones en temas de bioprospección.

En otros estudios se han utilizado métodos computacionales para extraer y priorizar información etnobotánica de literatura de conocimiento biomédico. En [Sharma et al., 2016] se realizó un estudio para poder relacionar información de especies de plantas provenientes de manuales de uso en ciertos países con conceptos

relacionados a enfermedades de la literatura biomédica indexada en MEDLINE. En esta investigación se obtuvieron resúmenes y títulos de artículos de MEDLINE utilizando como consulta un conjunto de plantas de origen micronesio. Se encontraron relaciones de 129 plantas de 180 en total, 19,798 citas donde se menciona a alguna de estas plantas de las cuales contienen 18,322 conceptos de MEDLINE. Un total de 22,425 co-relaciones entre plantas y conceptos fueron encontrados.

Por otra parte, en [Sharma and Sarkar, 2013] realizaron un estudio similar al anterior centrándose en plantas que tuvieran un uso potencial en terapias (fitoterapias). Se hace uso de un enfoque basado en conceptos para cubrir el conocimiento localizado dentro de literatura biomédica. Se busca recuperar asociaciones entre plantas y enfermedades humanas, centrándose en la identificación de fitoterapias descritas en MEDLINE. Se utilizaron descriptores y conceptos proporcionados por estos recursos. La identificación de este tipo de relaciones puede ser útil para enfoques de bioprospección y en la exploración de drogas. Los resultados obtenidos muestran 22,050 relaciones entre plantas y enfermedades, obteniendo valores de precisión de 0.78 y de recuerdo de 0.70 indicando que este enfoque puede ser utilizado para obtener relaciones entre conceptos extraídos de manuales o documentos informales y conceptos medicinales descritos en documentos científicos.

3.3. Discusión

En la clasificación de textos cortos se han utilizados varios enfoques para enriquecer la información que originalmente puede ser incompleta. Los métodos basados en corpus tienen la ventaja de no necesitar de información externa para enriquecer la representación de los datos, aunque esa misma característica puede ser una desventaja también debido a que sin información externa no se puede agregar información útil que no se encuentre en el corpus.

Por otro lado, los métodos basados en motores de búsqueda cuentan con todo el Internet para obtener información útil, pero esto conlleva a el uso constante de Internet lo que puede ser un proceso bastante lento. Además de que se puede obtener mucha información que no es relevante para el dominio que se está abordando.

Por último, los métodos basados en recursos semánticos dependen principalmente en la información disponible en el mismo, ya que si la información no es tan amplia o dedicada al dominio en particular no puede ser de utilidad para la tarea que se está abordando.

Nuestro trabajo se basa en obtener atributos de tipo semántico y sintáctico que sean relevantes para poder clasificar de manera correcta oraciones que contengan usos medicinales de plantas, al tratarse de oraciones, se ha optado por manejar este problema como uno de clasificación de textos cortos. Si bien se ha visto que la información semántica tiene mayor relevancia que la información sintáctica en este tipo de problemas, buscamos encontrar información de ambos tipos que nos pueda ser útil, para ello se hará uso de n-gramas para obtener atributos sintácticos y se usara un recurso externo para obtener atributos semánticos.

Con respecto a los trabajos relacionados con el dominio en particular de plantas medicinales se han abordado estudios de relación entre plantas y conceptos médicos que se encuentran en bases de datos indexadas de MEDLINE. Estos trabajos tienen como base el idioma inglés y han sido tratados como tareas de extracción de información. Nuestro enfoque aparte de estar basado en el idioma español se busca tratar la tarea como un problema de clasificación. Otra diferencia es que la fuente de los datos está basada en información obtenida de Internet la cual presenta un lenguaje más informal, a diferencia de los trabajos presentados que buscan encapsular la problemática a un lenguaje mas compacto y formal como lo es el usado en investigación científica.

Capítulo 4

Clasificación de Oraciones de Plantas Medicinales

En este capítulo se presenta la propuesta de trabajo para la resolución de la clasificación de oraciones de uso de plantas medicinales, utilizando información sintáctica y semántica. El enfoque se basa en el enriquecimiento de la representación base a nivel léxico, agregándole atributos de tipo sintáctico y semántico.

En el capítulo se describe en primer lugar el enfoque en general; posteriormente, se detalla el enfoque propuesto para cada una de las representaciones.

Como se ha indicado en el capítulo 3, en este trabajo la clasificación de oraciones se puede ver como una tarea de clasificación de textos cortos, donde es importante abordar el principal problema que es la falta de información. Las oraciones que se tienen recopiladas tienen una longitud promedio de 15 palabras, la oración con la menor cantidad de palabras tiene alrededor de 6 palabras. Al ser las oraciones de tan pequeña longitud conlleva a una pobre representación al momento de realizar la clasificación. Por este motivo es necesario enriquecer la representación de las oraciones agregando de alguna otra forma información que sea útil para poder categorizar

de manera correcta las oraciones de tipo medicinal.

El enfoque que se tomó en este trabajo consiste en dividir los dos tipos de información que se le puede extraer a las palabras, la información de tipo sintáctico y de tipo semántico. Una vez teniendo esa información se puede combinar para enriquecer la información léxica base.

Como se muestra en la figura 4.1, el proceso consta de 3 partes principales. En la primera parte se realizó la recopilación de oraciones donde este presente el nombre de una planta. Estas oraciones se obtuvieron mediante consultas a la web. Seguido de la parte de representación del texto donde se realizó un preprocesamiento a la información para poder obtener los atributos de tipo sintáctico y semántico. La última parte fue la de clasificación y prueba donde se clasificaron oraciones utilizando los 3 tipos de información así como sus combinaciones.

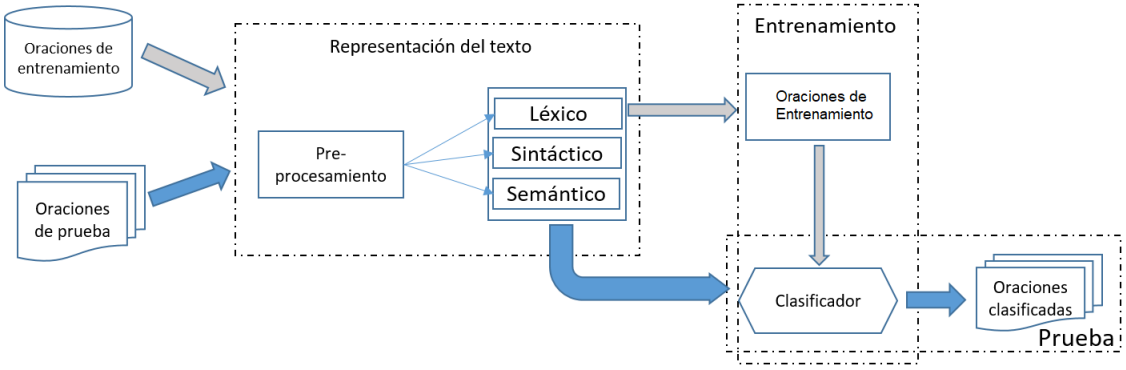

Figura 4.1: Diagrama del enfoque propuesto.

4.1. Representaciones del texto

Para la representación de la información se uso el modelo vectorial donde cada documento es representado como un vector de pesos de $|v|$ elementos los cuales

conforman el vocabulario total.

$$d_1 =< w_1, w_2, ..., w_{|v|} > \tag{4.1}$$

$$d_2 =< w_1, w_2, ..., w_{|v|} > \tag{4.2}$$

Donde w_1 es el peso del termino t_1, con cada tipo de información el tamaño del vocabulario cambia ya que cada representación genera diferentes atributos.

4.1.1. Representación léxica

La primera representación es de tipo léxico, la cual está conformada por las palabras de un lenguaje en específico. Con esta información se pretende realizar el primer experimento, el cual será la base de los experimentos posteriores. Se busca conocer los resultados obtenidos mediante solo el uso de las palabras que se encuentran en las dos clases definidas (Medicinal, No medicinal).

Para este fin se realizará la clasificación de las oraciones obtenidas solamente utilizando las palabras como información. Por lo tanto el tamaño del vector de los documentos esta definido por el total de las palabras que componen la colección de oraciones.

$$d_1 =< w_1, w_2, ..., w_{|v|} > \text{donde} |v| \text{ es el total de palabras de la colección.} \tag{4.3}$$

Como se muestra en el ejemplo siguiente una oración es preprocesada y posteriormente representada en el modelo vectorial, donde cada palabra se le asigna un valor en este caso binario si es parte de la oración o no. Los elementos p_i, p_j y p_k son palabras que no están en esta oración y pero pertenecen al vocabulario.

Oración original -> la albahaca se suele usar para bajar la fiebre durante distintas infecciones

Oración limpia -> albahaca suele usar bajar fiebre distintas infecciones

		albahaca	suele	usar	bajar	fiebre	distintas	infecciones	p_i	p_j	p_k	$p_{\|v\|}$
Oración representada en El modelo vectorial ->	d_1	1	1	1	1	1	1	1	0	0	0	0

4.1.2. Representación Sintáctica

Si bien en muchas tareas de clasificación el uso de la información sintáctica no es considerada relevante, creemos que en esta tarea puede ser de importancia.

Analizando las oraciones se notó que en algunas de ellas se pueden notar ciertos patrones al momento de describir usos de las plantas, tales patrones consisten en $n - gramas$ de palabras. Algunos de ellos se presentan a continuación:

1. *se utiliza la* espinosilla para aliviar trastornos de tipo eruptivo como la erisipela la rubeola sarampión.
2. *para controlar la* diarrea infantil se realiza una infusión en partes iguales de aceite de oliva y la planta de la amapola.
3. en algunos países la "angelica" *es utilizada para* *aliviar los dolores* nerviosos como lo son las neuralgias migrañas.

Se observó que en las oraciones se utilizan los mismos verbos o sinónimos tales como: "utilizar", "emplear", "usar" para describir un uso como se muestra en las oraciones 1 y 3, además de estar en medio de preposiciones y artículos. Se observaron otros patrones como en 3 donde un verbo es seguido de un articulo y un nombre común. Por otro lado en las oraciones que no describen un uso medicinal se noto que estos patrones no son tan comunes o se encuentran otros diferentes lo que nos llevo a pensar que estos patrones pueden hacer diferencia entre las clases a clasificar.

Al observar estos patrones, se optó por analizar las oraciones mediante el uso de un etiquetador de partes del habla con el cual se puede obtener la categoría a la que pertenecen cada una de las palabras de las oraciones.

En la figura 4.2 se muestra el diagrama del proceso para la extracción de los atributos sintácticos.

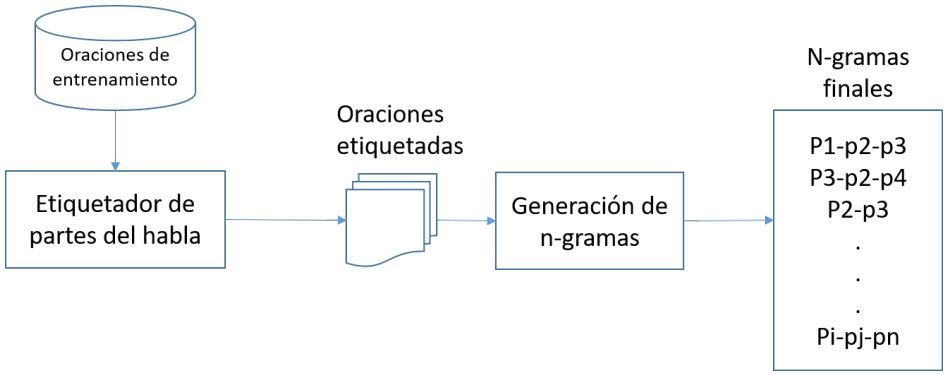

Figura 4.2: Extracción de la información sintáctica

El proceso esta descrito de la siguiente manera:

1. Obtenidas las oraciones de entrenamiento se procede a ingresarlas al etiquetador de partes del habla, como resultado obtendremos las oraciones divididas por palabra y su etiqueta respectiva.

2. Como segundo paso se realiza la generación de $n-gamas$ donde se indica la cantidad n que indicara de cuantas palabras será la secuencia. Obteniendo al final los $n-grama$ por cada oración.

Estas secuencias nos pueden aportar información acerca de la estructura con la que está conformada cada tipo de oración (medicinal, no medicinal) esperando que se pueda diferenciar entre estas dos clases de oraciones de una manera mejor.

Como se indico a principio del capítulo la representación se diferencia por el tamaño del vocabulario, para este caso $|v|$ =# de $n-gramas$, para unigramas es de 158, bigramas es de 1923 y para trigramas es de 7396.

$$d_1 =< w_1, w_2, ..., w_{|v|} > \text{donde} |v| \text{ es la cantidad de n-gramas generados.} \quad (4.4)$$

Se muestra un ejemplo de estos $n-gramas$ con la siguiente oración:

Oración original ->	la albahaca se suele usar para bajar la fiebre durante distintas infecciones
N-gramas = 3 ->	NCFP000_CS_NCMS000, NCMS000_SP_DA00S0, SP_DA00S0_PR0CS00, PR0CS00_DA0MS0_NCMS000, DA0MS0_NCMS000_SP, SP_DA0FS0_NCFS000, DA0FS0_NCFS000_SP, SP_NCFS000_AQ0FS00

4.1.3. Representación Semántica

La información semántica es utilizada para poder obtener información de contexto o dominio, es decir, se puede obtener información más allá de las palabras originales. Con esto se pueden encontrar relaciones entre las oraciones que no pueden ser obtenidas con la información léxica. Se propuso obtener información de este tipo para enriquecer las oraciones que pertenecen a la clase medicinal.

Las palabras están conectadas mediante relaciones semánticas, estas relaciones pueden ser de diferentes tipos como se comentó en el capítulo 3. La relación que nos interesa para este trabajo es la relación del hiperónimo de una palabra, que es aquella que es más general que otra y abarca su significado. Por ejemplo “mueble” es el hiperónimo de “silla” o “mesa”.

El objetivo es obtener un hiperónimo que sea compartido por varias palabras base, con esto se puede obtener información de la temática, en este caso que se refieran al ámbito médico.

El procedimiento mostrado en la figura 4.3 consistió de los siguientes pasos:

- Obtener el vocabulario. Se obtuvo del conjunto de entrenamiento las palabras únicas para evitar realizar consultas duplicadas al recurso semántico y ahorrar tiempo en este proceso.
- Consulta de categorías. En este paso se obtienen los hiperónimos de cada una de las palabras del vocabulario.
- Filtrado de categorías. Se elegirán aquellas categorías que estén relacionadas

al dominio medico o medicinal. esta selección se realizó a criterio propio.

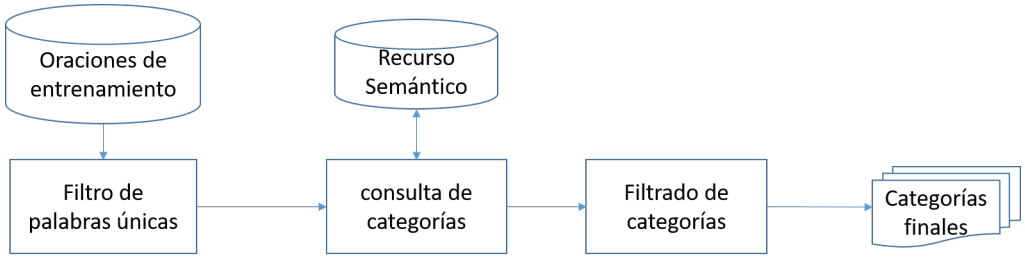

Figura 4.3: Extracción de la información semántica

Para realizar este proceso de generalización se propone el uso de un recurso semántico, en este caso BabelNet. Se eligió BabelNet debido a que es un recurso disponible para el idioma español idioma en el que se encuentran los datos utilizados en este trabajo, además de que contiene información proveniente tanto de Wikipedia como de la red semántica más conocida WordNet.

Debido a la temática se busca generalizar las palabras a un término que tengan en común, en este caso se buscara su hiperónimo directo en BabelNet como se muestra en la figura 4.4.

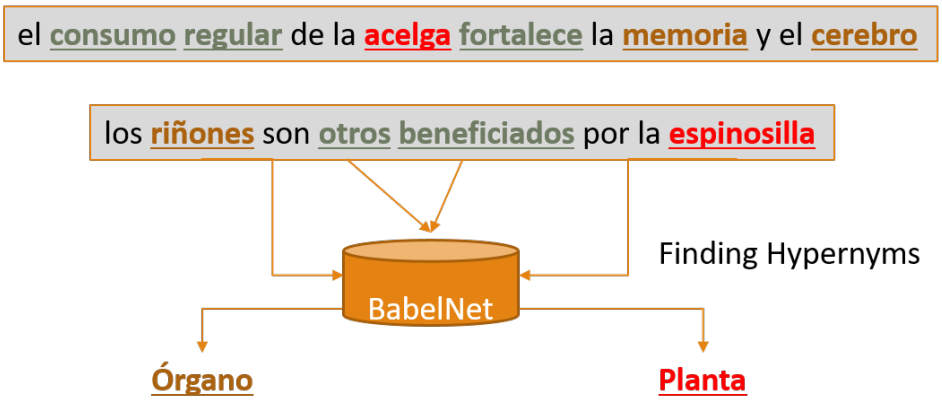

Figura 4.4: Generalización de las palabras mediante hiperónimos

De esta manera las palabras pueden ser sustituidas por su hiperónimo directo haciendo que las oraciones tengan más elementos en común y de este modo una

mayor relación al momento de la clasificación. Para la clasificación se utilizó la misma configuración que la de los experimentos anteriores.

Para la información semántica el vocabulario es: $|v| = 274$ que son los hiperónimos elegidos.

$$d_1 =< w_1, w_2, ..., w_{|v|} > \text{donde} |v| \text{ es el total de hiperónimos.} \tag{4.5}$$

Se muestra un ejemplo de los hiperónimos obtenidos para la siguiente oración:

Oración original ->	la albahaca se suele usar para bajar la fiebre durante distintas infecciones
Hiperónimos ->	Hierba, descripcion_de_espécie, táxon, tomar, cambiar, síntoma, signo_clínico, síntoma, insalubridad, enfermedad_infecciosa, medicina

Capítulo 5

Experimentos y resultados

En este capítulo se describen los experimentos realizados en esta investigación, así como los resultados obtenidos con cada una de las representaciones y la combinación de las mismas.

En primera instancia se describe la construcción del corpus, ya que para la tarea propuesta no se encuentra alguno disponible bajo los criterios del idioma español e Internet como fuente.

Partiendo del experimento base el cual solo usa información de tipo léxica, es decir únicamente las palabras originales del conjunto de datos. Añadiendo posteriormente la información sintáctica generada mediante $n-gramas$, buscando encontrar patrones frecuentes en las oraciones que pueden ser de utilidad para diferenciar entre oraciones de diferentes clases. Para obtener la información semántica se hará uso del recurso semántico BabelNet para generalizar palabras que sean similares en el contexto medicinal.

Se espera que la información sintáctica y semántica contribuyan y mejoren los resultados obtenidos usando solo información léxica. Se mostrarán los resultados obtenidos con cada una de estas representaciones, así como la combinación de éstas.

Se realizarán experimentos alternativos para poder generalizar el método propuesto mediante la clasificación de oraciones que sean de tipo descriptivo, de localización y de otro tipo de usos diferentes al medicinal. Esto se realizará con la intención de probar que el método puede ser adaptado según las necesidades que se tengan y que no está ligado exclusivamente con el dominio medicinal.

5.1. Construcción de la colección de datos

Por el momento no se encontró una colección de datos relacionada a plantas medicinales en el idioma español por lo que se optó por construirla. Para la construcción de esta colección de datos se obtuvo una lista de 250 plantas del sitio web de la biblioteca digital de la medicina tradicional mexicana [UNAM, 2009]. Mediante el buscador de Google se realizaron consultas al buscador por cada una de las plantas en la lista como se muestra en la figura 5.1 donde se realizó una consulta para la planta "manzanilla".

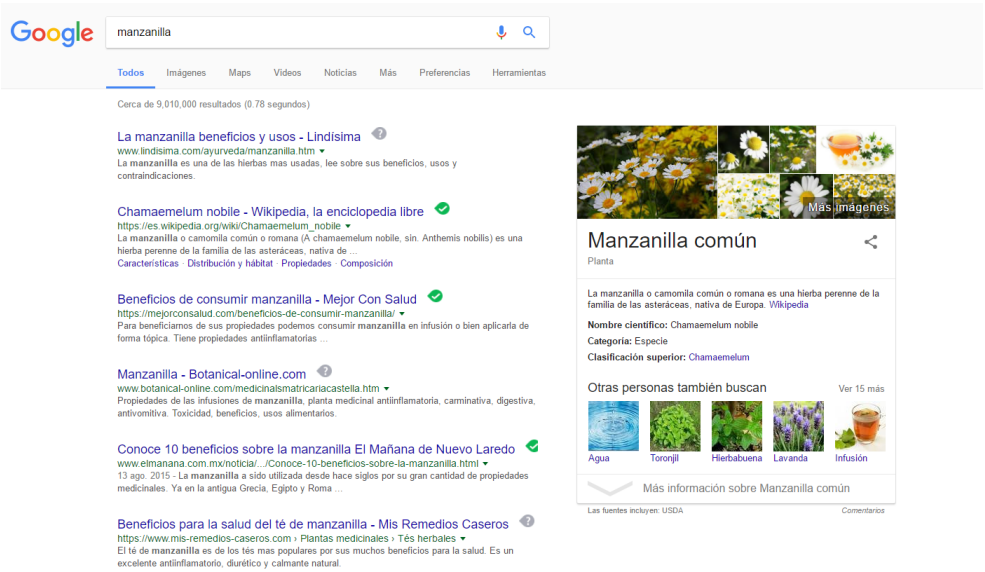

Figura 5.1: Resultados devueltos por la consulta "Manzanilla" en Google.

Se consultaron algunos de los enlaces devueltos por el buscador, buscando oraciones que cumplieran con los siguientes criterios:

- Mención de la planta de manera explícita.
- Las oraciones deben estar en el idioma español.

De la lista de plantas para realizar las consultas se encontraron oraciones que no cumplían con el criterio de mención explícita de la planta debido al formato de las páginas que las contenían. Esto se debía a que los autores se referían a la planta por el título de la página y no se colocaba el nombre de la planta en el contenido de la misma.

En la tabla 5.1 se muestran las plantas con el mayor número de oraciones obtenidas para la colección de datos.

Planta	Cantidad	Planta	Cantidad
Ajo	22	Yerba mate	14
Albahaca	22	Cebolla	14
Diente de león	20	Apio	13
Eucalipto	20	Romero	13
Hierbabuena	19	Aguacate	12
Borraja	16	Canela	11

Tabla 5.1: Plantas con mayor número de oraciones.

Algunos ejemplos de oraciones son las siguientes:

- En muchos lugares de Europa las flores del perejil son utilizadas para adornar platos o como colorante.
- La forma más común de usar la hierbabuena es haciendo infusión con sus hojas.

- El ajenjo aumenta la secreción de jugos biliares descongestionando el hígado y mejorando sus funciones.
- El jengibre se utiliza en la mayor parte de las cocinas del mundo a través de la cocina asiática.
- Consumir las semillas de chabacano en infusión o molidas en un mortero para tos o estreñimiento.

En general las oraciones utilizan verbos similares ya sea para describir un uso medicinal u otro diferente, además de que las oraciones de uso medicinal no comparten palabras en común por lo que es necesario utilizar la semántica para encontrar una relación.

5.1.1. Etiquetado de las oraciones

Una vez que se obtuvieron las oraciones se procedió a realizar el etiquetado de las mismas asumiendo los siguientes criterios:

- Clase Medicinal.
  - Debe mencionar de manera explícita la enfermedad o síntoma a curar.
  - Debe mencionar de manera explícita la parte del cuerpo o el área a tratar.
- Clase No Medicinal.
  - Otro uso. Oraciones donde se describan el uso de las plantas para un uso diferente al medicinal, por ejemplo: culinario, industrial, construcción, etc.
  - Localización. Oraciones donde se menciona el origen o lugar de crecimiento de la planta.

- Descripción. Oraciones donde se menciona alguna planta ya sea de forma general o específica.

Se obtuvieron 2000 oraciones en total, de las cuales 1000 oraciones describen un uso medicinal de alguna de las plantas de la lista, estas oraciones representan la clase positiva etiquetadas como: "medicinal". Las 1000 oraciones restantes representan la clase negativa etiquetadas como: "no medicinal".

Las oraciones de la clase "no medicinal" consisten en diferentes tipos de oraciones como se muestra a continuación:

Tipo de oración	Contenido
Otro uso	el aguacate es ampliamente conocido por su capacidad humectante en el mundo de la estética.
Otro uso	muchos soldados franceses murieron luego de azar conejos con las ramas secas de la adelfa
Localización	el achiote es un arbusto de de la familia de las bixaceas que crece en las regiones intertropicales
Localización	la planta de alcaparra es originaria de las costas occidentales del mediterraneo
Descripción	la artemisa es una planta perenne la cual alcanza hasta los 3 metros de altura y sus tallos son angulares
Descripción	el aconito es una de las plantas mas toxicas conocidas por el hombre.

Tabla 5.2: Oraciones que componen la clase 'no medicinal'"

Como se puede observar en la tabla 5.2 las oraciones mas parecidas a las de la clase "medicinal" son las que describen otros usos seguidas de las que indican una descripción y las oraciones que se diferencian mas son las de tipo localización.

en la tabla 5.3 se indica el total de oraciones por tipo.

Tipo de oración	Cantidad
Otro uso	343
Localización	162
Descripción	495

Tabla 5.3: Tipo de oraciones de la clase "No medicinal".

5.2. Experimentos para la clase Medicinal

Estos experimentos corresponden al tema principal de este trabajo, el clasificar de manera correcta oraciones donde se describa un uso medicinal de una planta.

En primer lugar, se realizó el experimento base, con la idea de clasificar las oraciones utilizando únicamente las palabras que las componen. Posteriormente se indican los experimentos extrayendo la información sintáctica y semántica además de los experimentos combinando los 3 tipos de información.

5.2.1. Experimento léxico

En esta fase se realizó el experimento utilizando las palabras que conforman las oraciones, este experimento conforma la base de los resultados del cual se partirá para mejorar mediante la integración de los otros tipos de información.

Pre Procesamiento

Este primer experimento se realizó la limpieza de las oraciones de la siguiente manera:

- Conversión de los caracteres a minúscula.
- eliminación de los signos de puntuación.
- eliminación de palabras que consisten de un solo carácter.
- eliminación de caracteres diferentes a letras.

Con el conjunto de datos procesado, se realizó el experimento utilizando 3 de los clasificadores más utilizados en clasificación de textos: K−vecinos más cercanos

(KNN), Máquinas de soporte vectorial (SVM) y Bayes multinomial (BM). Para el caso de KNN se utilizó un $K = 5$ y se realizaron experimentos con 2 tipos de pesado de términos: Frecuencia del termino (FT) y el pesado binario (PB). Se realizó esto para poder observar el desempeño de los clasificadores con diferentes configuraciones y se utilizó validación cruzada de 10 pliegues

Si bien se realizó la clasificación mediante los tres clasificadores mencionados anteriormente, se muestran resultados del clasificador SVM por ser el que mejores resultados obtuvo. Los resultados de los otros dos clasificadores pueden ser encontrados en los anexos al final de este trabajo.

Los resultados mostrados en las tablas consisten de las siguientes columnas:

- Pesado. Indica el tipo de pesado utilizado en la representación. FT (pesado por frecuencia) y Binario (se encuentra o no el término).
- Atributos. Muestra el total de atributos de la representación.
- Clase. Indica la clase que se evaluó en la clasificación.
- Precisión. Valor obtenido de precisión por la clase, entre paréntesis la desviación estándar de los datos.
- Recuerdo. Valor obtenido de recuerdo por la clase, entre paréntesis la desviación estándar de los datos.
- F-measure. Valor obtenido de f-measure por la clase, entre paréntesis la desviación estándar de los datos.

Pesado	Atributos	Clase	Precisión	Recuerdo	F-Measure
FT	5409				
		Medicinal	0.845 (0.041)	0.829 (0.089)	0.836 (0.065)
		No Medicinal	0.913 (0.058)	0.934 (0.006)	0.922 (0.027)
Binario	5409				
		Medicinal	**0.855 (0.032)**	**0.834 (0.094)**	**0.843 (0.064)**
		No Medicinal	**0.913 (0.058)**	**0.934 (0.006)**	**0.922 (0.027)**

Tabla 5.4: Resultados de la clasificación utilizando solo la parte léxica.

En la tabla 5.4 se puede apreciar una ligera mejora en los resultados con pesado binario del pesado por frecuencia, aunque es mínima.

Se realizó un segundo experimento en el cual el conjunto de datos tuvo un proceso de lematización, el cual consiste en eliminar las conjugaciones de las palabras para representarlas en su forma base o raíz. De este modo se pueden generalizar palabras que de manera conjugada son tomadas como diferentes para el clasificador. La lematización se realizó utilizando el software Freeling, una vez realizado esto se procedió a realizar la clasificación nuevamente con la misma configuración presentada anteriormente.

Pesado	Atributos	Clase	Precisión	Recuerdo	F-Measure
FT	3900				
		Medicinal	**0.891 (0.021)**	**0.873 (0.073)**	**0.881 (0.048)**
		No Medicinal	**0.927 (0.027)**	**0.927 (0.027)**	**0.927 (0.027)**
Binario	3900				
		Medicinal	0.864 (0.010)	0.823 (0.123)	0.839 (0.070)
		No Medicinal	0.910 (0.044)	0.918 (0.018)	0.914 (0.031)

Tabla 5.5: Resultados de la clasificación utilizando solo la parte léxica utilizando lematización de las palabras.

En este nuevo experimento el pesado que obtuvo un mejor desempeño fue el basado en frecuencia.

Comparando ambos resultados obtenidos podemos notar que los datos con lematización obtienen un mejor desempeño para la clase "Medicinal" y resultados similares para la clase "No medicinal".

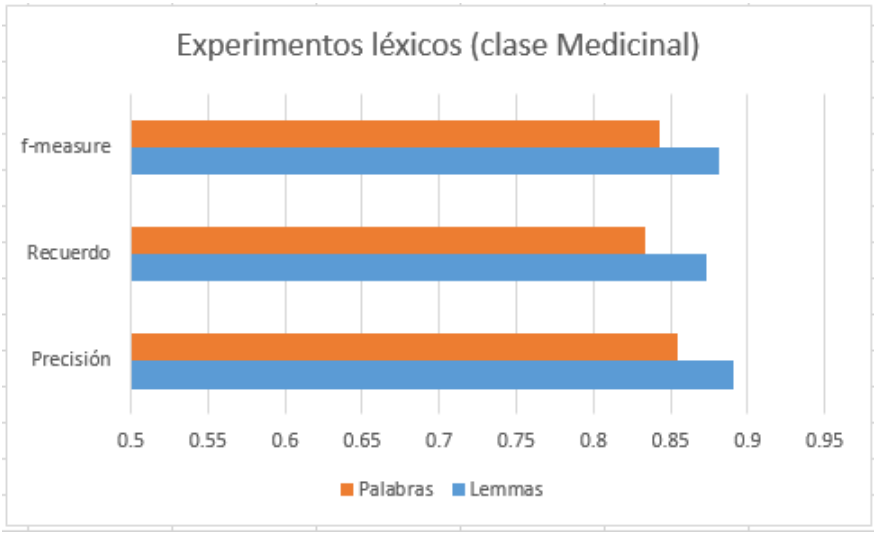

Figura 5.2: Comparación de la clase medicinal de ambos experimentos.

Además de la mejora en la clase medicinal para cada una de las tres medidas, se redujo la cantidad de atributos gracias al proceso de lematizado.

Las palabras que tienen mayor información mutua para la clase "medicinal" son las que se muestran en la figura 5.3 y las palabras con mayor información mutua para la clase "no medicinal" se muestran en la figura 5.4.

Figura 5.3: Palabras con mayor información mutua para la clase "medicinal".

Como se puede observar en la figura 5.3 las palabras relacionadas a la clase "medicinal" son aquellas propias del dominio medico. Palabras que describen enfermedades, síntomas, partes del cuerpo. Se observa que las palabras de esta clase, no se encuentran en la clase "no medicinal".

Figura 5.4: Palabras con mayor información mutua para la clase “no medicinal”

Para la clase “no medicinal” se incluyen palabras que describen principalmente países, palabras relacionadas a otros usos y localizaciones.

5.2.2. Experimento Sintáctico

Para el experimento sintáctico se utilizó el software de etiquetado de partes del habla provisto por Freeling, se busca representar las palabras por su categoría. De esta forma se puede generalizar las palabras y de esa manera encontrar los patrones sintácticos que son usados para describir usos medicinales de plantas.

Se optó por realizar experimentos utilizando los pesados de términos como en el experimento léxico, además de obtener los $n - gramas$ entre los rangos de $1 - 3$. Los mejores resultados obtenidos por cada $n - grama$ se muestra a continuación.

Nuevamente se muestran los resultados del clasificador SVM al ser el que obtuvo los resultados más altos. Los resultados mostrados en la tabla 5.6 consiste de las siguientes columnas:

- $n-grama$. uni-gramas(Experimento utilizando solo los uni-gramas), bi-gramas (experimento utilizando solo bi-gramas) y tri-gramas (experimento solo utilizando tri-gramas).
- Pesado. Indica el tipo de pesado utilizado en la representación. FT (pesado por frecuencia) y Binario (se encuentra o no el término).
- Atributos. Muestra el total de atributos de la representación.
- Clase. Indica la clase que se evaluó en la clasificación.
- Precisión. Valor obtenido de precisión por la clase, entre paréntesis la desviación estándar de los datos.
- Recuerdo. Valor obtenido de recuerdo por la clase, entre paréntesis la desviación estándar de los datos.
- F-measure. Valor obtenido de f-measure por la clase, entre paréntesis la desviación estándar de los datos.

$n-grama$	Pesado	Atributos	Clase	Precisión	Recuerdo	F-Measure
$uni-gramas$	FT	158				
			Medicinal	0.727 (0.034)	0.644 (0.284)	0.655 (0.181)
			No Medicinal	0.764 (0.042)	0.715 (0.195)	0.730 (0.125)
$bi-gramas$	FT	1923				
			Medicinal	0.736 (0.061)	0.695 (0.195)	0.708 (0.133)
			No Medicinal	0.0.816 (0.097)	0.837 (0.017)	0.824 (0.058)
$tri-gramas$	FT	7396				
			Medicinal	**0.782 (0.032)**	**0.729 (0.189)**	**0.746 (0.118)**
			No Medicinal	**0.927 (0.027**)	**0.927 (0.027)**	**0.927 (0.027)**

Tabla 5.6: Resultados de la clasificación utilizando información sintáctica mediante $n-gramas$ de palabras.

Como se puede observar en la tabla 5.6 los resultados más altos son obtenidos cuando los $n - grmas$ son ajustados a secuencias de 3 palabras. Esto es debido a que los $tri - gramas$ aportan mayor información para diferenciar entre clases de lo que lo hacen los $bi - gramas$ o $uni - gramas$.

En la tabla 5.7 se muestran los $tri - gramas$ mas discriminatorios en ambas clases, la tabla esta compuesta por las siguientes columnas:

- tri-grama. Tri-grama de etiquetas de parte del habla.
- Medicinal. Cantidad de veces que el tri-grama ocurrió en la clase "medicinal".
- No medicinal. Cantidad de veces que el tri-grama ocurrió en la clase "no medicinal".
- Ejemplo. Ejemplos léxicos de los tri-gramas.

tri-grama	Medicinal	No medicinal	Ejemplo
sp000 vmn0000 da0000	118	29	
			para aliviar los
			para controlar la
vmn0000 da0000 nc0p000’	63	20	
			aliviar los dolores
			aprovechar los beneficios
vsip000 vmp0000 sp000	60	8	
			es utilizada para
			es recomendado para
vsip000 di0000 nc0s000	16	193	
			es un arbol
			es una planta
di0000 nc0s000 aq0000	52	200	
			una hierba aromatica
			una planta originaria

Tabla 5.7: Tri-gramas mas significativos para ambas clases.

Se observa en la tabla 5.7 que los tri-gramas mas relevantes para la clase “medicinal” son aquellos que contienen verbos y preposiciones, mientras que los tri-gramas relevantes para la clase “no medicinal” tienen nombres comunes y adjetivos.

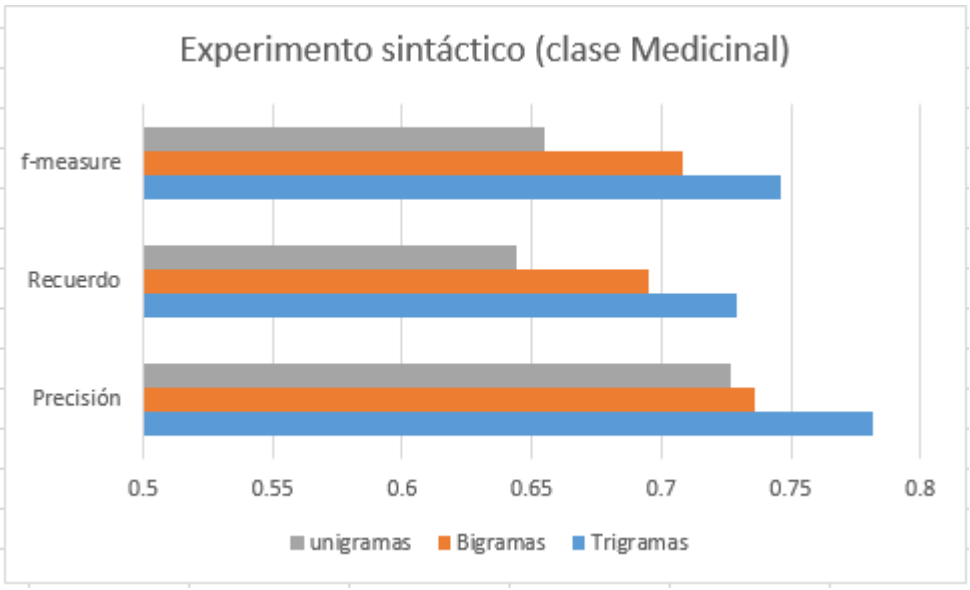

Figura 5.5: Experimento sintáctico utilizando $n - gramas$.

Para la clase medicinal se obtienen mejores resultados utilizando los $tri - gramas$ lamentablemente utilizando únicamente la información sintáctica extraída de las palabras no supera al experimento léxico. Se realizaron mas experimentos combinando los $n - gramas$, se combinaron unigramas, bigramas y trigramas, pero los resultados obtenidos no mejoraron el resultado obtenido utilizando únicamente $tri - gramas$ por lo que no se colocaron en esta sección, pero pueden ser consultados en los anexos.

5.2.3. Experimento Semántico

Como ya se ha indicado en anteriores capítulos, la información semántica es de utilidad para agregarle un dominio o temática a la información léxica.

Utilizando BabelNet como recurso semántico se obtienen hiperónimos de las palabras y así relacionar oraciones que antes no era posible.

En la tabla 5.8 se muestra el resultado de usar los hiperónimos en la clasifica-

ción, la tabla esta compuesta por las siguientes columnas:

- Pesado. Indica el tipo de pesado utilizado en la representación. FT (pesado por frecuencia) y Binario (se encuentra o no el término).
- Atributos. Muestra el total de atributos de la representación.
- Clase. Indica la clase que se evaluó en la clasificación.
- Precisión. Valor obtenido de precisión por la clase, entre paréntesis la desviación estándar de los datos.
- Recuerdo. Valor obtenido de recuerdo por la clase, entre paréntesis la desviación estándar de los datos.
- F-measure. Valor obtenido de f-measure por la clase, entre paréntesis la desviación estándar de los datos.

Pesado	Atributos	Clase	Precisión	Recuerdo	F-measure
FT	12981				
		Medicinal	**0.847 (0.018)**	**0.808 (0.128)**	**0.823 (0.076)**
		No medicinal	**0.938 (0.015)**	**0.901 (0.081)**	**0.917 (0.035)**
Binario	12981				
		Medicinal	0.806 (0.061)	0.795 (0.095)	0.800 (0.079)
		No medicinal	0.855 (0.055)	0.855 (0.055)	0.855 (0.055)

Tabla 5.8: Resultados obtenidos de la clasificación con hiperónimos de las palabras.

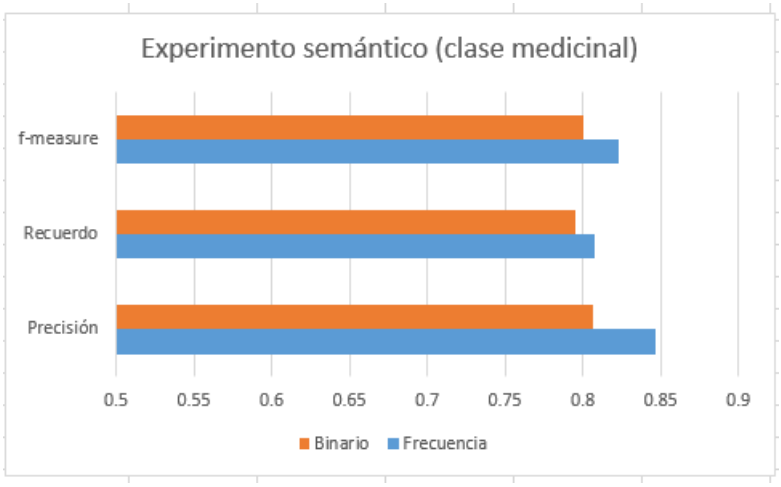

Figura 5.6: Resultados para la clase medicinal del experimento semántico

En la figura 5.6 se observa que la representación utilizando el pesado de frecuencia obtiene mejores resultados que utilizando el pesado binario.

5.2.4. Combinación de la información

En este último experimento para la clase medicinal lo que se hizo fue combinar la representación léxica, sintáctica y semántica para observar si estas combinaciones pueden superar al experimento base, ya que los resultados obtenidos por las representaciones sintáctica y semántica por si solos no pueden superar esos resultados.

En primera instancia se realizará lo que se conoce como "fusión temprana" lo cual consiste en unir los atributos de cada una de las representaciones en una sola matriz que será proporcionada al clasificador como se muestra en la figura 5.7.

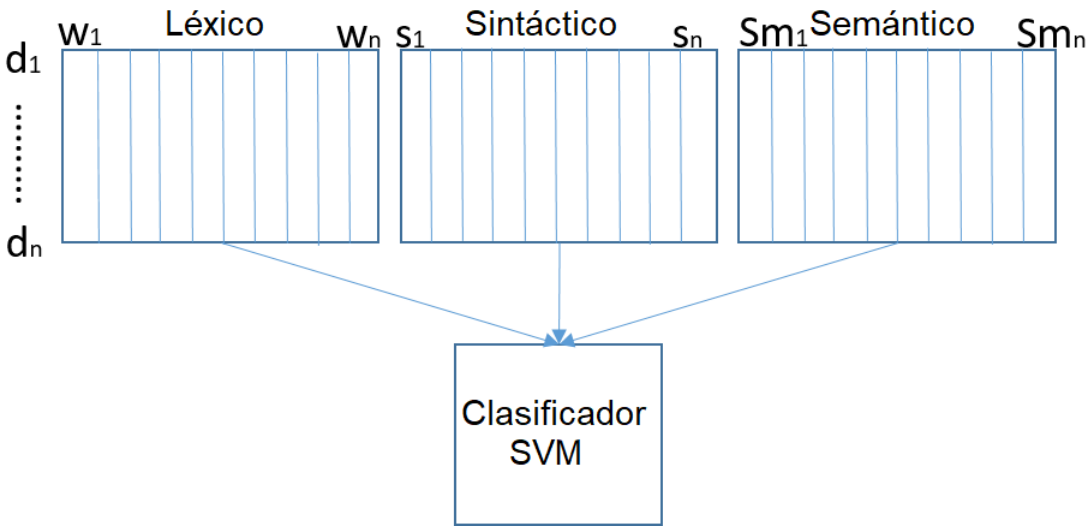

Figura 5.7: Combinación de representaciones mediante ”fusión temprana“.

Para realizar esta combinación de atributos, se eligieron las tres representaciones de la siguiente manera:

- Información léxica. Para esta representación se eligió el conjunto de datos que obtuvo el mejor resultado, este fue el conjunto con los datos que pasaron por el proceso de lematizado.
- Información Sintáctica. Se eligió la representación que consistió en $tri-gramas$ de etiquetas POS la cual obtuvo mejores resultados para la clase medicinal.
- Información Semántica. Para esta representación se optó por los hiperónimos con pesado de frecuencia ya que obtuvo los mejores resultados para la clase de interés.

Para la clasificación se utilizó la misma configuración que los experimentos anteriores, se hicieron las siguientes combinaciones de representaciones:

- Información léxica (L) + información sintáctica (P)

- información léxica(L) + información semántica (S)
- información léxica(L) + información sintáctica (P) + información semántica (S)

Los resultados obtenidos se muestran en las siguientes tablas:

En primer lugar la tabla 5.9 muestra los resultados obtenidos para la combinación de información léxica y sintáctica.

Tipo	Pesado	Atributos	Clase	Precisión	Recuerdo	F-measure
L+P	FT	5804				
			Medicinal	0.761 (0.106)	0.774 (0.054)	0.766 (0.080)
			No medicinal	**0.850 (0.100)**	**0.882 (0.018)**	**0.861 (0.043)**
	Binario	5804				
			Medicinal	**0.779 (0.128)**	**0.810 (0.010)**	**0.788 (0.062)**
			No medicinal	0.804 (0.115)	0.834 (0.006)	0.814 (0.057)

Tabla 5.9: Resultados obtenidos para la combinación de información léxica y sintáctica.

Combinando la información léxica y sintáctica no es suficiente para supurar los resultados obtenidos únicamente con la información léxica. En la tabla 5.10 se muestran los resultados de la combinación léxica y semántica.

Tipo	Pesado	Atributos	Clase	Precisión	Recuerdo	F-Measure
L+S	FT	5592				
			Medicinal	**0.950 (0.026)**	**0.905 (0.085)**	**0.924 (0.032)**
			No medicinal	**0.932 (0.015)**	**0.922 (0.042)**	**0.926 (0.028)**
	Binario	5592				
			Medicinal	0.908 (0.017)	0.856 (0.116)	0.876 (0.054)
			No medicinal	0.915 (0.047)	0.928 (0.008)	0.921 (0.028)

Tabla 5.10: Resultados de la combinación de la información léxica y semántica.

Con la combinación de la información léxica y semántica se logra superar al experimento básico, especialmente para la clase medicinal la cual es la de interés en esta investigación. En la tabla 5.11 se muestra el resultado de combinar los tres tipos de información en la clasificación.

Tipo	Pesado	Atributos	Clase	Precisión	Recuerdo	F-Measure
L+P+S	FT	5987				
			Medicinal	0.887 (0.033)	0.878 (0.058)	0.882 (0.046)
			No medicinal	**0.937 (0.002)**	**0.916 (0.056)**	**0.926 (0.030)**
	Binario	5987				
			Medicinal	**0.889 (0.046)**	**0.894 (0.034)**	**0.891 (0.040)**
			No medicinal	0.873 (0.062)	0.885 (0.025)	0.878 (0.043)

Tabla 5.11: Resultados de la combinación de la información léxica, sintáctica y semántica.

Al parecer el anexar la información sintáctica disminuye la efectividad del clasificador para la clase medicinal, lo que nos puede indicar que la estructura de las oraciones es muy similar para ambas clases.

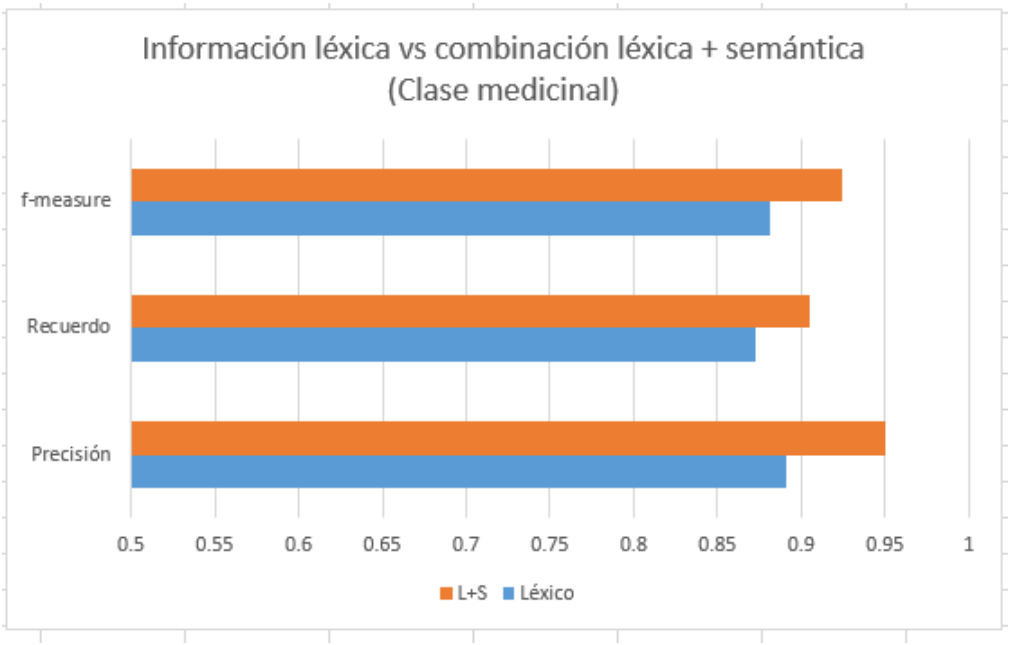

Figura 5.8: Comparativa de las combinaciones realizadas para clase medicinal.

Se puede observar en la figura 5.8 que para la clase medicinal, los mejores resultados son obtenidos por la combinación de la información semántica y léxica.

5.3. Experimento: reduciendo el conjunto de entrenamiento

En este experimento se busca encontrar la cantidad mínima de oraciones de entrenamiento con la que se obtengan resultados satisfactorios clasificando oraciones de uso medicinal.

Este experimento tiene como fundamento una aplicación realista en la cual se tenga poca información disponible para el entrenamiento del clasificador o solo se utilice la información necesaria y se ahorre tiempo en la parte de construcción del conjunto de entrenamiento.

Este experimento tiene la siguiente configuración:

- Pesado basado en frecuencias. Al ser el pesado con mejor desempeño en los experimentos anteriores.
- Clasificador SVM. De la misma manera fue el que obtuvo el mejor desempeño anteriormente.
- Se utilizaron los siguientes conjuntos de entrenamiento con mejor desempeño en la clasificación.
  - Léxico. Basado en las palabras lematizadas.
  - Sintáctico. Basado en $tri-gramas$ de etiquetas POS.
  - Semántico. Basado en hiperónimos.
  - Combinación de información. Basado en la información léxica y semántica.
- validación cruzada a 10 pliegues.
- Reducción de datos.
  - 100 %. 1000 oraciones medicinales, 1000 oraciones no medicinales.
  - 50 %. 500 oraciones medicinales, 500 oraciones no medicinales.
  - 25 %. 250 oraciones medicinales, 250 oraciones no medicinales.
  - 12 %. 125 oraciones medicinales, 125 oraciones no medicinales.
  - 6 %. 62 oraciones medicinales, 62 oraciones no medicinales.
  - 3 %. 31 oraciones medicinales, 31 oraciones no medicinales.
- El conjunto de prueba consistió de 200 oraciones de clase "Medicinal" y 200 oraciones de clase "No Medicinal" como en el experimento principal.

En las gráficas siguientes se muestran los resultados obtenidos con la reducción de datos. En la figura 5.9 se muestra el resultado de la reducción del conjunto de entrenamiento para la información léxica, se reporta el f-measure debido a que esta medida engloba tanto la precisión y el recuerdo.

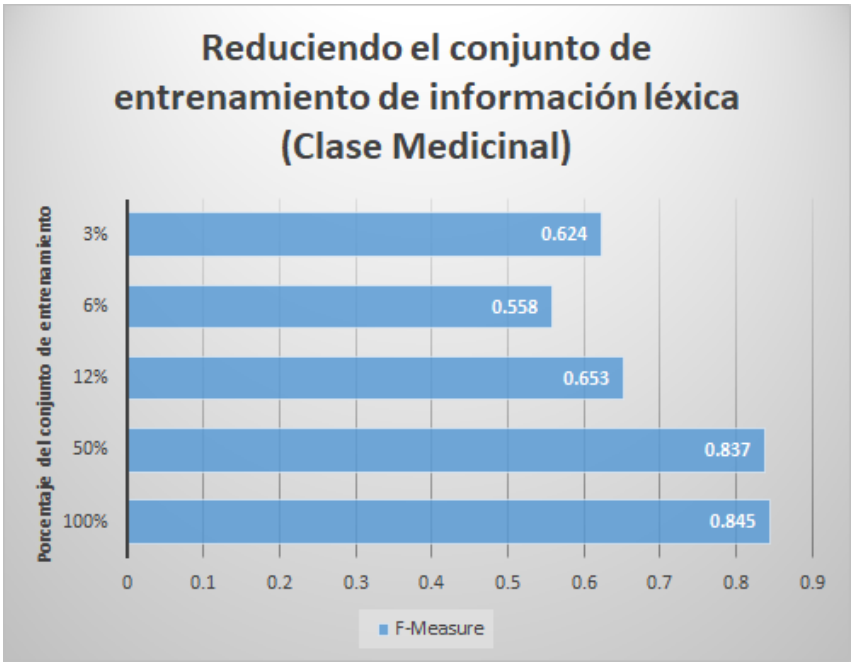

Figura 5.9: Reducción de datos de entrenamiento del experimento léxico.

usando el 6 % de los datos del conjunto de entrenamiento se obtuvo un f-measure de 0.558, mostrando una caída en la clasificación del 34 % respecto a usar el 100 % de los datos de entrenamiento. La clasificación se mantiene en resultados aceptables con el 50 % de los datos.

De manera similar en la figura 5.10 se muestran los resultados del experimento de reducción de el conjunto de entrenamiento para la información sintáctica.

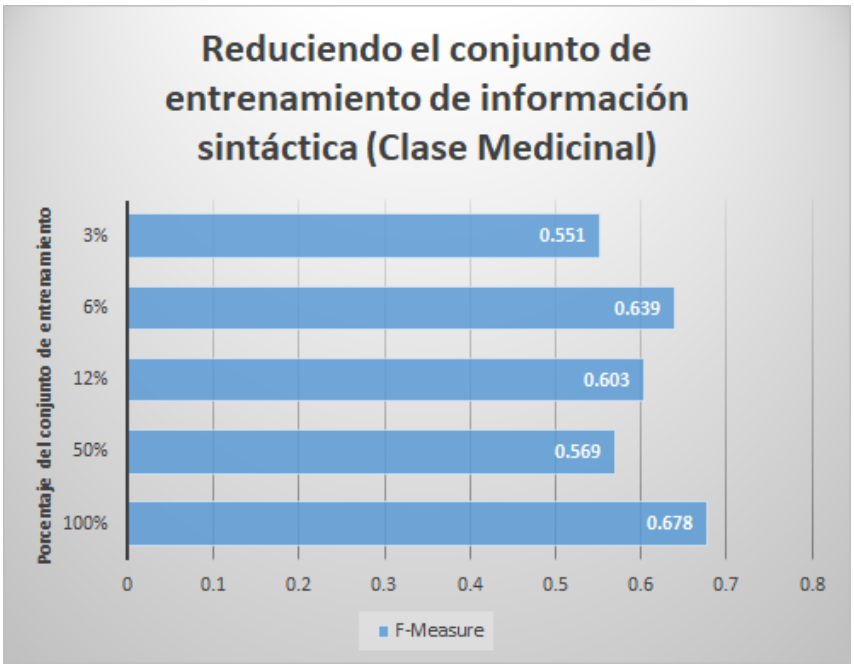

Figura 5.10: Reducción de los datos de entrenamiento para el experimento sintáctico

Para el experimento sintáctico los resultados fueron inconsistentes ya que en cantidades más reducidas se obtienen mejores resultados que con mayor cantidad de datos, siendo esta representación la que obtiene resultados inferiores a los del experimento base.

En el siguiente experimento se redujo el conjunto de entrenamiento para la información semántica, los resultados se pueden observar en la figura 5.11.

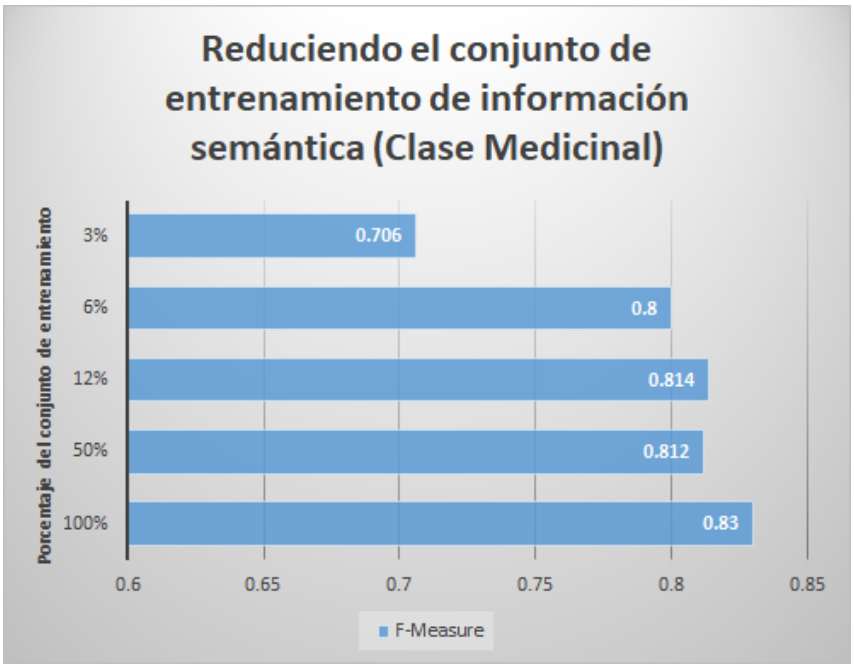

Figura 5.11: Reducción del conjunto de entrenamiento para la información semántica

La representación semántica es la más robusta respecto a la cantidad de información necesaria para entrenar el modelo, ya que utilizando un conjunto de entrenamiento reducido hasta el 6 % se obtienen resultados que alcanzan el 80 % de F-Measure.

Por último se presentan en la figura 5.12 los resultados obtenidos con la representación combinada de información léxica y semántica.

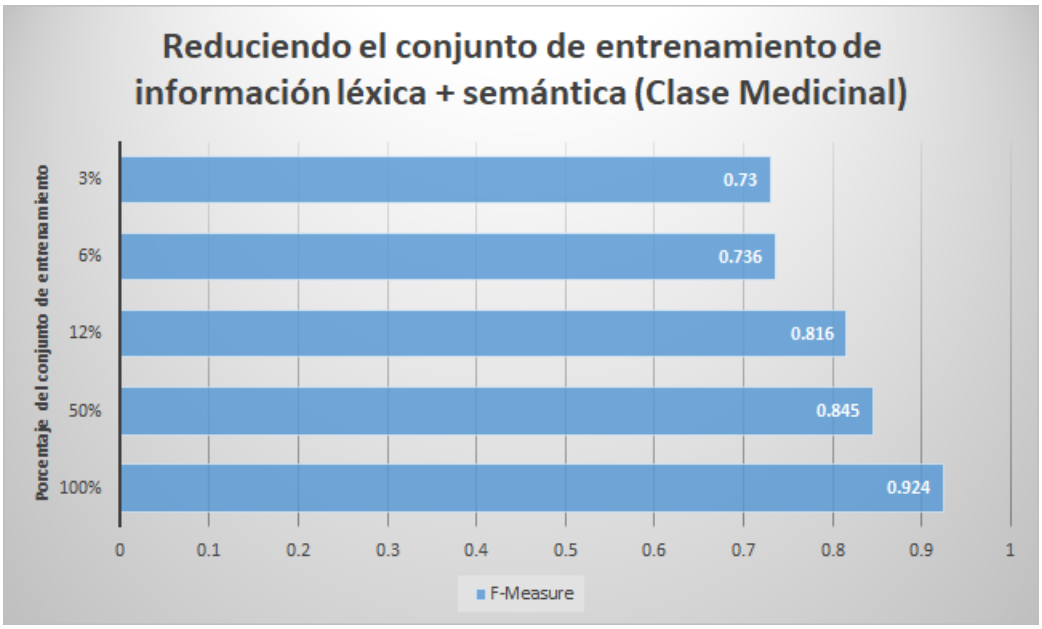

Figura 5.12: Reducción del conjunto de entrenamiento para la combinación de información léxica y semántica

De las combinaciones de información, la que obtuvo mejor desempeño fue aquella que contenía información léxica y semántica, con la reducción de datos de entrenamiento a un 12 % se obtienen resultados superiores al 80 % de F-Measure.

5.4. Experimentos para las otras clases

En este experimento se busca generalizar el método propuesto, realizando la clasificación para clases diferentes a la medicinal. Para realizar esto se tomaron las oraciones de la clase no medicinal, estas oraciones como se mencionó en la sección de creación del conjunto de datos, la clase no medicinal se compone de 3 tipos diferentes de oraciones.

- Otro uso. Oraciones donde se describan el uso de las plantas para un uso diferente al medicinal (culinario, industrial, construcción, etc).

- Localización. Oraciones donde se menciona el origen o lugar de crecimiento de alguna planta.
- Descripción. Oraciones donde se mencione a alguna planta ya sea de forma general o específica y no se ajuste a los dos criterios anteriores.

Cada una de estos tipos de oraciones se tomarán como la clase positiva y el resto de oraciones serán tomadas como la clase negativa.

Cabe aclarar que los hiperónimos están orientados a la clase medicinal y son los mismos utilizados en anteriores experimentos. Por lo que no están orientados a alguna de las clases en particular de estos experimentos.

Estos experimentos tienen la siguiente configuración:

- Pesado basado en frecuencias. Al ser el pesado con mejor desempeño en los experimentos anteriores.
- Clasificador SVM. De la misma manera fue el que obtuvo el mejor desempeño anteriormente.
- Se utilizaron los siguientes conjuntos de entrenamiento con mejor desempeño en la clasificación.
  - Léxico. Basado en las palabras lematizadas.
  - Semántico. Basado en hiperónimos.
  - Combinación de información. Basado en la información léxica y semántica.
- validación cruzada a 10 pliegues.

5.4.1. Clase "Descripción"

La clase descripción como se comentó son las oraciones que describen alguna información relacionada a alguna planta que no sea un uso medicinal, otro uso y no describa el origen o lugar de crecimiento de la planta.

Para esta clasificación solo se usarán las representaciones que tuvieron mejor desempeño: léxica, semántica y la combinación de ambas.

Los conjuntos de entrenamiento y prueba consistieron de la siguiente forma:

- Conjunto de entrenamiento. 1134 oraciones (408 oraciones de clase descripción y 726 de la clase negativa).
- Conjunto de prueba. 125 oraciones(45 oraciones de clase descripción y 80 de la clase negativa).

En la tabla 5.12 se muestran los resultados de esté experimento, las columnas corresponden como se detalla a continuación:

- Enfoque. Tipo de información que se utilizo para la clasificación.
- Clase. Clases que se evaluaron en la clasificación.
- Precisión. Valor obtenido por la clase para esta medida.
- Recuerdo.Valor obtenido por la clase para esta medida.
- F-measure.Valor obtenido por la clase para esta medida.

Enfoque	Clase	Precisión	Recuerdo	F-Measure
BOW	Descripción	0.767 (0.052)	0.758 (0.092)	0.762 (0.072)
	Oraciones negativas	0.800 (0.080)	0.812 (0.012)	0.805 (0.047)
Hiperónimos	Clase	Precisión	Recuerdo	F-Measure
	Descripción	**0.783 (0.074)**	**0.790 (0.035)**	**0.786 (0.055)**
	Oraciones negativas	**0.871 (0.030)**	**0.867 (0.045)**	**0.869 (0.038)**
Hiperónimos + palabras	Clase	Precisión	Recuerdo	F-Measure
	Descripción	0.724 (0.092)	0.732 (0.043)	0.727 (0.068)
	Oraciones negativas	0.826 (0.049)	0.826 (0.049)	0.826 (0.049)

Tabla 5.12: Resultados obtenidos para la clase “Descripción”.

Para la clase descripción los mejores resultados se obtienen mediante el uso únicamente de los hiperónimos, superando a las otras representaciones.

5.4.2. Clase “Localización”

La clase localización consiste de oraciones donde se habla de los orígenes y lugares de crecimiento de las plantas, es la clase que contiene una menor cantidad de oraciones.

Los conjuntos de entrenamiento y prueba consistieron de la siguiente forma:

- Conjunto de entrenamiento. 1133 oraciones (407 oraciones de clase descripción y 726 de la clase negativa).
- Conjunto de prueba. 125 oraciones(45 oraciones de clase descripción y 80 de la clase negativa).

En la tabla 5.13 se muestran los resultados de esté experimento, las columnas corresponden como se detalla a continuación:

- Enfoque. Tipo de información que se utilizo para la clasificación.
- Clase. Clases que se evaluaron en la clasificación.
- Precisión. Valor obtenido por la clase para esta medida.
- Recuerdo.Valor obtenido por la clase para esta medida.
- F-measure.Valor obtenido por la clase para esta medida.

Enfoque	Clase	Precisión	Recuerdo	F-Measure
BOW	Localización	**0.958 (0.024)**	**0.933 (0.058)**	**0.945 (0.042)**
	Oraciones negativas	**0.928 (0.019)**	**0.808 (0.183)**	**0.855 (0.114)**
Hiperónimos	Clase	Precisión	Recuerdo	F-Measure
	Localización	0.870 (0.085)	0.830 (0.142)	0.848 (0.115)
	Oraciones negativas	0.858 (0.089)	0.799 (0.174)	0.824 (0.135)
Hiperónimos + palabras	Clase	Precisión	Recuerdo	F-Measure
	Localización	0.903 (0.079)	0.924 (0.049)	0.913 (0.064)
	Oraciones negativas	0.901 (0.055)	0.835 (0.147)	0.863 (0.105)

Tabla 5.13: Resultados obtenidos para la clase “Localización”.

La mejor representación para esta clase es la que utiliza únicamente información léxica. Esto puede deberse a que esta clase es la que más se diferencia de las otras clases que componen la clase no medicinal y la clase medicinal por lo que la información semántica no aporta información que sea útil.

5.4.3. Clase “Otros usos”

La clase “‘otros usos” contiene oraciones que describen otros usos diferentes al medicinal, esta clase es la que está más cercana a la clase medicinal.

Los conjuntos de entrenamiento y prueba consistieron de la siguiente forma:

- Conjunto de entrenamiento. 1133 oraciones (407 oraciones de clase descripción y 726 de la clase negativa).
- Conjunto de prueba. 125 oraciones(45 oraciones de clase descripción y 80 de la clase negativa).

En la tabla 5.14 se muestran los resultados de esté experimento, las columnas corresponden como se detalla a continuación:

- Enfoque. Tipo de información que se utilizo para la clasificación.
- Clase. Clases que se evaluaron en la clasificación.
- Precisión. Valor obtenido por la clase para esta medida.
- Recuerdo.Valor obtenido por la clase para esta medida.
- F-measure.Valor obtenido por la clase para esta medida.

Enfoque	Clase	Precisión	Recuerdo	F-Measure
BOW	Otro uso	0.725 (0.075)	0.653 (0.270)	0.669 (0.188)
	Oraciones negativas	0.786 (0.047)	0.717 (0.217)	0.739 (0.142)
Hiperónimos	Clase	Precisión	Recuerdo	F-Measure
	Otro uso	0.749 (0.067)	0.682 (0.241)	0.701 (0.165)
	Oraciones negativas	**0.887 (0.002)**	**0.822 (0.145)**	**0.846 (0.080)**
Hiperónimos + palabras	Clase	Precisión	Recuerdo	F-Measure
	Descripción	**0.766 (0.052)**	**0.688 (0.246)**	**0.709 (0.163)**
	Oraciones negativas	0.804 (0.070)	0.780 (0.133)	0.790 (0.102)

Tabla 5.14: Resultados obtenidos para la clase “Otros usos”.

Los resultados muestran que utilizando la combinación de hiperónimos + palabras se obtienen los mejores resultados para la clase "otros usos". El hecho de que los hiperónimos estén orientados a la clase medicinal ayuda a esta clase, debido a que contienen oraciones similares en estructura.

5.5. Discusión

En esta sección se presentaron cada uno de los experimentos realizados en este trabajo. Estos experimentos se dividieron en 3.

- Experimentos para la clase medicinal.
- Experimentos reduciendo el conjunto de entrenamiento.
- Experimentos para otras clases.

El experimento principal buscando clasificar oraciones donde se describa el uso medicinal de una planta se partió de una base léxica que obtuvo un F-Measure de 0.881, se mejoró este resultado mediante la combinación de información léxica y semántica con un 0.924 de la misma medida hablando de la clase "Medicinal".

El segundo experimento tenía como propósito el definir la cantidad mínima del conjunto de entrenamiento con el cual la clasificación obtuviera resultados satisfactorios. Se realizaron experimentos para cada una de las representaciones y las combinaciones de estas, utilizando solo la parte léxica se necesita el 50 % de los datos que son alrededor de 1000 oraciones contando ambas clases. La información sintáctica obtuvo resultados irregulares con menor cantidad de datos ya que con 50 % de los datos obtuvo un f-measure de 0.569 mientras que con el 6 % obtuvo 0.639.

El experimento mas robusto fue aquel que utilizó la información semántica o hiperónimos, con esta representación se obtuvieron valores de F-Measure de 0.8

para el conjunto de entrenamiento reducido hasta solo utilizar el 6 %, alrededor de 30 oraciones de clase “medicinal” y 30 oraciones de clase “no medicinal”.

La combinación de información léxica y semántica fue la segunda más robusta alcanzando un 0.81 % de f-measure para el conjunto de entrenamiento con tan solo el 12 % de los datos.

Capítulo 6

Conclusiones y trabajo futuro

En este trabajo se abordó la tarea de clasificación de oraciones donde se describa un uso medicinal, para ello se hizo uso de la información léxica, sintáctica y semántica. Se realizaron varios experimentos con los cuales se buscó resolver esta tarea con cada una de las representaciones y con la combinación de éstas.

La tarea se abordó como un problema de clasificación de textos cortos debido a que las oraciones consisten en no más de 30 palabras. Al orientar este trabajo al idioma español se creó un conjunto de datos mediante la consulta a Internet de una lista de plantas. Esto se realizó con la intención de obtener oraciones que contuvieran palabras más del dominio público y menos científicas.

6.1. Conclusiones

Con la realización de este trabajo se puede concluir lo siguiente:

- Utilizando únicamente la información léxica se obtienen resultados aceptables para la cantidad de oraciones utilizadas en el conjunto de entrenamiento.

- La información sintáctica en forma de *trigramas* de etiquetas POS se creyó en un principio que obtendría mejores resultados debido a que se detectaron ciertos patrones en las oraciones principalmente en las oraciones medicinales, pero no fue el caso y esta representación por si sola fue la que obtuvo los resultados más bajos de todos los experimentos.

- La información semántica agregada consistió en el hiperónimo directo encontrado en BabelNet, por si solo obtuvo resultados cercanos al experimento base.

- de las combinaciones realizadas la que obtuvo mejores resultados y supero al experimento base fue la compuesta por la información léxica y semántica. Desafortunadamente la información sintáctica no aportó información útil para este trabajo lo que redujo los resultados de las combinaciones donde esta intervino, como lo fue combinada con la información léxica y la combinación de las tres representaciones.

- En el experimento de reducción del conjunto de entrenamiento se pudo observar que el reducir 50 % los datos de entrenamiento es suficiente para obtener resultados favorables con la representación léxica, la información sintáctica aun con el 100 % de los datos obtiene resultados bajos por lo que el reducir el conjunto de entrenamiento reduce aún más los resultados, por otro lado la representación semántica es la representación más estable ya que aun reduciendo a 6 % los datos de entrenamiento se obtienen resultados superiores a 0.8 para la medida de F-Measure. La combinación de información léxica y semántica es la segunda más estable obteniendo valores similares, pero con el 12 % del conjunto de entrenamiento.

- El ultimo experimento consistió en observar cómo se comportaba el método para otras clases. Para realizar esto se utilizaron las oraciones que conforman la clase "No medicinal", ya que estas oraciones están divididas por tres tipos: localización, otros usos y descripción. Los resultados obtenidos nos demuestran

que aun utilizando información semántica relacionada a la clase medicinal se obtienen resultados satisfactorios para 2 de los tipos de oraciones (Descripción y otros usos), no así para la clase localización que obtuvo un mejor resultado utilizando la representación léxica. Debido a que esta clase es la que contiene oraciones muy diferentes a las de la clase medicinal.

6.2. Trabajo a futuro

Con los resultados vistos en este trabajo se propone el siguiente trabajo a futuro:

- Incrementar el conjunto de datos de entrenamiento y prueba.
- Obtener información semántica de diferente manera, para ello se puede utilizar otro recurso semántico.
- Aplicar el método propuesto para otro tipo de colecciones de datos y dominios diferentes.
- Aplicar métodos semi-supervisados para ir enriqueciendo el conjunto de datos de entrenamiento con nuevas oraciones que ya hayan sido clasificadas correctamente como medicinales o no.

Apéndice A

Tablas de Resultados

Datos del Conjunto de entrenamiento

- 1000 oraciones de clase medicinal
- 1000 oraciones de clase no medicinal

Datos del Conjunto de Prueba

- 200 oraciones de clase medicinal
- 200 oraciones de clase no medicinal

Pesado de términos

- Pesado basado en frecuencia (TF)
- Pesado binario (binario)

Los resultados fueron obtenidos mediante el clasificador de maquinas de soporte vectorial (SVM).

A.1. Tablas de resultados para la clase medicinal

En esta sección se muestran los resultados obtenidos en la clasificación para la clase "Medicinal".

Experimento Léxico

Tipo	Pesado	Atributos	Clase	Precisión	Recuerdo	F-measure
Palabras	FT	5409				
			Medicinal	0.845 (0.041)	0.829 (0.089)	0.836 (0.065)
			No medicinal	0.913 (0.058)	0.934 (0.006)	0.922 (0.027)
	Binario	5409				
			Medicinal	**0.855 (0.032)**	**0.834 (0.094)**	**0.843 (0.064)**
			No medicinal	**0.913 (0.058)**	**0.934 (0.006)**	**0.922 (0.027)**
Lemmas	FT	3900				
			Medicinal	**0.891 (0.021)**	**0.873 (0.073)**	**0.881 (0.048)**
			No medicinal	**0.927 (0.027)**	**0.927 (0.027)**	**0.927 (0.027)**
	Binario	3900				
			Medicinal	0.864 (0.010)	0.823 (0.123)	0.839 (0.070)
			No medicinal	0.910 (0.044)	0.918 (0.018)	0.914 (0.031)

Tabla A.1: Resultados obtenidos para la información léxica.

Experimento sintáctico

- 1-2gramas. Combinación de unigramas y bigramas de etiquetas POS.
- 1-2-3gramas. Combinación de unigramas, bigramas y trigramas.

Tipo	Pesado	Atributos	Clase	Precisión	Recuerdo	F-measure
Unigramas de POS	FT	158				
			Medicinal	0.727 (0.034)	0.644 (0.284)	0.655 (0.181)
			No medicinal	0.764 (0.042)	0.715 (0.195)	0.730 (0.125)
	Binario	158				
			Medicinal	0.595 (0.128)	0.567 (0.287)	0.567 (0.217)
			No medicinal	0.742 (0.067)	0.711 (0.171)	0.722 (0.122)
Bigramas de POS	FT	1923				
			Medicinal	0.736 (0.061)	0.695 (0.195)	0.708 (0.133)
			No medicinal	0.816 (0.097)	0.837 (0.017)	0.824 (0.058)
	Binario	1923				
			Medicinal	0.723 (0.056)	0.670 (0.230)	0.683 (0.153)
			No medicinal	0.864 (0.024)	0.838 (0.098)	0.849 (0.062)
Trigramas de POS	FT	7396				
			Medicinal	**0.782 (0.032)**	**0.729 (0.189)**	**0.746 (0.118)**
			No medicinal	**0.927 (0.027)**	**0.927 (0.027)**	**0.927 (0.027)**
	Binario	7326				
			Medicinal	0.771 (0.042)	0.725 (0.185)	0.740 (0.119)
			No medicinal	0.913 (0.033)	0.913 (0.033)	0.913 (0.033)
1-2gramas	FT	2081				
			Medicinal	0.758 (0.058)	0.725 (0.165)	0.737 (0.115)
			No medicinal	0.823 (0.091)	0.842 (0.022)	0.831 (0.057)
	Binario	2081				
			Medicinal	0.734 (0.047)	0.675 (0.235)	0.688 (0.152)
			No medicinal	0.859 (0.022)	0.828 (0.108)	0.841 (0.067)
1-2-3gramas	FT	9478				
			Medicinal	0.754 (0.035)	0.689 (0.229)	0.705 (0.144)
			No medicinal	0.810 (0.079)	0.816 (0.056)	0.813 (0.068)
	Binario					
			Medicinal	0.744 (0.029)	0.664 (0.264)	0.678 (0.165)
			No medicinal	0.817 0.081)	0.826 (0.046)	0.821 (0.064)

Tabla A.2: Resultados obtenidos para el experimento sintáctico.

Experimento semántico

- 1-2gramas. Combinacion de unigramas y bigramas de Hiperónimos.
- 1-2-3gramas. Combinacion de unigramas, bigramas y Hiperónimos.

Tipo	Pesado	Atributos	Clase	Precisión	Recuerdo	F-measure
Hiperónimos	FT	274				
			Medicinal	0.840 (0.060)	0.840 (0.060)	0.840 (0.060)
			No medicinal	0.864 (0.024)	0.838 (0.098)	0.849 (0.062)
	Binario	274				
			Medicinal	0.849 (0.031)	0.824 (0.104)	0.834 (0.068)
			No medicinal	0.906 (0.015)	0.887 (0.067)	0.896 (0.042)
Bigramas de hiperónimos	FT	5724				
			Medicinal	0.845 (0.041)	0.829 (0.089)	0.836 (0.065)
			No medicinal	0.902 (0.027)	0.893 (0.053)	0.897 (0.040)
	Binario	5724				
			Medicinal	0.836 (0.049)	0.825 (0.085)	0.830 (0.067)
			No medicinal	0.904 (0.041)	0.908 (0.028)	0.906 (0.035)
Trigramas de hiperónimos	FT	12981				
			Medicinal	**0.847 (0.018)**	**0.808 (0.128)**	**0.823 (0.076)**
			No medicinal	**0.938 (0.015)**	**0.901 (0.081)**	**0.917 (0.035)**
	Binario	12981				
			Medicinal	0.806 (0.061)	0.795 (0.095)	0.800 (0.079)
			No medicinal	0.855 (0.055)	0.855 (0.055)	0.855 (0.055)
1-2gramas	FT	5999				
			Medicinal	0.847 (0.018)	0.808 (0.128)	0.823 (0.076)
			No medicinal	0.938 (0.015)	0.901 (0.081)	0.917 (0.035)
	Binario	5999				
			Medicinal	0.830 (0.047)	0.815 (0.095)	0.821 (0.071)
			No medicinal	0.927 (0.027)	0.927 (0.027)	0.927 (0.027)
1-2-3gramas	FT	18980				
			Medicinal	0.847 (0.018)	0.808 (0.128)	0.823 (0.076)
			No medicinal	0.938 (0.015)	0.901 (0.081)	0.917 (0.035)
	Binario	18980				
			Medicinal	0.853 (0.020)	0.818 (0.118)	0.832 (0.071)
			No medicinal	0.872 (0.039)	0.864 (0.064)	0.868 (0.051)

Tabla A.3: Resultados del experimento semántico.

Experimento de combinación de representaciones

- L+P. Combinación léxica y sintáctica.
- L+S. Combinación léxica y semántica.
- L+P+S. Combinación léxica, sintáctica y semántica.

Tipo	Pesado	Atributos	Clase	Precisión	Recuerdo	F-measure
L+P	FT	5804				
			Medicinal	0.761 (0.106)	0.774 (0.054)	0.766 (0.080)
			No medicinal	0.850 (0.100)	0.882 (0.018)	0.861 (0.043)
	Binario	5804				
			Medicinal	0.779 (0.128)	0.810 (0.010)	0.788 (0.062)
			No medicinal	0.804 (0.115)	0.834 (0.006)	0.814 (0.057)
L+S	FT	5592				
			Medicinal	**0.950 (0.026)**	**0.905 (0.085)**	**0.924 (0.032)**
			No medicinal	**0.932 (0.015)**	**0.922 (0.042)**	**0.926 (0.028)**
	Binario	5592				
			Medicinal	0.908 (0.017)	0.856 (0.116)	0.876 (0.054)
			No medicinal	0.915 (0.047)	0.928 (0.008)	0.921 (0.028)
L+P+S	FT	5987				
			Medicinal	0.887 (0.033)	0.878 (0.058)	0.882 (0.046)
			No medicinal	0.937 (0.002)	0.916 (0.056)	0.926 (0.030)
	Binario	5987				
			Medicinal	0.889 (0.046)	0.894 (0.034)	0.891 (0.040)
			No medicinal	0.873 (0.062)	0.885 (0.025)	0.878 (0.043)

Tabla A.4: Resultados del experimento de combinación de representaciones.

A.2. Tablas de resultado del experimento de reducción del conjunto de entrenamiento

Información léxica

Datos	N. Oraciones	Clase	Precisión	Recuerdo	F-Measure
100 %	450	Medicinal	0.843 (0.010)	0.767 (0.187)	0.790 (0.100)
	990	No medicinal	0.937 (0.018)	0.932 (0.032)	0.934 (0.025)
50 %	225	Medicinal	0.826 (0.012)	0.737 (0.217)	0.760 (0.118)
	495	No medicinal	0.942 (0.005)	0.926 (0.046)	0.933 (0.026)
25 %	112	Medicinal	0.741 (0.021)	0.648 (0.288)	0.660 (0.180)
	247	No medicinal	0.911 (0.018)	0.897 (0.057)	0.904 (0.038)
12 %	56	Medicinal	0.780 (0.002)	0.683 (0.263)	0.701 (0.155)
	123	No medicinal	0.865 (0.004)	0.807 (0.147)	0.828 (0.078)
6 %	28	Medicinal	0.693 (0.043)	0.598 (0.338)	0.598 (0.226)
	61	No Medicinal	0.754 (0.051)	0.710 (0.190)	0.724 (0.126)
3 %	14	Medicinal	0.707 (0.014)	0.572 (0.392)	0.555 (0.270)
	30	No Medicinal	0.757 (0.007)	0.653 (0.293)	0.666 (0.180)

Tabla A.5: Resultados para la representación léxica utilizando solo palabras.

Datos	N. Oraciones	Clase	Precisión	Recuerdo	F-Measure
100 %	450	Medicinal	0.875 (0.000)	0.827 (0.127)	0.845 (0.068)
	990	No medicinal	0.927 (0.027)	0.927 (0.027)	0.927 (0.027)
50 %	225	Medicinal	0.870 (0.002)	0.817 (0.137)	0.837 (0.073)
	495	No medicinal	0.946 (0.009)	0.936 (0.036)	0.941 (0.023)
25 %	112	Medicinal	0.787 (0.001)	0.693 (0.253)	0.712 (0.148)
	247	No medicinal	0.921 (0.008)	0.902 (0.062)	0.911 (0.036)
12 %	56	Medicinal	0.779 (0.021)	0.642 (0.322)	0.653 (0.195)
	123	No medicinal	0.899 (0.042)	0.811 (0.171)	0.839 (0.077)
6 %	28	Medicinal	0.862 (0.138)	0.580 (0.420)	0.558 (0.282)
	61	No medicinal	0.922 (0.050)	0.835 (0.155)	0.864 (0.064)
3 %	14	Medicinal	0.714 (0.032)	0.618 (0.318)	0.624 (0.207)
	30	No medicinal	0.830 (0.027)	0.722 (0.242)	0.746 (0.130)

Tabla A.6: Resultados de la representación léxica utilizando palabras lematizadas.

Información sintáctica

Datos	N. Oraciones	Clase	Precisión	Recuerdo	F-Measure
100 %	450	Medicinal	0.744 (0.029)	0.664 (0.264)	0.678 (0.165)
	990	No medicinal	0.859 (0.006)	0.797 (0.157)	0.818 (0.083)
50 %	225	Medicinal	0.612 (0.112)	0.571 (0.311)	0.569 (0.226)
	495	No medicinal	0.841 (0.026)	0.742 (0.222)	0.767 (0.117)
25 %	112	Medicinal	0.658 (0.081)	0.600 (0.300)	0.603 (0.208)
	247	No medicinal	0.770 (0.066)	0.751 (0.131)	0.759 (0.099)
12 %	56	Medicinal	0.697 (0.064)	0.640 (0.260)	0.650 (0.175)
	123	No medicinal	0.782 (0.032)	0.729 (0.189)	0.746 (0.118)
6 %	28	Medicinal	0.688 (0.068)	0.630 (0.270)	0.639 (0.183)
	61	No medicinal	0.684 (0.084)	0.646 (0.226)	0.656 (0.161)
3 %	14	Medicinal	0.589 (0.127)	0.556 (0.316)	0.551 (0.236)
	30	No medicinal	0.658 (0.081)	0.600 (0.300)	0.603 (0.208)

Tabla A.7: Resultado obtenidos con la información sintáctica.

Datos	N. Oraciones	Clase	Precisión	Recuerdo	F-Measure
100 %	450	Medicinal	0.858 (0.008)	0.813 (0.133)	0.830 (0.074)
	990	No medicinal	0.947 (0.008)	0.921 (0.061)	0.933 (0.027)
50 %	225	Medicinal	0.847 (0.005)	0.793 (0.153)	0.812 (0.085)
	495	No medicinal	0.927 (0.005)	0.896 (0.076)	0.910 (0.037)
25 %	112	Medicinal	0.842 (0.017)	0.798 (0.138)	0.814 (0.081)
	247	No medicinal	0.902 (0.003)	0.862 (0.102)	0.878 (0.052)
12 %	56	Medicinal	0.849 (0.009)	0.777 (0.177)	0.800 (0.094)
	123	No medicinal	0.862 (0.021)	0.782 (0.182)	0.806 (0.092)
6 %	28	Medicinal	0.726 (0.049)	0.665 (0.245)	0.678 (0.159)
	61	No medicinal	0.862 (0.038)	0.756 (0.216)	0.783 (0.108)
3 %	14	Medicinal	0.772 (0.014)	0.688 (0.248)	0.706 (0.149)
	30	No Medicinal	0.760 (0.093)	0.763 (0.083)	0.761 (0.088)

Tabla A.8: Resultados de la información semántica.

Datos	N. Oraciones	Clase	Precisión	Recuerdo	F-Measure
100 %	450	Medicinal	0.761 (0.106)	0.774 (0.054)	0.766 (0.080)
	990	No medicinal	0.850 (0.100)	0.882 (0.018)	0.861 (0.043)
50 %	225	Medicinal	0.723 (0.098)	0.718 (0.118)	0.721 (0.108)
	495	No medicinal	0.818 (0.120)	0.854 (0.026)	0.829 (0.050)
25 %	112	Medicinal	0.731 (0.086)	0.717 (0.137)	0.723 (0.113)
	247	No medicinal	0.782 (0.103)	0.798 (0.038)	0.788 (0.071)
12 %	56	Medicinal	0.592 (0.142)	0.580 (0.220)	0.583 (0.183)
	123	No medicinal	0.774 (0.088)	0.777 (0.077)	0.776 (0.083)
6 %	28	Medicinal	0.646 (0.091)	0.595 (0.295)	0.598 (0.208)
	61	No medicinal	0.789 (0.032)	0.739 (0.179)	0.755 (0.112)
3 %	14	Medicinal	0.668 (0.076)	0.610 (0.290)	0.615 (0.200)
	30	No Medicinal	0.741 (0.021)	0.648 (0.288)	0.660 (0.180)

Tabla A.9: Resultados de la combinación de información léxica y sintáctica.

Datos	N. Oraciones	Clase	Precisión	Recuerdo	F-Measure
100 %	450	Medicinal	0.950 (0.026)	0.905 (0.085)	0.924 (0.032)
	990	No medicinal	0.932 (0.015)	0.922 (0.042)	0.926 (0.028)
50 %	225	Medicinal	0.875 (0.000)	0.827 (0.127)	0.845 (0.068)
	495	No medicinal	0.921 (0.008)	0.902 (0.062)	0.911 (0.036)
25 %	112	Medicinal	0.847 (0.005)	0.793 (0.153)	0.812 (0.085)
	247	No medicinal	0.927 (0.005)	0.896 (0.076)	0.910 (0.037)
12 %	56	Medicinal	0.867 (0.019)	0.792 (0.172)	0.816 (0.086)
	123	No medicinal	0.897 (0.006)	0.852 (0.112)	0.869 (0.056)
6 %	28	Medicinal	0.775 (0.033)	0.719 (0.199)	0.736 (0.124)
	61	No medicinal	0.836 (0.015)	0.788 (0.148)	0.805 (0.086)
3 %	14	Medicinal	0.764 (0.042)	0.715 (0.195)	0.730 (0.125)
	30	No Medicinal	0.761 (0.050)	0.720 (0.180)	0.734 (0.120)

Tabla A.10: Resultados de la combinación de información léxica y semántica.

Datos	N. Oraciones	Clase	Precisión	Recuerdo	F-Measure
100 %	450	Medicinal	0.881 (0.030)	0.868 (0.068)	0.874 (0.050)
	990	No medicinal	0.942 (0.005)	0.926 (0.046)	0.933 (0.026)
50 %	225	Medicinal	0.875 (0.014)	0.843 (0.103)	0.856 (0.060)
	495	No medicinal	0.910 (0.044)	0.918 (0.018)	0.914 (0.031)
25 %	112	Medicinal	0.807 (0.023)	0.754 (0.174)	0.771 (0.104)
	247	No medicinal	0.867 (0.059)	0.875 (0.035)	0.870 (0.047)
12 %	56	Medicinal	0.858 (0.008)	0.813 (0.133)	0.830 (0.074)
	123	No medicinal	0.839 (0.039)	0.819 (0.099)	0.828 (0.070)
6 %	28	Medicinal	0.721 (0.063)	0.675 (0.215)	0.688 (0.146)
	61	No medicinal	0.893 (0.035)	0.888 (0.048)	0.890 (0.042)
3 %	14	Medicinal	0.744 (0.029)	0.664 (0.264)	0.678 (0.165)
	30	No Medicinal	0.841 (0.026)	0.742 (0.222)	0.767 (0.117)

Tabla A.11: Resultados de la combinación de información léxica, sintáctica y semántica.

A.2.1. Tablas de resultados del experimento de clasificación de otras clases

Tipo de Oración	Cantidad de Oraciones
Otro uso	343
Descripción	453
Localización	162

Tabla A.12: Número de oraciones por clase.

Clase "Otros usos"

Oraciones de clase "otro uso"

- Training. 1133 oraciones (309 oraciones de otro uso, 824 oraciones negativas).
- Test. 125 oraciones (34 oraciones de otro uso, 91 oraciones negativas)

Enfoque	Clase	Precisión	Recuerdo	F-Measure
BOW	Otro uso	0.725 (0.075)	0.653 (0.270)	0.669 (0.188)
	Oraciones negativas	0.786 (0.047)	0.717 (0.217)	0.739 (0.142)
Hiperónimos	Clase	Precisión	Recuerdo	F-Measure
	Otro uso	0.749 (0.067)	0.682 (0.241)	0.701 (0.165)
	Oraciones negativas	**0.887 (0.002)**	**0.822 (0.145)**	**0.846 (0.080)**
Hiperónimos + palabras	Clase	Precisión	Recuerdo	F-Measure
	Descripción	**0.766 (0.052)**	**0.688 (0.246)**	**0.709 (0.163)**
	Oraciones negativas	0.804 (0.070)	0.780 (0.133)	0.790 (0.102)

Tabla A.13: Resultados de la clasificación de la clase Otros usos.

Clase "Descripción"

Oraciones de clase "Descripción"

- Training. (408 oraciones de descripción 726 oraciones negativas).
- Test. (45 oraciones de descripción, 80 oraciones negativas).

Enfoque	Clase	Precisión	Recuerdo	F-Measure
BOW	Descripción	0.767 (0.052)	0.758 (0.092)	0.762 (0.072)
	Oraciones negativas	0.800 (0.080)	0.812 (0.012)	0.805 (0.047)
Hiperónimos	Clase	Precisión	Recuerdo	F-Measure
	Descripción	0.783 (0.074)	0.790 (0.035)	0.786 (0.055)
	Oraciones negativas	0.871 (0.030)	0.867 (0.045)	0.869 (0.038)
Hiperónimos + palabras	Clase	Precisión	Recuerdo	F-Measure
	Descripción	0.724 (0.092)	0.732 (0.043)	0.727 (0.068)
	Oraciones negativas	0.826 (0.049)	0.826 (0.049)	0.826 (0.049)

Tabla A.14: Resultados de la clasificación de la clase Descripción.

Clase "Localización"

Oraciones de clase "Localización"

- Training. (146 oraciones de localización, 987 oraciones negativas).
- Test. (17 oraciones de localización, 109 oraciones negativas).

Enfoque	Clase	Precisión	Recuerdo	F-Measure
BOW	Localización	0.958 (0.024)	0.933 (0.058)	0.945 (0.042)
	Oraciones negativas	0.928 (0.019)	0.808 (0.183)	0.855 (0.114)
Hiperónimos	Clase	Precisión	Recuerdo	F-Measure
	Localización	0.870 (0.085)	0.830 (0.142)	0.848 (0.115)
	Oraciones negativas	0.858 (0.089)	0.799 (0.174)	0.824 (0.135)
Hiperónimos + palabras	Clase	Precisión	Recuerdo	F-Measure
	Localización	0.903 (0.079)	0.924 (0.049)	0.913 (0.064)
	Oraciones negativas	0.901 (0.055)	0.835 (0.147)	0.863 (0.105)

Tabla A.15: Resultados de la clasificación de la clase Localización.

Bibliografía

[Barguil et al., 2016] Barguil, Suarez, Rueda, Ramos, Reguero, González, and Barreto (2016). Bioprospectus: Biodiversity data integration and search to support bioprospecting of the industrial uses of plants.

[Batool et al., 2013] Batool, R., Khattak, A. M., Maqbool, J., and Lee, S. (2013). Precise tweet classification and sentiment analysis. In *2013 IEEE/ACIS 12th International Conference on Computer and Information Science (ICIS)*, pages 461–466.

[Béchara et al., 2015] Béchara, H., Costa, H., Taslimipoora, S., Guptaa, R., Orasana, C., Pastorb, G. C., and Mitkova, R. (2015). Miniexperts: An svm approach for measuring semantic textual similarity. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 96–101.

[Dai et al., 2006] Dai, H. K., Zhao, L., Nie, Z., Wen, J.-R., Wang, L., and Li, Y. (2006). Detecting online commercial intention (oci). In *Proceedings of the 15th international conference on World Wide Web*, pages 829–837. ACM.

[Fellbaum, 1998] Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Bradford Books.

[Ferrando et al., 2016] Ferrando, A., Beux, S., Mascardi, V., and Rosso, P. (2016). Identification of disease symptoms in multilingual sentences: an ontology driven-

approach. In *ECIR 2016 Workshop on Modeling, Learning and Mining for Cross/Multilinguality (MultiLingMine), Padua, Italy*, pages 6–15.

[Franco-Salvador et al., 2012] Franco-Salvador, M., Gupta, P., and Rosso, P. (2012). Detección de plagio translingüe utilizando el diccionario estadístico de babelnet. *Computación y Sistemas*, 16(4):383–390.

[Franco-Salvador et al., 2016] Franco-Salvador, M., Kar, S., Solorio, T., and Rosso, P. (2016). Uh-prhlt at semeval-2016 task 3: Combining lexical and semantic-based features for community question answering. *Proceedings of SemEval*, 16:814–821.

[Gutiérrez et al., 2013] Gutiérrez, Y., Castaneda, Y., González, A., Estrada, R., Piug, D. D., Abreu, J. I., Pérez, R., Fernández Orquín, A., Montoyo, A., Munoz, R., et al. (2013). Umcc_dlsi: reinforcing a ranking algorithm with sense frequencies and multidimensional semantic resources to solve multilingual word sense disambiguation. Association for Computational Linguistics.

[Harish et al., 2010] Harish, B. S., Guru, D. S., and Manjunath, S. (2010). Representation and classification of text documents: A brief review. *IJCA, Special Issue on RTIPPR (2)*, pages 110–119.

[Hotho et al., 2005] Hotho, A., Nürnberger, A., and Paaß, G. (2005). A brief survey of text mining. In *Ldv Forum*, volume 20, pages 19–62.

[Iroju and Olaleke, 2015] Iroju, O. G. and Olaleke, J. O. (2015). A systematic review of natural language processing in healthcare. *International Journal of Information Technology and Computer Science (IJITCS)*, 7(8):44.

[Islam and Inkpen, 2008] Islam, A. and Inkpen, D. (2008). Semantic text similarity using corpus-based word similarity and string similarity. *ACM Trans. Knowl. Discov. Data*, 2(2):10:1–10:25.

[Islam et al., 2012] Islam, A., Milios, E., and Kešelj, V. (2012). *Text Similarity Using Google Tri-grams*, pages 312–317. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Jain and Pise, 2015] Jain, R. and Pise, N. (2015). Feature selection for effective text classification using semantic information. *International Journal of Computer Applications*, 113(10).

[Jensen et al., 2014] Jensen, K., Panagiotou, G., and Kouskoumvekaki, I. (2014). Correction: Integrated text mining and chemoinformatics analysis associates diet to health benefit at molecular level. *PLoS computational biology*, 10(1).

[jin Tang et al., 2013] jin Tang, H., feng Yan, D., and Tian, Y. (2013). Semantic dictionary based method for short text classification. *The Journal of China Universities of Posts and Telecommunications*, 20:15 – 19.

[Joachims, 1998] Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *Machine learning: ECML-98*, pages 137–142.

[Kenter and de Rijke, 2015] Kenter, T. and de Rijke, M. (2015). Short text similarity with word embeddings. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 1411–1420, New York, NY, USA. ACM.

[Li et al., 2017] Li, J., Cai, Y., Cai, Z., Leung, H., and Yang, K. (2017). *Wikipedia Based Short Text Classification Method*, pages 275–286. Springer International Publishing, Cham.

[Lim-Cheng et al., 2014] Lim-Cheng, N. R., Richmond, C., Co, J., Gaudiel, C., Umadac, D., and Victor, N. (2014). Semi-automatic population of ontology of philippine medicinal plants from on-line text. In *DLSU Research Congress, De La Salle University, Manila, Philippines*, pages 6–8.

[Lochter et al., 2016] Lochter, J. V., Zanetti, R. F., Reller, D., and Almeida, T. A. (2016). Short text opinion detection using ensemble of classifiers and semantic indexing. *Expert Systems with Applications*, 62:243–249.

[Meng et al., 2013] Meng, W., Lanfen, L., Jing, W., Penghua, Y., Jiaolong, L., and Fei, X. (2013). *Improving Short Text Classification Using Public Search Engines*, pages 157–166. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Michel et al., 2011] Michel, Jean-Baptiste, and Shen (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

[Navigli and Ponzetto, 2010] Navigli, R. and Ponzetto, S. P. (2010). Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 216–225, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Navigli and Ponzetto, 2012] Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

[Organization, 1999] Organization, W. H. (1999). *WHO monographs on selected medicinal plants*, volume 2. World Health Organization.

[Pokou et al., 2016] Pokou, Y. J. M., Fournier-Viger, P., and Moghrabi, C. (2016). Authorship attribution using small sets of frequent part-of-speech skip-grams. In *FLAIRS Conference*, pages 86–91.

[Posadas-Durán et al., 2015] Posadas-Durán, J., Markov, I., Gómez-Adorno, H., Sidorov, G., Batyrshin, I., Gelbukh, A., and Pichardo-Lagunas, O. (2015). Syntactic n-grams as features for the author profiling task. *Working Notes Papers of the CLEF*.

[Rigutini and Maggini, 2004] Rigutini, L. and Maggini, M. (2004). *Automatic text processing: Machine learning techniques*. PhD thesis, Ph. d. thesis, University of Siena.

[Sebastiani, 2002] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.

[Sharma et al., 2016] Sharma, V., Law, W., Balick, M. J., and Sarkar, I. N. (2016). Identifying plant-human disease associations in biomedical literature: A case study. *AMIA Summits on Translational Science Proceedings*, 2016:84.

[Sharma and Sarkar, 2013] Sharma, V. and Sarkar, I. N. (2013). Leveraging concept-based approaches to identify potential phyto-therapies. *Journal of biomedical informatics*, 46(4):602–614.

[Shrestha, 2011] Shrestha, P. (2011). Corpus-Based methods for Short Text Similarity. In *Rencontre des Étudiants Chercheurs en Informatique pour le Traitement automatique des Langues*, volume 2, page 297, Montpellier, France.

[Silvaa et al., 2016] Silvaa, T. P., Santosb, I., Hidalgoc, J. M. G., and Almeidaa, T. A. (2016). Text normalization and semantic indexing to enhance sms spam filtering. *Knowledge-Based Systems*.

[Song et al., 2014] Song, G., Ye, Y., Du, X., Huang, X., and Bie, S. (2014). Short text classification: A survey. *Journal of Multimedia*, 9(5):635–643.

[Suganya et al., 2013] Suganya, S., Gomathi, C., et al. (2013). Syntax and semantics based efficient text classification framework. *International Journal of Computer Applications*, 65(15).

[Takeda et al., 2017] Takeda, M., Kobayashi, N., and Shiina, H. (2017). Classification of short comments by weighted tree kernels using the hierarchy of wikipedia. In *Proceedings of the 11th International Conference on Ubiquitous Information Management and Communication*, IMCOM '17, pages 84:1–84:5, New York, NY, USA. ACM.

[Thessen et al., 2012] Thessen, A. E., Cui, H., and Mozzherin, D. (2012). Applications of natural language processing in biodiversity science. *Advances in bioinformatics*, 2012.

[Thomas et al., 2001] Thomas, M. B., Lin, N., and Beck, H. H. (2001). A database model for integrating and facilitating collaborative ethnomedicinal research. *Pharmaceutical biology*, 39(sup1):41–52.

[UNAM, 2009] UNAM (2009). Biblioteca digital de la medicina tradicional mexicana. urlhttp://www.medicinatradicionalmexicana.unam.mx/atlas.php. Accedido 28-10-2017.

[Wang et al., 2014] Wang, F., Wang, Z., Li, Z., and Wen, J.-R. (2014). Concept-based short text classification and ranking. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 1069–1078, New York, NY, USA. ACM.

[Wei et al., 2010] Wei, K., Zhang, R., and Xu, X. (2010). Search-based short-text classification. In *5th International Conference on Pervasive Computing and Applications*, pages 297–301.

[Zhang and Wu, 2015] Zhang, X. and Wu, B. (2015). Short text classification based on feature extension using the n-gram model. In *2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, pages 710–716.