EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Tecnológico Nacional de México

Centro Nacional de Investigación y Desarrollo Tecnológico

Tesis de Maestría

Exploración para la identificación automática de palabras con polaridad

presentada por

Lic. Carlos Alberto Álvarez Vázquez

como requisito para la obtención del grado de

Maestro en Ciencias de la Computación

Director de tesis

Dr. Noé Alejandro Castro Sánchez

Codirector de tesis

Dr. Héctor Jiménez Salazar

Cuernavaca, Morelos, México. Julio de 2020.

cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico

Cuernavaca, Mor., 24/julio/2020

OFICIO No. DCC/108/2020
Asunto: Aceptación de documento de tesis
CENIDET-AC-004-M14-OFICIO

C. DR. GERARDO VICENTE GUERRERO RAMÍREZ
SUBDIRECTOR ACADÉMICO
PRESENTE

Por este conducto, los integrantes de Comité Tutorial del **C. Lic. Carlos Alberto Álvarez Vázquez**, con número de control M16CE066, de la Maestría en Ciencias de la Computación, le informamos que hemos revisado el trabajo de tesis de grado titulado **"Exploración sobre la identificación automática de palabras con polaridad"** y hemos encontrado que se han atendido todas las observaciones que se le indicaron, por lo que hemos acordado aceptar el documento de tesis y le solicitamos la autorización de impresión definitiva.

Dr. Noé Alejandro Castro Sánchez
Doctor en Ciencias de la Computación
08701806
Director de tesis

Dr. Héctor Jiménez Salazar
Doctor en Ciencias en Ingeniería Eléctrica
Co-director de tesis

Dr. Juan Gabriel González Serna
Doctor en Ciencias de la Computación
7820329
Revisor 1

Dra. Andrea Magadán Salazar
Doctorado en Ciencias Computacionales
10654097
Revisor 2

C.c.p. Depto. Servicios Escolares
Expediente / Estudiante
JGGS/lmz

EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Centro Nacional de Investigación y Desarrollo Tecnológico

"2020, Año de Leona Vicario, Benemérita Madre de la Patria"

Cuernavaca, Morelos 27/julio/2020
OFICIO No. SAC/ 240/2020
Asunto: Autorización de impresión de tesis

CARLOS ALBERTO ÁLVAREZ VÁZQUEZ
CANDIDATO AL GRADO DE MAESTRO EN CIENCIAS
DE LA COMPUTACIÓN
P R E S E N T E

Por este conducto tengo el agrado de comunicarle que el Comité Tutorial asignado a su trabajo de tesis titulado ***"Exploración sobre la identificación automática de palabras con polaridad"***, ha informado a esta Subdirección Académica, que están de acuerdo con el trabajo presentado. Por lo anterior, se le autoriza a que proceda con la impresión definitiva de su trabajo de tesis.

Esperando que el logro del mismo sea acorde con sus aspiraciones profesionales, reciba un cordial saludo.

A T E N T A M E N T E
Excelencia en Educación Tecnológica®
"Conocimiento y tecnología al servicio de México"

CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO SUBDIRECCIÓN ACADÉMICA

DR. GERARDO VICENTE GUERRERO RAMÍREZ
SUBDIRECTOR ACADÉMICO

C.c.p. M.E. Guadalupe Garrido Rivera. Jefa del Departamento de Servicios Escolares
Expediente
GVGR/CHG

Resumen

El presente trabajo se centra en la identificación automática de palabras con polaridad en textos; esta es una de las actividades que aparece con frecuencia en métodos del análisis de sentimientos. Este campo de estudio analiza opiniones, sentimientos, actitudes y emociones de las personas hacia entidades tales como productos, servicios, organizaciones, individuos, problemas, entre otros. Para ello se requiere contar con listas de palabras, también denominadas Léxicos afectivos, que se encuentran asociadas a emociones positivas o negativas. Por lo regular, estas listas de palabras con polaridad se recopilan manualmente y ello conlleva la utilización de recursos económicos, competencias lingüísticas, así como tiempo. Además, los métodos existentes que realizan esta tarea están dirigidos a un idioma en específico, otros enfocados en un dominio en particular o usan algún recurso como base.

En este trabajo se propone la utilización de métodos computacionales que logren minimizar la intervención manual en la construcción y/o enriquecimiento de léxicos afectivos. Se proponen diversos experimentos que muestran cómo el contexto en el que se presentan las palabras con polaridad, conformado por información lingüística, puede ser usado para identificar de manera automática vocablos con propiedades afectivas.

El método que se propone se apoya de un pequeño número de palabras que expresan polaridad (palabras denominadas “semillas”) para obtener un conjunto de modelos que representan la estructura a nivel sintáctico de estas, los cuales se usan para buscar en un texto vocablos que se encuentren en un contexto similar. El método se apoya en la similitud coseno entre los vectores de los contextos de las semillas y de las palabras del texto para determinar el grado de vinculación entre los mismos.

La evaluación del método utiliza un léxico afectivo para obtener la cantidad de palabras presentes en un corpus y se obtiene una línea tope, es decir, la cantidad máxima de palabras con polaridad que debe extraer el sistema. Los resultados obtenidos muestran variaciones en precisión y cobertura, logrando en algunos casos identificar palabras con polaridad con una precisión de hasta 0.98 y cobertura de 0.18. Aunque no se logran cubrir la cantidad de palabras, el método puede ser mejorado y superar lo obtenido hasta este punto.

El método propuesto en este trabajo tiene varios campos de aplicación; por mencionar algunos, puede ser utilizado para generar o enriquecer léxicos afectivos en idiomas que no cuentan con ellos (por ejemplo, idiomas indígenas), así como las diversas tareas que busca resolver el análisis de sentimientos.

Abstract

This work focuses on the automatic identification of polarity words in texts; one of the activities that appears frequently in methods of sentiment analysis. This field analyses people's opinions, feelings, attitudes and emotions towards entities such as products, services, organizations, individuals, problems, among others. To make possible this kind of tasks, it is necessary to have lists of words, also called affective lexicons, that are associated with positive or negative emotions. These polarized word lists are usually collected manually, and this involves the use of financial resources, language skills, and time. Furthermore, existing methods that perform this task are directed at a specific language, others focused on a particular domain, or use some resource as starting point.

In this work we propose the use of computational methods to achieve minimize manual intervention in the construction and / or enrichment of affective lexicons. We present a collection of experiments that show how the context in which polarity words are presented, and their linguistic information, can be used to automatically identify words with affective properties.

The proposed method relies on a small number of words that express polarity (words called "seeds") to obtain a set of models that represent their syntactic structure, which are used to search words that could be presented in a similar context into a text. The method uses the cosine similarity between the vectors of the seed contexts and the words of the text to determine their connection degree.

To evaluate the method, we use an affective lexicon to obtain the number of polarity words presented in the corpus to set the top line, that is, the maximum number of words with a polarity that the system must extract. The results obtained show variations in precision and recall, achieving in some cases to identify polarity words with a precision of up to 0.98 but a recall of 0.18. Although the number of words cannot be covered, the method can be improved and exceed what has been obtained up to this point.

The method proposed in this work has several fields of application; to mention a few, it can be used to generate or enrich affective lexicons in languages that do not have them (for example, indigenous languages), as well as the various tasks that the sentiment analysis try to solve.

Agradecimientos

Al Consejo Nacional de Ciencia y Tecnología (CONACyT) por su Programa Nacional de Posgrados de Calidad (PNPC) por medio del cual me fue otorgada una beca con número (CVU/Becario): 523223/605712, para ser estudiante de tiempo completo y desarrollar este trabajo de investigación.

Al Tecnológico Nacional de México (TecNM) y al campus Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), por la oportunidad de pertenecer a su gran comunidad y por facilitar sus instalaciones, gracias por permitirme el desarrollo de capacidades y habilidades. Al personal administrativo y académico, quienes con su trabajo diario permiten el desarrollo de nuestras actividades.

A mi director de tesis, el Dr. Noé Alejandro Castro Sánchez, por apoyarme, por todas las ideas y las charlas acerca del desarrollo del trabajo, por siempre estar para ayudarme y animarme con la investigación, por sus consejos e indicaciones, por su respaldo en las iniciativas que se propusieron, por enseñarme cuan interesante es el campo del Procesamiento del Lenguaje Natural y la infinidad de tareas que pueden realizarse, por mostrarme que es posible lograrlo, y por soportarme.

A mi co-director, por sus aportes al motivarme a experimentar, por sus ideas, por sus conocimientos y su forma de hablar acerca del PLN que me ayudo a comprender y amar más el campo, por sus correcciones y por su disponibilidad.

A mi comité tutorial conformado por la Dra. Andrea Magadán Salazar, quien me aconsejo, me dio palabras de aliento, por sus correcciones, por creer en mí, por su clase de Reconocimiento de patrones; y el Dr. Juan Gabriel González Serna, quien me exigió, me corrigió, me hizo esforzarme, por sus clases, por sus consejos a mis compañeros y a mí, nos hizo ver que si se podía si nos dedicábamos, a ellos quienes dedicaron parte de su tiempo a las revisiones de esta investigación y ayudaron a la mejora del mismo.

Dedicatoria

A Dios, por las bendiciones, los aprendizajes, las experiencias, las personas que conocí, por permitirme llegar hasta este punto - 2 Timothy 4:7.

A mi madre por sus palabras de aliento y su apoyo incondicional, gracias por darme estudios y permitirme llegar a este punto, por todo el esfuerzo y trabajo. Te quiero mamá.

A mis hermanos, que siempre me han alentado a ser mejor persona, me han apoyado de muchas formas. Los quiero mucho.

A mis sobrinos, gracias por su energía y sus palabras de aliento para lograr más. Tengo fe en que ustedes lograran cosas increíbles.

A los doctores que me guiaron en esta investigación por apoyarme, alentarme, corregirme, y por creer en mí.

Índice

Índice de tablas

Índice de figuras

Índice de ecuaciones

Capítulo 1

Introducción

El procesamiento de lenguaje natural (PLN) es parte del área de Inteligencia Artificial en conjunto con la Lingüística. Su base conceptual se enfoca en procesar un texto con la intensión de conocer su contenido y producir una respuesta que ayude al ser humano en la resolución de algún problema. El análisis de sentimientos y minería de opinión son tareas que se realizan dentro del PLN muy demandadas en el contexto de la comunicación a través de internet. La minería de opinión se centra en procesar los comentarios emitidos, por ejemplo, sobre productos o servicios, extrayendo una lista de atributos y agregando a cada uno de estos una polaridad (pobre, bueno, regular, etc.); mientras que el análisis de sentimientos es, en ciertos aspectos, paralelo a la minería de opinión, abordando el tratamiento computacional de opiniones, sentimientos y subjetividad en textos; por lo que se puede decir que ambos términos denotan el mismo campo de estudio y el uso de los mismos puede ser intercambiable. Como resultado del auge del PLN, se dio que un gran número de empresas de diferente tamaño incluyeran los procesos de análisis de sentimientos y minería de opinión dentro de su misión de trabajo (Pang & Lee, 2008).

Para implantar los métodos utilizados en el análisis de sentimientos y lograr generar respuestas eficaces se requiere un conjunto de recursos, dentro de los cuales destacan los léxicos afectivos. Un léxico afectivo se define como un diccionario de términos con connotaciones subjetivas asociadas con un grado de polaridad. Existen distintos tipos de léxicos, ser de propósito general, otros específicos de un dominio, o algunos más especializados, es decir, con un mayor grado de características.

Hasta ahora, los diferentes trabajos para la creación de estos léxicos afectivos han abordado los procesos de forma manual, lo que conlleva la utilización de recursos económicos, competencias lingüísticas, así como tiempo; por otra parte, algunos otros parten de corpus anotados, léxicos afectivos o recursos léxico-semánticos, por ejemplo WordNet, para obtener los lexicones de opinión de manera automática o semiautomática (López, Cruz, & Enríquez, 2016).

Con la presente tesis se aborda el problema de extracción automática de palabras con polaridad de textos escritos, apoyados en la hipótesis de que a partir de un pequeño número de palabras que expresen polaridad (palabras que denominaremos “semillas”) y un conjunto de modelos que representen la estructura sintáctica alrededor de este conjunto de palabras, son necesarios y suficientes para encontrar otros vocablos con polaridad dentro de un texto; ayudando así en la creación o expansión de léxicos afectivos sin la necesidad de recurrir a procesos manuales ni apoyándose de otros recursos como diccionarios o corpus anotados.

Este documento se divide de la siguiente manera:

- en el capítulo 1 se indica el problema por el que se origino esta investigación, la justificación y objetivos de la misma;
- el capítulo 2, se describen algún conceptos que deben conocerse para una mejor comprensión del contenido del trabajo;
- para el capítulo 3, se presentan los trabajos relacionados más representativos

obtenidos de la literatura, son trabajos que tienen un enfoque similar al de esta investigación o que aportaron ideas y técnicas para realizar experimentos;

- en el capítulo 4 se habla de las características consideradas para identificar palabras con polaridad, se presentan los diferentes experimentos que se realizaron y sus resultados;
- el capitulo 5 se presentan las pruebas realizadas al método de identificación de palabras con polaridad y los resultados de estas;
- para el capitulo 6 se describen las conclusiones de este proyecto de investigación y trabajos futuros que se pueden derivar;
- y por ultimo se presentan la bibliografía y anexos.

1.1 Planteamiento del problema

Conocer, entender e inferir lo que las personas quieren o desean es, actualmente, pieza clave en la toma de decisiones; resolver esta tarea ha sido motivación para que aumente el interés en el análisis de sentimientos y minería de opinión por parte de compañías e instituciones educativas (Pang & Lee, 2008).

Sea en la industria o en la investigación, la mayor parte de métodos utilizados en el análisis de sentimientos y minería de opiniones requiere de recursos lingüísticos tales como bases de datos, herramientas de software para procesamiento del lenguaje natural, corpus y lexicones afectivos. Un lexicón se puede pensar como un tipo de diccionario expandido con información cuantitativa, la cual tiene un formato que puede ser procesado por una computadora (Manning & Schütze, 1999). Ahora bien, un léxico afectivo se define como un diccionario de términos con connotaciones subjetivas asociadas con un grado de polaridad (López et al., 2016).

El problema es que este tipo de diccionarios afectivos, a menudo, son generados manualmente y ello conlleva la utilización de recursos económicos, competencias lingüísticas, así como tiempo para generarlos, sin omitir la existencia de errores naturales en todo proceso manual. En vista de estas dificultades se han generado algunos métodos para la creación, o más precisamente, la ampliación de léxicos afectivos a partir de textos o incluso de otros diccionarios; algunos de estos métodos están dirigidos a un idioma en específico y otros enfocados a un dominio particular, utilizando todos un léxico afectivo o un corpus anotado como base.

Dada la importancia e impacto que tiene la construcción de los léxicos afectivos, en este trabajo se lleva a cabo el desarrollo de un método para la identificación automática de palabras con polaridad en el texto escrito. Se planteó, al contrario de los métodos creados hasta el momento, apoyarse únicamente de un pequeño número de palabras que expresen emociones y un conjunto de modelos que representen el contexto lingüístico más representativos de dichas palabras emocionales, al mismo tiempo, se intentó desligar al método de un dominio específico. Con la implantación de este método es posible encontrar nuevos vocablos con polaridad dentro de un texto; con el objetivo de ayudar en la creación de léxicos afectivos.

1.2 Justificación

La identificación de la connotación subjetiva de las palabras dentro de un texto es una de las técnicas básicas del análisis de sentimientos. A partir de esta técnica y con la utilización de varios recursos, la minería de opinión, o también llamada análisis de sentimientos, logra clasificar textos, frases y palabras en diferentes polaridades, siendo estas clases la negativa, positiva y neutral; en algunos casos incluyen diferentes intensidades (muy positiva y muy negativa). Dentro de las tareas donde la minería de opinión se aplica resaltan la determinación de la opinión de los comentarios acerca de un producto dado, la clasificación de Tweets o de comentarios de redes sociales según su tipo de emoción, los resúmenes de críticas, el filtrado de mensajes, etcétera.

Los trabajos relacionados generalmente se enfocan en la identificación de palabras con polaridad, generando o expandiendo lexicones afectivos. Sin embargo, los enfoques hasta ahora establecidos requieren de una serie de recursos para lograr su objetivo; además, están dirigidos a un dominio específico o están disponibles para un lenguaje diferente al español. Por ello, es importante generar trabajos en el área para cubrir la mayor cantidad de dominios y dar un enfoque al idioma español.

En esta tesis, se considera proveer un método funcional para la identificación automatizada de palabras con polaridad sin utilización de recursos léxico-semánticos como base, sino aplicando únicamente características que componen a una palabra con polaridad, como su categoría gramatical, su lema y su contexto.

1.3 Objetivos

Los objetivos tanto general como específico que se han planteado en este trabajo de investigación, se mencionan a continuación.

1.3.1 Objetivo general

Desarrollar un método que, a partir en un conjunto reducido de vocablos con polaridad, denominados “palabras semilla“, identifique de manera automática otras palabras con similar propiedad afectiva dentro de un texto escrito.

1.3.2 Objetivos específicos

- Recopilar recursos léxicos (diccionario de palabras con polaridad, corpus de opiniones etiquetado, corpus sin anotar).
- Identificar y extraer un conjunto reducido de palabras semilla a partir de los recursos léxicos, y posteriormente asignarles una polaridad de forma manual.
- Identificar los patrones contextuales a nivel sintáctico de las palabras semilla extraídos.
- Localizar en un corpus sin anotar nuevas palabras emocionales y hallar nuevos patrones apoyándose de las semillas y modelos contextuales previamente identificados.
- Realizar pruebas en un par de dominios específicos con la finalidad de comparar el desempeño del método.
- Evaluar los resultados de las pruebas.

Capítulo 2

Marco teórico

En este capítulo se presentan algunos conceptos que el lector debe conocer para comprender el contenido del documento.

Lenguaje

Un sistema simbólico abstracto, gobernado por una relación de forma-significado la cual asigna contenido proposicional a cada combinación de símbolos que tienen forma de oración gramaticalmente correcta (Kamp & Reyle, 1993).

Procesamiento de lenguaje natural

Es el uso de computadoras para el procesamiento de textos o discursos en lenguaje natural (Geman & Johnson, 2004).

Palabras con polaridad

Llamadas también palabras de sentimiento, o de opinión, o palabras que contienen opinión. Para un mejor manejo del término, estas pueden referirse tanto a palabras individuales como a frases. Las palabras de opinión positiva se usan para expresar estados deseados mientras que las palabras de sentimiento negativo se usan para expresar cualidades o estados no deseados. Algunos ejemplos de palabras de sentimiento positivo son: hermoso, maravilloso, bueno, y asombroso. Ejemplos de palabras de opinión negativas son: malo, pobre y terrible (Liu, 2012).

Léxico afectivo

Es la colección de palabras con polaridad, también conocidas como palabras de opinión o palabras de sentimientos. Es también llamado léxico de opinión (Liu, 2012).

Corpus

Una colección de piezas de lenguaje que se seleccionan y ordenan de acuerdo con criterios lingüísticos explícitos para ser utilizadas como muestra del lenguaje (Sinclair, 1996).

Texto de grandes dimensiones, llamado también corpus lingüístico (Jorge-botana, Olmos, & León, 2007).

Corpus anotado

Un corpus anotado es un corpus en el que los datos se enriquecen con anotaciones lingüísticas que pueden ser de diferentes niveles: morfológico (en general asociando lema y categoría a las formas), sintáctica (constituyentes y/o dependencias), léxico-semántico,

anotación de la modalidad, la polaridad o la correferencia, entre otros (Castellón & Juarros, 2014).

Atributos de una palabra

Según la Real Academia Española (2016), la palabra es una unidad lingüística dotada generalmente de significado, que se separa de las demás mediante pausas potenciales en la pronunciación y blancos en la escritura.

Así, además de la definición antes mencionada se emplean otros atributos que son útiles computacionalmente. Debido a que en la implementación se utiliza la librería de *Freeling* se adopta el objeto *Word* de esta y sus atributos se describen a continuación.

- Forma de palabra. Es la forma "original" en que una palabra aparece en un texto y necesaria para crear un nuevo objeto Word.
- Lema. Es la forma única común a todas las posibles variaciones de una misma palabra. En sustantivos se utiliza el masculino en singular de la palabra y para los verbos su forma en infinitivo. Por ejemplo, el lema de "tomar", "tomemos" o "tomen" es "tomar"; el lema de "perro", "perros" o "perrito" es "perro".
- Etiqueta gramatical. Es una serie de símbolos que proveen información gramatical de la forma de palabra. Indica por ejemplo si es un verbo, sustantivo, adjetivo, etc.
- Forma fonética. Es la representación del sonido de la palabra por medio de símbolos de un alfabeto fonético. Se utiliza una serie de reglas fonológicas para generar la forma fonética, a este proceso se le llama "transcripción fonética".
- Forma semántica. Es la representación del significado de la palabra. Esta forma semántica no es descriptiva, es decir, no es un texto en el que se describe el significado de la palabra. La forma más común de representar la forma semántica es a través de un synset, el cual es un código único (ver 2.10) para cada concepto del mundo.

Se menciona en (Villarejo Martínez, 2016).

Etiquetado gramatical

En inglés, *Part-of-Speech tagging*o o *PoS tagging,* es la asignación de etiquetas gramaticales a cada una de las palabras de un texto. Este proceso se realiza dependiendo del contexto en que se encuentra cada una (Villarejo Martínez, 2016).

FreeLing

Es una biblioteca escrita en C++ que proporciona funcionalidades de análisis lingüístico (análisis morfológico, detección de entidades nombradas, etiquetado gramatical, análisis sintáctico, desambiguación del Sentido de Palabra, etiquetado de roles semánticos, etc.) para una variedad de idiomas (inglés, español, portugués, italiano, alemán, entre otros) (Padró & Stanilovsky, 2017).

WordNet

WordNet es una extensa base de datos en la que verbos, sustantivos, adjetivos y adverbios se encuentran agrupados en conjuntos de sinónimos cognitivos llamados *synsets*. Un synset es un signo lingüístico se compone de dos elementos: el significante y el significado. El primero corresponde a la representación escrita o hablada de un concepto, es decir, la palabra. El segundo corresponde al concepto en sí, a la imagen mental que se tiene del concepto.

A nivel computacional, para representar los significados en lugar de imágenes mentales se utilizan códigos llamados synsets. Un synset es un código único que representa un concepto en específico, por ejemplo, el synset de "perro" es 02084071-n. A este synset se le asocia un conjunto de lemas que adquieren este significado. En el caso de este synset, "perro" y "can" serían los lemas asociados. La estructura de WordNet la convierte en una herramienta útil para la lingüística computacional y el procesamiento del lenguaje natural (Villarejo Martínez, 2016).

Palabras semilla

Este término se utiliza en varios trabajos del estado del arte, los cuales estudian sus propiedades y las relaciones que tienen con otras palabras; se usan como datos iniciales para entrenamiento de algún algoritmo, como entrada de algún método para generar un comportamiento, o para identificar alguna unidad de comunicación, sean: texto completo, párrafo, oración o palabras.

En este trabajo el término refiere al proceso de inducción: a partir de un conjunto de palabras etiquetadas con una polaridad, y sus contextos dentro del corpus, servirán para identificar nuevas palabras con similares propiedades afectivas.

Como lo definen algunos autores es:

- Una lista corta de palabras con polaridad manualmente elegidas (Yazidi, Bai, Hammer, & Engelstad, 2015).
- Palabras etiquetadas con polaridad, usualmente seleccionadas manualmente (Yu, Deng, & Li, 2013).

Disparador emocional

Es una palabra o concepto que expresa una idea que, dependiendo del mundo de interés del lector, factores culturales, educativos y sociales, puede conducir a una interpretación emocional del contenido del texto. Ejemplos de desencadenantes emocionales son libertad, salario, empleo, venta, orgullo, etc. (Balahur & Montoyo, 2008).

Indicadores emocionales

Tramos de texto (palabras individuales o cadenas de palabras consecutivas) que transmiten contenido emocional en una oración, llamados también marcadores (Aman & Szpakowicz, 2007).

Palabras candidatas

A lo largo de este documento se utiliza el término “palabras candidatas” para referirse a algunos de los vocablos que satisfacen los contextos de las palabras semillas y que serán identificadas por el método propuesto.

Una palabra candidata es una palabra clasificada de acuerdo con la polaridad de sentimiento de una palabra semilla (Yu et al., 2013).

Análisis de la semántica latente (LSA)

Es un método de representación de textos. Un modelo plausible de la adquisición y la representación del conocimiento. Empleada para modelar algunos fenómenos cognitivos, además de aplicaciones más directas como son la corrección de textos en el ámbito académico, para medidas de cohesión y coherencia textual, para simular modelos de usuarios potenciales en usabilidad WEB o como complemento a las ontologías.

Para llevar a cabo la técnica, se procesa un corpus lingüístico. El corpus se representa en una matriz cuyas filas contiene todos los términos distintos del corpus (palabras) y las columnas representan una ventana contextual en la que aparecen esos términos (habitualmente párrafos). De este modo, la matriz contiene sencillamente el número de veces que cada término aparece en un documento. Esta matriz sufre una ponderación que resta importancia a las palabras excesivamente frecuentes y la aumenta a las palabras moderadamente infrecuentes con la idea de que las palabras demasiado frecuentes no sirven para discriminar bien la información importante del párrafo y las moderadamente infrecuentes sí. El siguiente paso es someter esta matriz ponderada a un algoritmo llamado Descomposición del Valor Singular (SVD), variante del análisis factorial. El SVD se aplica con la idea de reducir el número de dimensiones de la matriz original en un número mucho más manejable, sin que se pierda la información sustancial de la matriz original (Jorge-botana et al., 2007).

Espacio semántico

Representación de textos mediante vectores, tanto de términos como de documentos, que contienen información sustancial para la formación de conceptos. Una ventaja de estas representaciones, es que son susceptibles a comparaciones por medio de cosenos, distancias euclídeas u otras medidas de similitud (Jorge-botana et al., 2007).

N-grama

Los n-gramas tradicionales son secuencias de elementos tal como aparecen en un documento. En este caso la letra *N* indica cuántos elementos se deben tomar en cuenta, es decir, la longitud de la secuencia o de n-grama. Por ejemplo, existen bigramas (2-gramas), trigramas (3-gramas), 4-gramas, 5-gramas, etc. De esa manera, si se habla de unigramas, es decir, de n-grama construidos de un solo elemento, es lo mismo que hablar de palabras.

Vamos a ver un ejemplo de los n-gramas tradicionales de palabras. De la frase: *Juan lee un libro interesante* se pueden sacar los siguientes bigramas (2-gramas): *Juan lee, lee un, un libro, libro interesante*. O los siguientes trigramas (3-gramas): *Juan lee un, lee un libro, un libro interesante,* etc. Se puede sustituir cada palabra por su lema o por su clase gramatical

y construir los n-gramas correspondientes. Como se observa, el procedimiento es muy sencillo, pero se usa con gran éxito en los sistemas de lingüística computacional (Sidorov, 2013a).

Similitud de coseno entre vectores

Para expresar formalmente la similitud, se utiliza la medida del coseno del ángulo entre los vectores: menor el ángulo, mayor es su coseno; es decir, mayor es la similitud entre vectores — y de esa manera de los objetos mismos que estamos comparando—. Para calcular la similitud de coseno entre dos vectores, llamémoslos $V = [\ V\_1, V\_2, ..., V\_n\ ]$ y $U = [\ U\_1, U\_2, ..., U\_n\ ]$, se utiliza el producto interno (producto punto) de los vectores normalizados. La normalización consiste en la división del resultado entre la longitud de cada vector o, qué es lo mismo, de la multiplicación de sus longitudes. La longitud se llama la norma euclidiana y se denomina, por ejemplo, para el vector V como $\| V \|$.

Entonces la fórmula final para el cálculo de similitud de coseno consiste en obtener el producto punto de los dos vectores y después aplicar la norma euclidiana. De manera general (1):

$$sim(V, U) = \frac{\sum_{n=1}^{m}(V_n \times U_n)}{\| V \| \times \| U \|} \quad (1)$$

En este caso, la similitud de coseno nos indica que tan parecidos son los vectores V y U. Para los valores positivos, el coseno está en el rango de 0 a 1. Note que la similitud de coseno está definida exactamente para dos objetos (dos vectores). La similitud del objeto consigo mismo sería igual a 1 (Sidorov, 2013a).

Precisión, cobertura y medida F1

Precisión y especificidad (*precision* y *recall*, en inglés), miden la viabilidad de una hipótesis de manera formal. La combinación armónica de esas dos medidas se llama medida *F*1. Se recomienda utilizar esta última para comparación de los métodos. Son conceptos relativamente sencillos.

Ejemplo: Se tiene una colección de documentos y una petición. También se tiene un sistema que estamos evaluando.

El sistema genera la respuesta, es decir, nos presenta algunos documentos recuperados considerados relevantes para la petición; vamos a llamarlos “todos recuperados”.

Entre esos documentos algunos son recuperados correctamente, es decir, son documentos relevantes: “relevantes, recuperados”; mientras que otros son errores de recuperación cometidos por el sistema son documentos “no relevantes, recuperados”. Eso quiere decir que los documentos que llamados “todos recuperados” consisten en “relevantes, recuperados” y “no relevantes, recuperados”. De aquí surge el concepto de precisión: ¿qué tan buena es la respuesta con respecto a sí misma? ¿Cuántos documentos en la respuesta son recuperados de manera correcta? Precisión (P) es la relación de “relevantes recuperados” con respecto a

“todos recuperados” (2).

$$P = \frac{relevantes\ recuperados}{todos\ recuperados} \quad (2)$$

Por ejemplo, la precisión es 1, cuando todos los documentos en la respuesta son correctamente recuperados.

Existen, sin embargo, otros documentos que son relevantes para la petición, pero el sistema no los recuperó; vamos a llamarlos “relevantes, no recuperados”.

De aquí surge, el concepto de especificidad (*recall*, R). ¿Qué tan buena (específica) fue la respuesta del sistema con respecto a la colección? ¿Pudo recuperar la mayoría de los documentos relevantes o solo unos pocos?

Especificidad es la relación de los documentos “relevantes, recuperados” a todos los documentos relevantes (“todos relevantes”); este último conjunto incluye los documentos “relevantes, recuperados” y “relevantes, no recuperados” (3).

$$R = \frac{relevantes\ recuperados}{todos\ relevantes} \quad (3)$$

Especificidad es igual a 1 cuando el sistema recuperó todos los documentos relevantes. Siempre existe una relación entre precisión y especificidad: si se trata de aumentar uno de esos valores, el otro disminuye.

Finalmente, la fórmula para la medida $F1$ que combina la precisión P y la especificidad R es (4):

$$F1 = \frac{2 \times P \times R}{P + R} \quad (4)$$

La medida se llama $F1$ porque se da el mismo peso a la precisión y a la especificidad igual a uno. Se puede darles pesos diferentes, y se producen medidas relacionadas con $F1$ pero diferentes (Sidorov, 2013a).

Capítulo 3

Estado del arte

A continuación, se presentan 13 trabajos relacionados que se consideran los más representativos en la revisión de la literatura.

3.1 Búsqueda de palabras con polaridad de un dominio específico para la clasificación de sentimientos (Sharifi & Cohen, 2008)

El objetivo de este artículo se enfoca en extraer palabras con polaridad y con ellas determinar el sentimiento general de un texto a través del uso de campos aleatorios condicionales (CRF), estos son una forma de modelos gráficos no dirigidos utilizados para modelar datos secuenciales.

Para crear y probar el extractor de polaridad CRF, se tienen datos en los que las palabras individuales están etiquetadas con polaridad. Se utilizan tres lexicones afectivos, General Inquirer, Subjectivity Clues, SentiWordNet; con el clasificador Boosting como clasificador de bolsa de palabras se extraen las características de los tres léxicos afectivos anteriores, es decir, si se presenta en el lexicón se selecciona, los campos aleatorios condicionales (CRF) son usados como modelos de secuencia.

En este artículo no se muestra una comparación con otros métodos, por lo que únicamente se menciona el potencial de utilizar CRF para extraer palabras con polaridad.

Se llega a obtener una precisión de 70.9% al utilizar el léxico de Subjectivity Clues para la identificación de polaridad, no se presenta la cobertura.

Resulta muy interesante cómo se utilizan patrones contextuales y patrones de polaridad incorporando información lingüística.

3.2 ELS: un método a nivel-palabra para el análisis de sentimientos a nivel-entidad (Engonopoulos, Lazaridou, Paliouras, & Chandrinos, 2011)

Este artículo presenta un nuevo método para la clasificación de textos a nivel-entidad llamado ELS, esté utiliza un modelo de secuencia por campos aleatorios condicionales (CRF). El modelo de secuencia se usa para identificar el sentimiento de cada palabra en una crítica, y después ser utilizada para determinar el sentimiento de la entidad, según donde aparezca en el texto.

Se utilizó el corpus anotado “Customer Review” para implementar y probar el método. Este corpus contiene 314 reseñas de internet de 5 diferentes productos, y del cual cada elemento del conjunto de datos ha sido anotado con el sentimiento que expresa para las entidades mencionadas en el texto.

Para evaluar el rendimiento del método utilizaron sus propias anotaciones de los datos. No

se hizo una comparación contra otros enfoques a este nivel, a falta de disponibilidad.

En este artículo, en la tarea de clasificación de sentimientos utilizaron tres valores para etiquetar las palabras (positiva, negativa, neutral); mientras que en la extracción de opinión se tomó como una tarea de clasificación binaria (subjetiva, objetiva), con la clase objetiva que representa el sentimiento neutral y la clase subjetiva que representa el sentimiento positivo o negativo.

Los resultados para las tres clases (positiva, negativa, neutral) fueron los siguientes, Tabla 1 y 2:

Tabla 1. Resultados considerando precisión y cobertura en los niveles de polaridad.

	Positiva	**Negativa**	**Neutral**	**Total**
Precisión	62.6%	52.2%	53.7%	56.2%
Cobertura	62.8%	52.5%	53.2%	56.2%

Mientras que para las dos clases (subjetiva, objetiva) fueron:

Tabla 2. Resultados considerando precisión y cobertura en clases objetiva y subjetiva.

	Objetiva	**Subjetiva**	**Total**
Precisión	49.3%	80.3%	64.8%
Cobertura	54.8%	76.6%	65.7%

En este trabajo, se obtiene mejor precisión y cobertura al clasificar una opinión completa; aunque este no es el propósito de la presente tesis, el interés surge debido a los valores obtenidos a nivel palabra y a la técnica muy relacionada con lo que se pretende resolver en esta tesis.

3.3 Definición de disparador de emoción asociado a la cultura y aplicación a la clasificación de la valencia y la emoción en textos (Balahur & Montoyo, 2008)

Este artículo presenta un método para la identificación y clasificación de la polaridad y las emociones presentes en un texto, introduciendo un nuevo concepto denominado “disparador de emoción”. Se basan en tres teorías diferentes: la Teoría de la Relevancia de Pragmática, la Teoría de la Motivación de Maslow de Psicología y la Teoría de Necesidades de Neef de Economía, para así, construir de forma incremental una base de datos léxica de disparadores de emoción asociados a la cultura con la que se quiere trabajar.

El concepto de “disparador de emoción” se introduce de la motivación de basarse de los supuestos y principios de la Teoría de la Relevancia. Para la clasificación de los “disparadores

de emoción" y la creación de reglas de activación de una emoción se basan en la teoría de la motivación humana de Abraham Maslow y su pirámide. En paralelo, se aplica la matriz de Neef de las necesidades humanas fundamentales para crear un sistema de "disparadores emocionales" de satisfacciones-necesidades.

El método parte de la idea de que las palabras dentro de un texto no tienen polaridad, sino que se cargan de emoción dependiendo de la interpretación y el campo (mundo) de interés del lector, así como de la intención y campo de interés del autor. El campo o mundo de interés está constituido por necesidades generales y personales, factores de motivación, nociones que satisfacen estas necesidades, conocimiento de los hechos históricos y sociales, información de los medios de comunicación, etc. Le llaman a esta colección de factores "bolsa de conocimiento". Esta bolsa está constituida de conocimientos generales sobre las palabras y sus significados, términos afectivos generales, y disparadores de emociones. Estos últimos contienen términos que llevan en sí una emoción o un conjunto de emociones, cada uno en un cierto porcentaje; también, se expone que fueron propuestos por primera vez en este trabajo. En la colección se incluye además, el período, la cultura y el lugar.

La base de datos creada, parte de un conjunto inicial de términos, y es ampliada con la información de otros recursos léxicos como WordNet, NomLex. Se usa EuroWordNet para hacer el enlace entre idiomas y se completa y adapta a diversas culturas con bases de conocimiento específicas para cada lengua.

Después, utilizan la base de datos construida para buscar en textos la valencia (polaridad) y el significado afectivo.

Primero se analiza el texto de entrada y se obtiene para cada palabra la categoría gramatical, el lema y sus modificadores. Más adelante, se identifican los disparadores de emoción presentes en el texto, junto con sus correspondientes modificadores. Y se calcula la polaridad del texto sobre la base de los disparadores de emoción identificados y sus modificadores. Tanto para los disparadores de emoción obtenidos de la pirámide de Maslow, como para la matriz de Neef se calcula una puntuación denominada valencia ponderada del disparador de la emoción. Al final, la polaridad total del texto es igual a la suma de todas las valencias ponderadas de todos los disparadores de emoción en el texto. A continuación, se muestran los resultados obtenidos (ver Tabla 3).

Tabla 3. Comparativa entre los idiomas español e inglés.

	Precisión	**Cobertura**
Inglés	75.2	65.0
Español	71.1	66.1

Este artículo incrementó el interés por el uso de patrones a nivel sintáctico. Se habla de cómo existen modificadores, palabras que niegan, intensifican o disminuyen el significado de una emoción. Palabras que introducen a una palabra con polaridad, por ejemplo (más alegre, menos dolor).

3.4 Identificación de palabras semilla para la construcción de un léxico de orientación semántica por medio de un procedimiento supervisado (Vincze & Bestgen, 2011)

En este artículo se proponen realizar una optimización de las palabras semilla que tienen una orientación semántica y pertenecen a un léxico afectivo, y del cual se basan la mayoría de los métodos para la clasificación de textos. Se comparan 5 métodos automáticos que sirven para la construcción automática de lexicones con polaridad y a partir de esto se muestra la importancia de la selección de estas palabras semilla y el interés de identificarlas a través de un procedimiento supervisado.

Los cinco métodos que se compararon y que sirven para estimar automáticamente la polaridad de las palabras consisten: el primero, está basado en los enlaces de sinonimia entre los adjetivos, consiste en medir la distancia mínima, es decir el camino más corto, entre la palabra a la cual se le quiere atribuir un valor y las palabras semilla “bueno” y “malo”. La polaridad de un término es, entonces, igual a su distancia relativa con las dos palabras semillas.

El segundo y tercero, utilizan el análisis semántico latente (LSA) para construir un espacio semántico a partir de información estadística acerca de la co-ocurrencia de las palabras en el texto; el segundo lo emplea para estimar la distancia semántica entre palabras y 14 palabras semilla (7 positivas y 7 negativas), una palabra es por tanto más positiva si está más cercana a semillas positivas y más alejada de las negativas.

Mientras que el tercero, lo utiliza para identificar palabras frecuentemente asociadas con las palabras que quiere determinar su polaridad, atribuyendo a cada palabra la polaridad promedio de sus 30 vecinos próximos de los cuales su valencia es conocida

Estos tres métodos sirvieron de referencia para evaluar dos nuevas aproximaciones: una extensión del primer método y un método de aprendizaje supervisado de palabras semillas.

El cuarto método que se evaluó es una adaptación del primer método en el cual el número de pares de adjetivos de referencia es multiplicado por 7. Se tomaron los 7 pares de referencia del segundo método (7 positivos y 7 negativos).

Mientras que el quinto método fue el creado en el trabajo, derivado del segundo y tercer métodos arriba mencionados. Como primer paso se seleccionan las palabras semilla potenciales que son más extremas sobre la dimensión positiva-negativa basados en una norma de evaluación utilizada en el tercer método. Enseguida sobre la base de un espacio semántico obtenido por el LSA de una colección de textos, se calculan los vecinos entre cada una de las semillas potenciales y todas las palabras que se encuentren en la norma. Después, se utiliza un procedimiento de regresión a fin de construir un predictivo basado en las palabras semillas más eficaces en la predicción de la polaridad. Finalmente, se emplea este modelo predictivo para estimar la polaridad de los términos presentes en el espacio semántico, pero no en nivel inicial.

Los resultados obtenidos, al modificar el primer método son que se logra más del 80% de términos correctamente clasificados. Ese porcentaje se califica en la medida en que se calcula en un número limitado de palabras. Del quinto método desarrollado, el cual selecciona las palabras semilla por aprendizaje supervisado tuvo una eficacia cercana al 75%, sobrepasando al claramente al tercer método.

El objetivo prioritario de la investigación es identificar palabras semilla específicas que luego pudieran usarse en los otros métodos. Sin embargo, se selecciona el trabajo porque pone especial atención a la selección de semillas, se habla de cómo una combinación de semillas puede arrojar mejores resultados en comparación a seleccionar aquellas semillas que individualmente hacen una mayor contribución al método; además, esta tesis se apoya en una hipótesis semejante: ciertas características contextuales de las semillas podrán arrojar mejores resultados.

3.5 Identificación de palabras de opinión utilizando un modelo basado en optimización sin palabras semilla (Yu et al., 2013)

En este artículo se propone un modelo basado en optimización para la identificación de palabras de sentimiento (*Sentiment Word Identification,* SWI); pero en lugar de utilizar semillas, utiliza las etiquetas de los documentos.

El objetivo del modelo, llamado WEED, es explotar el fenómeno de "coincidencia de sentimientos". Lo primero que realizan es medir la importancia de las palabras que componen los documentos etiquetados semánticamente. La hipótesis que proponen es que, las palabras importantes están más relacionadas con la polaridad del documento, que las menos importantes. Lo segundo que hacen es estimar la polaridad de cada documento basándose en la importancia de sus palabras que los componen, junto con sus valores de opinión, y enseguida se compara la estimación de la polaridad con la polaridad real. Después de eso, se construye un modelo de optimización para todo el corpus con el fin de evaluar el error de estimación general, que se minimiza con los mejores valores de opinión de las palabras candidatas.

Para la realización de sus pruebas utilizaron dos conjuntos de datos, el primero es Cornell Movie Review Data 1, y el segundo Stanford Large Dataset 2. Y los compararon contra dos métodos que utilizan palabras semillas, uno de SO-PMI (Turney and Littman, 2003) y el otro de COM (Chen et al., 2012).

Se obtienen precisiones de 93.5% y 89.0% para los conjuntos de datos de Cornell y Stanford, respectivamente al obtener 10 palabras con polaridad; ahora bien, a lo que se presta atención de este trabajo, es cómo a medida que aumenta el tamaño de la lista de palabras de polaridad recomendadas (palabras candidatas), la precisión de todos los métodos disminuye en consecuencia, es algo relevante para esta tesis.

3.6 Clasificación de polaridad en textos con opiniones en español mediante análisis sintáctico de dependencias (Vilares Calvo, Alonso Pardo, & Gómez Rodríguez, 2013)

En este artículo se presenta un sistema de clasificación de polaridad para textos escritos en español, cuyas principales características son la utilización de diccionarios semánticos y de la estructura sintáctica de las oraciones para clasificar un texto subjetivo como positivo o negativo. Dentro de lo que conlleva, se menciona una segmentación, tokenización y etiquetado de los textos para a continuación obtener la estructura sintáctica de las oraciones mediante algoritmos de análisis de dependencias. La estructura sintáctica se emplea para tratar las construcciones lingüísticas: la intensificación (“muy”, “bastante”), las oraciones subordinadas adversativas (“pero”, “si no”) y la negación (“no”, “nunca”, “sin”).

Se logra obtener una precisión del 78.5%, los resultados muestran una mejora del rendimiento con respecto a los sistemas puramente léxicos y refuerzan la idea de que el análisis sintáctico es necesario a la hora de tratar construcciones lingüísticas en un entorno de minería de opinión o análisis del sentimiento, a fin de elaborar un método robusto y fiable.

Esta investigación aporto las primeras ideas para generar un prototipo y comenzar con experimentos, su punto de enfoque se dirige a cómo se usa la estructura sintáctica y se establece que la intensificación aparece con frecuencia en textos con contenido emocional.

3.7 Enriquecimiento automático de un léxico afectivo basado en relaciones semánticas obtenidas de un diccionario explicativo en español (Castro-Sánchez & López-Santiago, 2014)

Este trabajo está enfocado en estudiar las relaciones semánticas de sinonimia y heteronimia, determinando los patrones que son utilizados en las definiciones, de los sustantivos y verbos, presentadas en el diccionario de Real Academia de la Lengua Española. Estas definiciones tienen patrones que permiten, a partir de heurísticas, extraer relaciones de sinonimia e inclusión.

El método consiste, primero, en discriminar las entradas con un grupo de etiquetas de vigencia de uso y voces técnicas. Enseguida de filtrar, se buscaron patrones en las definiciones, identificando el uso de definiciones morfo-semántica y sinonímicas, y también relaciones de tipo hipónimo- hiperónimo.

Como resultado de la extracción de palabras de las definiciones morfo-semánticas se agregaron 125 nuevos términos al léxico afectivo. La precisión para la extracción de hiperónimos fue del 76%. Los hiperónimos encontrados y agregados al léxico afectivo sumaron 749 palabras.

Como conclusión del trabajo se observó que, la estructura de las definiciones del diccionario de la RAE hace idóneo el uso de patrones para la extracción de palabras. Además, se pudieron obtener buenos resultados en la identificación de sinonimia con un procesamiento sencillo y rápido, en comparación con otros métodos que, para obtener una precisión aceptable, se basan en corpus o en otros recursos.

3.8 Desarrollo de un servicio web para determinar la polaridad de textos de redes sociales en español (Baca Gómez, 2014)

En este trabajo se desarrolló una aplicación para detectar la polaridad de textos en español mediante la utilización del algoritmo de clasificación automática SMO (*Sequential Minimal Optimization*) y de la extracción de características a partir de un léxico afectivo en español. A falta de recursos léxicos disponibles para el idioma utilizado, también se lleva a cabo la creación de un léxico afectivo y de la creación de un corpus de comentarios de Facebook.

El método propuesto en este trabajo tiene un enfoque híbrido, es decir, se hace una combinación de la clasificación automática con un léxico afectivo. Lo primero que se realiza es un procesamiento del texto entrada, en esta parte se corrigen errores presentes, se hace un etiquetado gramatical, etc., utilizando la herramienta Freeling, dando como salida un texto corregido. Enseguida se extraen las características del texto que aportan información para la detección de la polaridad y la intensidad de esta, tomando en cuenta factores lingüísticos

como: la negación, modificadores de polaridad y expresiones con polaridad, etc., para lograr esto se utiliza el léxico afectivo.

Para la última parte se implantó el algoritmo de aprendizaje automático SMO. Se hizo un entrenamiento del algoritmo utilizando un corpus etiquetado con 5 categorías; ya entrenado el algoritmo se genera el modelo de clasificación, utilizando la librería Weka; la detección de polaridad se realiza a partir de este modelo, el cual recibe como parámetros el texto pre-procesado y el vector de características, obteniendo así la polaridad del texto.

En la evaluación, se obtuvo una precisión de 83.4% en pruebas de clasificación de 3 categorías (positivo, negativo y neutral) y un 56.6% en pruebas de clasificación de 5 categorías (muy positivo, positivo, neutro, negativo y muy negativo) utilizando un corpus de 1500 comentarios de Facebook, y 77.2% en las 3 categorías y 56.6% en las 5 categorías con un corpus de 3100 comentarios. Se expone que el algoritmo es más eficiente en comparación con los mencionados en estado del arte.

3.9 Etiquetado no supervisado de la polaridad de las palabras utilizando representaciones continuas de palabras (García-Pablos, Cuadros, & Rigau, 2015)

El enfoque de este trabajo es, partir de representaciones continuas de palabras (*word embeddings*) y calcular el valor de polaridad de palabras de cualquier dominio. Utilizan Word2vec para generar el modelo de vectores de palabras procedentes de un conjunto de datos de diferentes dominios.

Se entrenan tres modelos de Word2vec y generan algunos lexicones para un conjunto pequeño de palabras y expresiones de opinión de tres dominios (opiniones en inglés de restaurantes y laptops, y opiniones de cine). Específicamente, se anotaron 200 adjetivos tomados del conjunto de datos de cada dominio. Los resultados muestran que el 80% de los adjetivos del dominio de restaurantes anotados manualmente fueron correctamente etiquetados, mientras para las opiniones de laptops fue el 70% de los adjetivos. La tercera prueba se realizó con opiniones en español, y aunque los resultados fueron buenos, muchas de las palabras etiquetadas tenían una polaridad menos precisa, una posible razón fue que el corpus utilizado fue menos que el de las opiniones en inglés. Los resultados fueron bastante buenos a pesar de la naturaleza simple y sin supervisar del enfoque.

3.10 Ampliación de lexicones de opinión específicos de dominio usando representaciones continuas de palabras (López et al., 2016)

Este trabajo expone un método de ampliación de lexicones de opinión, extrayendo de documentos sin anotar, términos a incluir en un lexicón usado como base, a cada uno de estos términos se les agrega una estimación de su polaridad. Lo que se busca es asegurar la precisión, es decir, que la mayoría de los términos que se agreguen y sus polaridades sean correctos. El objetivo secundario que presentan es corroborar las bondades de las representaciones continuas de palabras (*word embeddings*) aplicándolas en tareas de clasificación de la polaridad, mediante el uso de la herramienta Word2vec.

El método consiste en entrenar clasificadores ternarios que deciden si una palabra candidata

tiene una polaridad, tomando como entrada representaciones continuas de palabras. Primero, en la fase de entrenamiento, se seleccionan las palabras positivas, negativas y neutras del lexicón base y se obtiene la representación continua de palabras utilizando Word2vec, eligiendo un modelo para la herramienta, construido de textos de un dominio específico o de un corpus genérico. Enseguida, se entrena un clasificador multiclase de tipo *Support Vector Machines* (SVM). Después, para las palabras de los textos no anotados, se obtienen sus representaciones continuas. Los vectores obtenidos son pasados por el clasificador que decidirá si la palabra tiene polaridad o es neutra.

Los resultados mostraron que dichas representaciones contienen información relativa a la polaridad de las palabras, el método fue capaz de ampliar los lexicones iniciales capturando la mitad del total de las palabras de opinión contenidas en los textos de partida, y con una alta precisión: de los dominios analizados, casi el 94% de las palabras seleccionadas eran palabras de opinión y su polaridad estaba correctamente asignada.

Resulta interesante como su trabajo parte de una hipótesis parecida a la de esta investigación, establecen que existe una relación similar entre palabras con misma polaridad, en este caso entre *word embedding*. Por otro lado, aunque se obtienen muy buenos resultados, requiere de otros recursos, en particular entrenamiento, lo cual difiere con el enfoque de la presente tesis.

3.11 Extracción de patrones de cambio de polaridad chinos en grandes corpus (Xu & Huang, 2016)

Este estudio está enfocado a extraer patrones que invierten, incrementan o cancelan la polaridad, a través de un enfoque semiautomático basado en la secuencia de minería de secuencias.

Para probar el método se utilizó el algoritmo para minería de secuencias PrefixSpan, procesando dos corpus de dominio (críticas de comida y criticas productos).

Se realizaron tres experimentos principales para extraer los patrones de cambio de polaridad:

- I: Cambiando las palabras o frases que aparecía después de los adverbios de grado (pistas) de positivas en negativas

- II: Cambiando las pistas negativas en buenos comentarios

- III: Cambiando entre pistas positivas y pistas negativas

Una de las observaciones que destaca en los experimentos fue la disminución hacia lo negativo. Muestran que la representación de lo "negativo" debe hacerse con más cuidado y eufemismosía; además, de detectar cambios de polaridad más sutiles a medida que se extraen patrones de cambio de polaridad.

Este trabajo muestra lo importante que es realizar un análisis del contexto en el proceso de identificación de la polaridad aparente de una palabra.

3.12 Análisis de polaridad de textos usando un algoritmo basado en Sentiwordnet (Tomar & Sharma, 2016)

El objetivo de este artículo es ofrecer una mejor estrategia de análisis de texto de opinión para el idioma inglés, que reconozca la polaridad de los mensajes de texto, incluidos los

mensajes positivos, negativos y neutros.

Lo que se menciona es que los potenciadores/intensificadores de polaridad y de negaciones afectan la polaridad general de una manera anormal, por lo que se discute, en no depender de la polaridad de una palabra en particular solo para obtener resultados precisos. Por ejemplo, en una oración como “no odiamos…”, la frase expresa algo positivo, pero la palabra en si misma tiene una polaridad negativa.

Lo que hacen es evaluar la polaridad de una oración mediante el uso de Sentiwordnet; y con la ayuda del etiquetador POS Stanford tokenizar. Luego seleccionan las categorías gramaticales de las palabras que podría afectar la polaridad de una oración. Y al final, a través de su propio algoritmo encuentran la polaridad general de la oración que también incluya aquellas palabras que podrían mejorar, invertir o disminuir la polaridad de la palabra correspondiente.

Como resultado al evaluar el método se presentan los siguientes valores:

- Al clasificar 2000 criticas de películas, 1000 positivas y 1000 negativas, se obtuvo una precisión de 73% para las positivas y 72% para las negativas.
- Al clasificar 200 criticas de hotel, se logró clasificar con una precisión de 73% para las positivas y de 69% para las negativas.

Al igual que otros trabajos mencionan que es importante considerar el contexto y el discurso en donde aparecen las palabras con polaridad.

3.13 Identificación de palabras de opinión y polaridad a partir de críticas en Tweets utilizando la Minería de Opinión basada en aspectos (Vadivukarassi, Puviarasan, & Aruna, 2017)

Este trabajo propone un modelo para detectar palabras de opinión de un usuario sobre las críticas de productos en Tweets utilizando el modelo de minería de opiniones basado en características/aspectos.

El método que proponen funciona de la siguiente manera: Como entrada se recibe la información del usuario en sobre los productos recopilados de los datos de Twitter. Después, se procesa el Tweet se realiza para eliminar los símbolos innecesarios, luego se usa la tokenización para dividir el grupo de Tweets en conjuntos únicos, y se hace un etiquetado gramatical (en inglés *part-of-speech, PoS*) para analizar cada oración y luego identificar los aspectos del producto y las palabras de opinión. Las palabras de opinión extraídas se utilizan para definir la polaridad (positiva o negativa). Finalmente, se resumen las opiniones para cada característica del producto en función de sus orientaciones.

En este artículo no se exponen resultados numéricos, sin embargo, destacan aspectos importantes a la hora de identificar palabras de opinión. Lo que se describe es que la mayoría de las palabras de opinión son palabras que expresan características sobre el aspecto de un producto; además se dice que son en su mayoría verbos, adverbios, adjetivos, adjetivos y arreglos verbales adverbiales.

También, se menciona que para cada palabra es necesario identificar su orientación semántica, de esta forma se puede predecir la orientación semántica de cada oración (Tweet). Y finalmente, hacen mención de que al momento obtener la información contextual de una oración, las negaciones deben manejarse de manera apropiada.

3.14 Tablas comparativas de los trabajos relacionados

Se presentan los trabajos que tienen un enfoque similar al de esta investigación (ver Tabla 4).

Tabla 4 Tabla comparativa de trabajos relacionados con la identificación de palabras con polaridad

Trabajo	**Idioma**	**Recursos y técnicas de solución**	**Precisión**	**Cobertura**
Búsqueda de palabras con polaridad de un dominio específico para la clasificación de sentimientos (Sharifi & Cohen, 2008)	Inglés	Uso de clasificador de bolsa de palabras y léxicos afectivos	70	-
Definición de disparador de emoción asociado a la cultura y aplicación a la clasificación de la valencia y la emoción en textos (Balahur & Montoyo, 2008)	Inglés, español	Uso de WordNet, EuroWordNet y disparadores de emoción (términos etiquetados con una emoción para un dominio en específico), conjunto de términos iniciales (semillas)	71	66
Identificación de palabras de opinión utilizando un modelo basado en optimización sin palabras semilla (Yu, Deng, & Li, 2013)	Inglés	Corpus anotados	93	-
Enriquecimiento automático de un léxico afectivo basado en relaciones semánticas obtenidas de un diccionario explicativo en español (Castro-Sánchez & López-Santiago, 2014)	Español	Léxicos afectivos, relaciones semánticas de sinonimia y heteronimia	76	-
Etiquetado no supervisado de la polaridad de las palabras utilizando representaciones continuas de palabras (García-Pablos, Cuadros, & Rigau, 2015)	Español, inglés	Representaciones continuas de palabras (*word embedding*)	80	-
Ampliación de lexicones de opinión específicos de dominio usando representaciones continuas de palabras (López et al., 2016)	Español	Representaciones continuas de palabras (*word embedding*), léxicos afectivos	94	-
Identificación de palabras de opinión y polaridad a partir de críticas en Tweets utilizando la Minería de Opinión basada en aspectos (Vadivukarassi, Puviarasan, & Aruna, 2017)	Inglés	Sentiwordnet, categorías gramaticales como aspectos/características	-	-

Por otro lado se presentan aquellos trabajos que aportaron ideas y técnicas interesantes para tomar en cuenta al momento de experimentar (ver Tabla 4).

Tabla 5 Tabla comparativa de trabajos base para desarrollo de ideas

Trabajo	**Idioma**	**Recursos y técnicas de solución**	**Precisión**	**Cobertura**
ELS: un método a nivel-palabra para el análisis de sentimientos a nivel-entidad (Engonopoulos, Lazaridou, Paliouras, & Chandrinos, 2011)	Inglés	Modelos de secuencia por campos aleatorios condicionales y corpus anotados	80	76
Identificación de palabras semilla para la construcción de un léxico de orientación semántica por medio de un procedimiento supervisado (Vincze & Bestgen, 2011)	Inglés	Análisis semántico latente (LSA), utilización de palabras semilla	-	-
Clasificación de polaridad en textos con opiniones en español mediante análisis sintáctico de dependencias (Vilares Calvo, Alonso Pardo, & Gómez Rodríguez, 2013)	Español	Análisis sintáctico de dependencias, corpus anotados	78	-
Desarrollo de un servicio web para determinar la polaridad de textos de redes sociales en español (Baca Gómez, 2014)	Español	Léxicos afectivos, algoritmo de clasificación automática SMO (*Sequential Minimal Optimization*)	83	-
Extracción de patrones de cambio de polaridad chinos en grandes corpus (Xu & Huang, 2016)	Chino	Algoritmo para minería de secuencias PrefixSpan	-	-
Análisis de polaridad de textos usando un algoritmo basado en Sentiwordnet (Tomar & Sharma, 2016)	Inglés	Evaluación de polaridad de oraciones con Sentiwordnet	73	-

Capítulo 4

Método de solución

Para lograr la resolución del problema planteado anteriormente, se propuso la realización de diversos experimentos que permitieran identificar la mejor combinación de variables lingüísticas para la extracción de nuevas palabras con polaridad. Se propuso el siguiente método de solución que consta de cuatro fases: 1) conformar recursos léxicos, 2) realizar experimentos para extraer palabras semillas y patrones contextuales, 3) realizar experimentos para identificar palabras candidatas y nuevos patrones, 4) probar y evaluar resultados considerando todos los módulos desarrollados y que permitieron obtener los mejores resultados en la fase experimental. El proceso completo se muestra en la Figura 1.

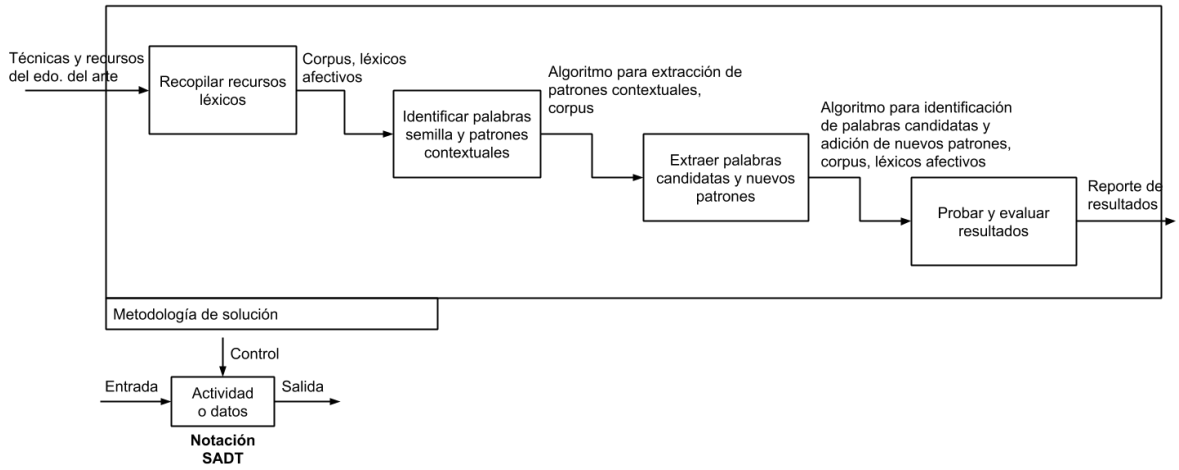

Figura 1 Método de solución

La primera fase consiste en enfocarse en la búsqueda de dos tipos de recursos léxicos: léxicos afectivos y corpus de texto, recursos que se usarán para realizar los experimentos.

La segunda fase consiste en realizar un análisis preliminar de los corpus utilizando las palabras emocionales de los recursos léxicos obtenidos. Para ello se analiza el contexto de las semillas (palabras emocionales) para precisar las características que deben de tener los patrones contextuales, es decir, las características que permitirían identificar a las palabras con polaridad.

Se puede ver el flujo de las actividades en la Figura 2:

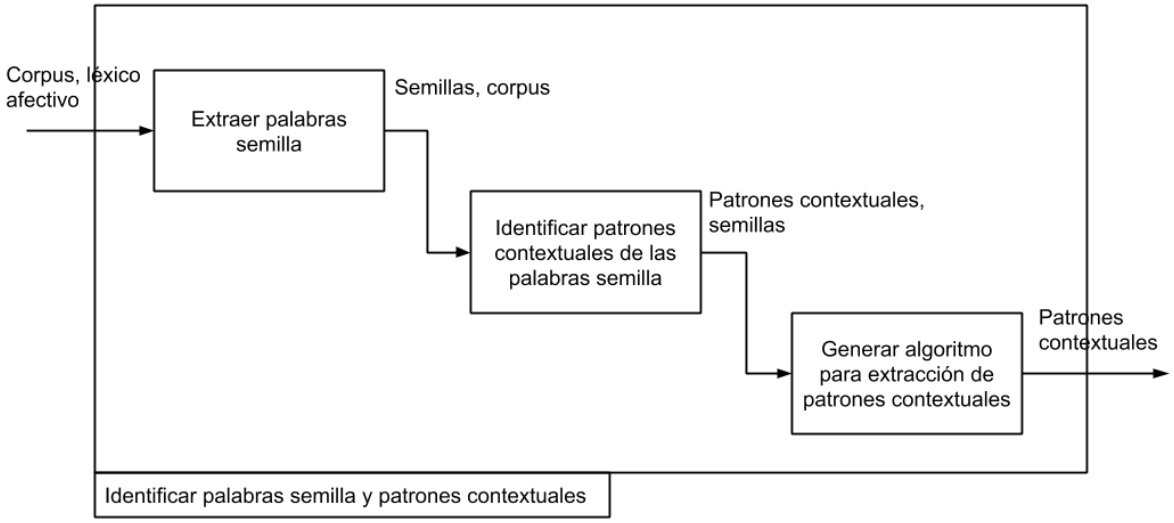

Figura 2 Actividades de la fase dos de la metodología

El objetivo de la tercera fase es realizar diversos experimentos que permitan identificar la mejor combinación de variables lingüísticas para obtener palabras candidatas con los resultados más altos con respecto a precisión y cobertura.

En la siguiente Figura 3 se muestra la secuencia de las tareas propuestas para esta actividad.

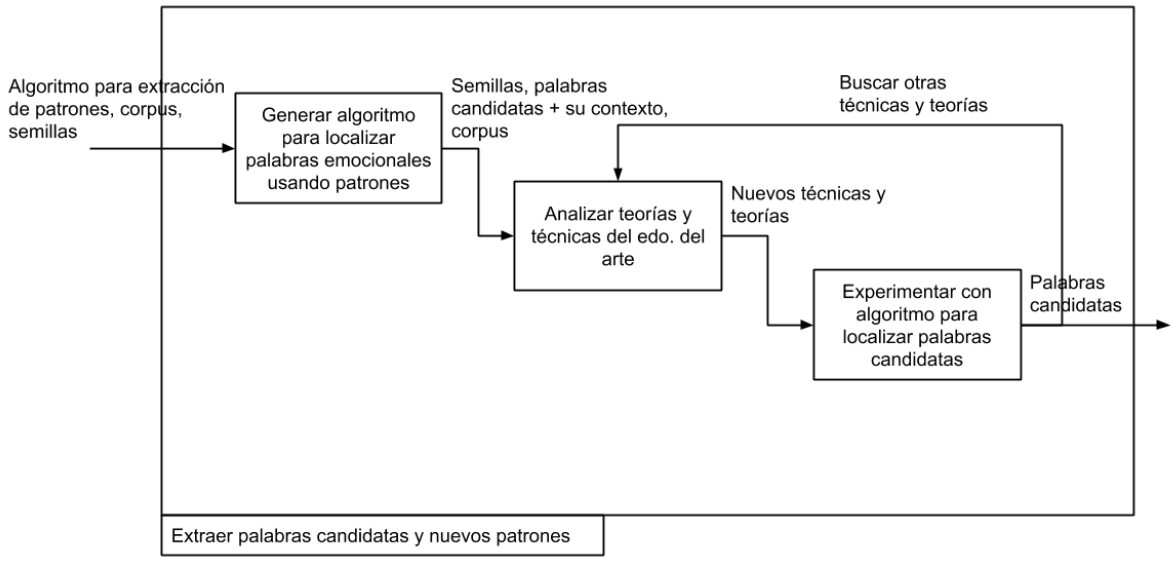

Figura 3 Nuevos técnicas y teorías

Para la cuarta fase se propone diseñar y realizar pruebas usando una aproximación de caja negra, en las cuales normalmente se evalúa el comportamiento de un sistema generando un catálogo de entradas y comparar las salidas esperadas con los resultados reales.

Algunas de las pruebas a realizar que se consideran son: comparar que el método propuesto pueda identificar las palabras con polaridad del corpus por cada uno de los corpus anotados o léxicos obtenidos y calcular la eficiencia del método usando medidas de evaluación propias de un sistema de extracción de información.

A continuación, se describen las acciones que se realizaron en cada una de las fases.

4.1 Recopilación de recursos léxicos

En esta actividad, se recuperaron recursos útiles para el diseño y desarrollo de los algoritmos y las pruebas.

Se obtuvieron dos léxicos afectivos, uno fue tomado de (Baca Gómez, 2014), y otro generado en el trabajo de (Sidorov, 2013b).

El léxico de (Baca Gómez, 2014) consta de 3550 palabras obtenidas a través de la traducción de diversos trabajos en el idioma inglés. Tienen una polaridad asignada, además, algunas de ellas tienen una emoción a la que se relacionan.

El léxico presentado por (Sidorov, 2013b) contiene 2036 palabras relacionadas con una emoción que puede ser alegría, ira, miedo, tristeza, sorpresa o disgusto. Cada una de las palabras está asociada con la medida del factor de probabilidad de uso afectivo (AFP) con respecto a la emoción con la cual está vinculada. Este recurso lo presentan como una herramienta para el análisis de sentimientos, aunque el objetivo estaba más enfocado al procesamiento de Tweets. Se menciona que puede aplicarse en cualquier texto, mientras no sea un análisis detallado.

En la Tabla 6, se muestran cómo están compuestos estos léxicos afectivos.

Tabla 6: Características de léxicos afectivos

Léxico afectivo	**Cantidad de palabras**
(Baca Gómez, 2014)	2,036
(Sidorov, 2013b)	3,550

Por otro lado, se generaron dos corpus en español sin anotar, el primero contiene cuentos infantiles y el segundo novelas en español.

Se presentan sus contenidos en la siguiente tabla.

Tabla 7: Características de corpus

Corpus	**Cantidad de palabras**
Novelas	89,281
Cuentos	14,036

Los cuentos utilizados son los siguientes: Caperucita Roja, El enano saltarín, El ganso de oro, El gato con botas, El gigante egoísta, El lobo y los cabritillos, El mago de Oz, El patito feo, El sastrecillo valiente, El soldadito de plomo, Hansel y Gretel, Rapunzel, Los tres cerditos, Pedro y el lobo, Pinocho y Ricitos de oro.

Las novelas son: El cuaderno dorado y El quinto hijo de la autora Doris Lessing y Más allá del

invierno de la autora Isabel Allende. Fueron elegidas por aparecer entre las más leídas en el sitio casadellibro.com.

4.2 Identificación de semillas y patrones contextuales

Los experimentos mostrados en esta sección, corresponden a la fase 2, que consiste en la identificación de las semillas y patrones contextuales. El primer paso fue el procesamiento del corpus, el cual consiste en analizar el texto con la herramienta *Freeling*, la cual permite lematizar y hacer un etiquetado gramatical; obteniendo para cada palabra su información morfosintáctica (lemas y etiquetas gramaticales).

Para instalar la herramienta se siguió lo que se indica en la documentación expuesta en https://freeling-user-manual.readthedocs.io/en/latest/. Al procesar los textos, se extrajeron las palabras del léxico afectivo del trabajo de (Baca Gómez, 2014) y se buscan en el corpus; al encontrarse alguna de las palabras se obtuvieron n-gramas con valores de n desde uno hasta cinco, tanto a la izquierda y como a la derecha de las mismas, esto con el objetivo de observar si existía un contexto a nivel sintáctico similar entre las palabras con polaridad.

Tal y como se menciona en (Alm, Roth, & Sproat, 2005), los textos narrativos con frecuencia se componen de contenido emocional, en ellos se expresan emociones como la felicidad, la ira, el amor, el odio, etcétera, estas son parte fundamental en el desarrollo de las historias por lo que se les da una particular importancia. También, en (Mohammad, 2011) se habla que desde hace mucho tiempo este tipo de textos han sido utilizados para transmitir emociones, ya sea de forma explícita o implícita. Algunos de los propósitos a los que hacen referencia, y por los cuales aplicar análisis de sentimientos a este tipo de textos, incluyen:

- Buscar las emociones que se expresan dentro de las obras, sea por capítulo o libros enteros.
- Aplicar análisis social, es decir observar cómo se han plasmado (utilizando palabras emocionales) a las personas o entidades, por ejemplo, a través del tiempo, o en una región en particular.
- Comparar obras literarias, por ejemplo, por géneros, o estilos de escritura de autores.
- Generar resúmenes de forma automática, capturando las emociones que se expresan en los textos.
- Analizar palabras con emoción y cómo influyen en la persuasión.

A continuación, se muestran ejemplos de lo que se obtuvo al procesar los textos, se puede apreciar como en la Tabla 8 y Tabla 9, una palabra con polaridad podría aparecer entre mismas clases gramaticales, en este caso *VS, RG* (Verbo, Adverbio) a la izquierda y *SP, DP* (Preposición, Determinante) a la derecha de la palabra con polaridad.

Tabla 8: Primer ejemplo de extracto de texto procesado

Palabra	una	niña	que	era	muy	**querida**	Por	su	abuelita
Categoría gramatical	DI	NC	PR	*VS*	***RG***	VM	***SP***	*DP*	N

Tabla 9: Segundo ejemplo de extracto de texto procesado

Palabra	El	gato	con	botas	que	se	sentía	muy	**complacido**	con	su	plan
Categoría gramatical	DA	NC	SP	NC	PR	P	*VM*	***RG***	VM	***SP***	*DP*	NC

Las cantidades de palabras con polaridad obtenidas al momento de procesar los léxicos y obtener los n-gramas son los que a continuación se presentan. Al procesar seis de los cuentos del corpus se obtuvo lo presente en la siguiente Tabla 10:

Tabla 10: Cantidad de palabras con polaridad presentes en extracto de cuentos

Texto procesado	Cuentos de 4,885 palabras
Léxico afectivo	**Cantidad de palabras con polaridad**
(Baca Gómez, 2014)	385
(Sidorov, 2013)	317

Y al procesar 3 capítulos de una novela (Más allá del invierno), se extrajo lo presente en la Tabla 11:

Tabla 11: Cantidad de palabras con polaridad en extracto de novela

Texto procesado	Capítulos de novela de 9,292 palabras
Léxico afectivo	**Cantidad de palabras con polaridad**
(Baca Gómez, 2014)	890
(Sidorov, 2013)	557

Enseguida, se seleccionaron las palabras con polaridad que aparecen con más frecuencia dentro de los textos procesados con el fin de que sirvieran para probar el método utilizándose como “semillas” para extraer los contextos y conformar los patrones. Se muestra en la Tabla 12 las palabras y la cantidad presente en los textos.

Tabla 12: Palabras presentes con mayor frecuencia en textos

Palabra con polaridad	Cantidad presente en texto	Palabra con polaridad	Cantidad presente en texto
vivir	52	lograr	24
dar	38	seguro	24
querer	38	caer	24
llegar	33	amor	20
quedar	30	miedo	18
malo	28	frío	18
bien	28	vida	17
bueno	26	mal	16
perder	24	temor	16
solo	24	dejar	16

La Tabla 13 y 14 muestran un ejemplo de 5 unigramas y 5 bigramas de las categorías gramaticales (ver Atributos de una palabra en el Capítulo 2) que aparecen usualmente a la izquierda y derecha de una palabra con polaridad. Esto se hizo con el fin de darse una idea de cómo aparecen los patrones contextuales y cómo podría desarrollarse el algoritmo.

Tabla 13: Cinco unigramas que aparecen con frecuencia alrededor de una palabra emocional

Contexto	**Categoría gramatical**	**Cantidad encontrada**
Izquierdo	DA	269
	NC	261
	SP	247
	VM	210
	RG	137
Derecho	SP	587
	Fc	242
	Fp	194
	VM	173
	CC	145

Tabla 14: Cinco bigramas que aparecen frecuentemente alrededor de una palabra con polaridad

Contexto	**Categoría gramatical**	**Cantidad encontrada**
Izquierdo	SP DA	130
	DA NC	114
	NC SP	104
	VM SP	64
	VM RG	48
Derecho	SP DQ	238
	Fp	187
	DA NC	89
	SP NC	70
	SP VM	67

4.2.1 Identificación de patrones contextuales

Este punto se enfoca en el desarrollo del algoritmo para identificar los patrones contextuales de las semillas.

La lógica del prototipo se basa en lo obtenido al procesar el corpus, por otra parte, se toma en cuenta la hipótesis y resultados presentados en (Xu & Huang, 2016), el cual refiere a la aparición frecuente de adverbios de intensidad modificando palabras con polaridad, ya sea que se presenten (en la mayoría de los casos para el idioma chino) inmediatamente antes de la palabra que modifican o se produzcan después de la palabra modificada (algunos casos para el idioma chino, pero más frecuente en idioma inglés). Por otra parte, en el artículo se alude a que la relación entre adverbios de intensidad y las palabras que modifican es adecuada para el análisis de texto automático a gran escala.

A continuación, se presentan ejemplos que se mencionan en el artículo:

- Estoy **muy** feliz. (El adverbio “muy” aparece antes de la palabra con polaridad "feliz").
- Soy la persona **más** feliz. (El adverbio se reemplaza por el superlativo).

- Te apoyo **mucho**. (El adverbio "mucho" aparece después de la palabra que modifica).

Partiendo de lo anterior, se genera una primera versión del algoritmo, este programa no utiliza ningún otro recurso, únicamente los patrones encontrados, a continuación se presenta su funcionamiento en la Figura 4.

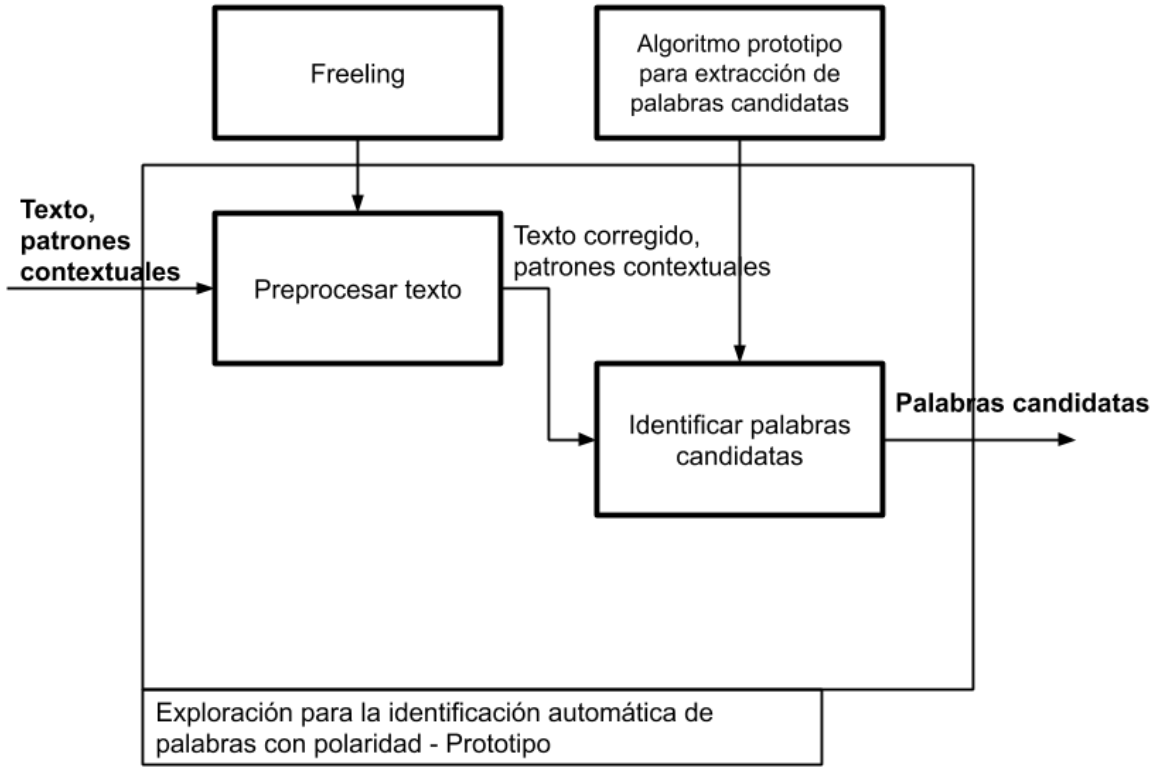

Figura 4 Prototipo de algoritmo para primera comprobación de hipótesis

El objetivo de esta primera versión fue comprobar de forma rápida la hipótesis planteada, intentando obtener algunas de las palabras con polaridad presentes en el grupo de los seis cuentos, se logra lo presente en la Tabla 15:

Tabla 15: Resultados de algoritmo prototipo para comprobación de hipótesis

Texto procesado	Cuentos de 4885 palabras				
Léxico utilizado	**Palabras candidatas**			**Precisión**	**Cobertura**
	Identificadas correctamente	Identificadas incorrectamente	No identificadas		
(Baca Gómez, 2014)	35	16	350	0.68	0.09
(Sidorov, 2013b)	19	32	298	0.37	0.05

4.2.2 Integración de semillas y extracción de patrones contextuales

Al limitar el conjunto de patrones y definirlos de forma manual, no se encontraban otras palabras presentadas en contextos diferentes a los ingresados, afectando así las cantidades de palabras identificadas Lo que se hizo entonces fue generar un algoritmo para que pudiera encontrar nuevos modelos, es decir que el algoritmo recibiera como entrada un conjunto de palabras "semilla" y a partir de su contexto sintáctico se extrajeran los posibles patrones que sirvieran para encontrar nuevos vocablos con polaridad.

La primera prueba de esta versión del algoritmo se hizo con patrones basados en unigramas, es decir modelos generados a partir de la categoría gramatical de la palabra a la izquierda y a la derecha de la palabra utilizada como semilla, un ejemplo de lo que se buscaba se presenta a continuación. En la Tabla 16 se presenta uno de los patrones obtenidos de la semilla "malo", como este adjetivo calificativo (AQ) se encuentra rodeado por un nombre (NC) y un verbo (VM).

Tabla 16: Ejemplo de contexto de semilla

Patrón contextual izquierdo	**Semilla**	**Patrón contextual derecho**
Lobo	malo	ando
NC	AQ	VM

Enseguida se presenta un ejemplo de los resultados del algoritmo al ingresar la semilla "malo" (ver Tabla 17): los patrones en que se encontraron las palabras y como aparecen en el texto.

Tabla 17: Ejemplo de contextos de resultados del algoritmo

Tipo de patrón	**Contexto izquierdo**	**Palabra candidata**	**Contexto derecho**
Idéntico	Bestia	**mala**	despertó
	NC	**AQ**	VM
Idéntico	Gigante	**egoísta**	decidió
	NC	**AQ**	VM
Similar	Lobo	**malo**	con
	NC	***AQ***	***SP***

En la Figura 5 se muestra la integración y flujo del funcionamiento completo del algoritmo para obtener palabras candidatas a presentar una polaridad.

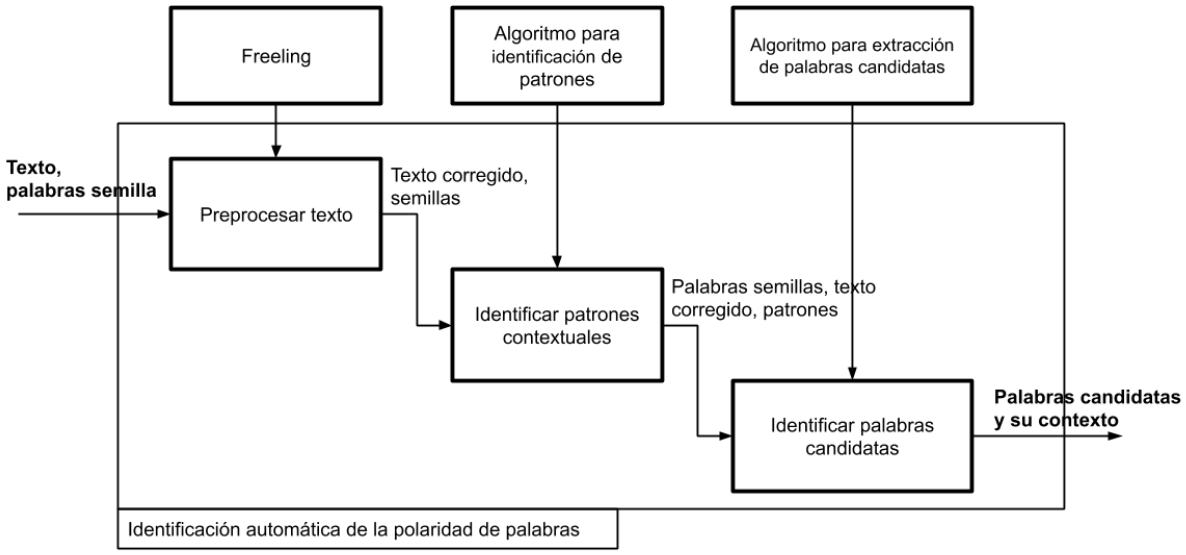

Figura 5 Flujo de funcionamiento de primera versión del algoritmo

Los resultados al experimentar con unigramas se presentan en la Tabla 18:

Tabla 18: Resultados al usar unigramas

Texto utilizado		Cuentos de 4885 palabras	
Léxico afectivo	**Contexto (unigramas)**	**Precisión**	**Cobertura (Recall)**
(Baca Gómez, 2014)	Izquierdo + derecho	0.11	0.29
	Izquierdo	0.10	0.53
(Sidorov, 2013b)	Izquierdo + derecho	0.07	0.29
	Izquierdo	0.07	0.53

Se realizó un segundo experimento utilizando bigramas (ver rendimiento en la Tabla 19), para realizar esta prueba se estableció que era mejor utilizar un archivo de configuración inicial el cual podría servir para indicar la cantidad de n-gramas anteriores y posteriores a evaluar, así se podrían realizar experimentos de diferentes cantidades de n-gramas sin tener que modificar directamente la lógica del algoritmo, por lo cual se agregó la función para soportar un archivo de configuración.

Tabla 19: Resultados al utilizar bigramas

Texto utilizado		Cuentos de 4,885 palabras	
Léxico afectivo	**Contexto (bigramas)**	**Precisión**	**Cobertura (Recall)**
(Baca Gómez, 2014)	Izquierdo + derecho	0.09	0.35
	Izquierdo	0.08	0.64
(Sidorov, 2013b)	Izquierdo + derecho	0.06	0.36
	Izquierdo	0.05	0.62

Los resultados de este experimento se dieron debido a que se comparaba que el contexto fuera idéntico, también se notó que la cobertura aumentaba cuando sólo se utilizaban patrones en posiciones anteriores a la palabra, lo que se observó fue que algunas de las categorías gramaticales presentadas en los patrones posteriores a la palabra eran signos de puntuación, es por esa razón que tampoco se hizo la prueba utilizando únicamente patrones posteriores.

Basándonos en los resultados obtenidos hasta ese punto, se observó que los patrones en los que se presentan las palabras con polaridad no eran completamente idénticos por lo que se optó por experimentar con una medida de similitud. Como se menciona en (Huang, 2008), el rendimiento de la similitud coseno, correlación de Jaccard y el coeficiente de Pearson son muy similares, y destacan sobre otras medidas. También se hace mención de que la medida de similitud coseno es una de las más populares en aplicarse a documentos de texto. Por ello, se modificó el algoritmo para representar a los patrones como un vector y encontrar su correlación, es decir, su similitud de coseno. Aunque no hay un estándar del valor aceptado para esta similitud, en este experimento se optó por utilizar una en la que cubriera al menos un 75%. Resultado de aplicar esta función dio lo que se muestra en la Tabla 20 y Tabla 21:

Tabla 20: Uso de unigramas y medidas de similitud

Texto utilizado		Cuentos de 4,885 palabras	
Léxico afectivo	**Contexto (unigramas)**	**Precisión**	**Cobertura (Recall)**
(Baca Gómez, 2014)	Izquierdo + derecho	0.05	0.59
	Izquierdo	0.9	0.63
(Sidorov, 2013b)	Izquierdo + derecho	0.08	0.63
	Izquierdo	0.06	0.63

Tabla 21: Utilización de bigramas y medida de similitud coseno

Texto utilizado		Cuentos de 4,885 palabras	
Léxico afectivo	**Contexto (bigramas)**	**Precisión**	**Cobertura (Recall)**
(Baca Gómez, 2014)	Izquierdo + derecho	0.07	0.61
	Izquierdo	0.08	0.66
(Sidorov, 2013b)	Izquierdo + derecho	0.05	0.59
	Izquierdo	0.05	0.64

Aunque la cobertura aumentó un poco, no fue un gran avance. De acuerdo con los resultados de los experimentos se observó que, efectivamente, palabras con polaridad, o al menos algunas de ellas pueden presentarse en un similar patrón sintáctico.

Después de realizar experimentos utilizando unigramas y bigramas se observó que los patrones en los cuales se presenta pueden variar y al convertir a vectores los patrones en los cuales se presentaban las palabras y medir su similitud mejoraron los resultados, pero sin llegar a ser aceptables.

De manera que se decidió hacer otros experimentos que permitieran identificar palabras candidatas con una precisión aceptable, las pruebas siguientes se hicieron sobre los textos recopilados y así obtener una mejor idea de optimización para el algoritmo.

4.3 Identificación de palabras candidatas

En esta sección se describen las actividades que sea realizaron para intentar incrementar la precisión del método al momento de identificar vocablos con polaridad. A continuación, se enuncia en forma general cómo fue dividido el trabajo:

- Identificación de la orientación de las palabras con polaridad dentro de una oración. Se buscó en qué posición aparece con más frecuencia una palabra con polaridad en lo que se refiere a la oración en la que se presenta, es decir, al inicio, en medio o al final de la oración.
- Análisis de la Ley de Zipf. Se experimentó con la ley con el propósito de conocer si era posible identificar palabras que tuvieran una polaridad con mayor facilidad.
- Evaluación de la técnica representación vectorial de palabras (en inglés, *word embeddings*). Utilizando la herramienta Word2vec se extrajeron los vectores con los cuales se puede representar una palabra en un documento, esto con el objetivo de utilizarlos en el algoritmo, reemplazando a los n-gramas de categorías gramaticales.
- Modificación al algoritmo utilizando los lemas que componen una oración. En esta actividad se hizo un cambio al algoritmo, ahora se usaron los lemas de la oración para generar vectores, y encontrar los que son similares.

4.3.1 Identificación de la orientación de las palabras con polaridad dentro de una oración

Se menciona en (Cheng & Lapata, 2016), que la mayoría de métodos se basan en características de ingeniería humana que se extraen de las oraciones tales como: posición y longitud de la oración, las palabras en el título, la presencia de nombres propios, características del contenido como la frecuencia de las palabras y características de evento como sustantivos de acción. Partiendo de lo anterior, se generó un algoritmo para obtener en qué posición aparece con más frecuencia una palabra con polaridad dentro de una oración, es decir, al inicio, en medio o al final de la oración. A continuación, se presentan algunos ejemplos de la posición de una palabra con polaridad en una oración.

- Su madre que sabía coser muy **bien** le había hecho una **bonita** caperuza roja que la niña nunca se quitaba, por lo que todos la llamaban Caperucita roja.
- Tan **triste** se puso aquella joven, que no tuvo más remedio que echarse a **llorar** durante toda la noche.
- Esa noche, mientras el viento rugía afuera arrastrando remolinos de nieve y colándose insolente por las rendijas, sintió el **miedo** visceral de la infancia.
- El gato desbordaba de **júbilo** y rápidamente fue a poner al tanto a su dueño, incapaz de comprender la estrategia de su felino.
- Le **dolía** cada fibra del cuerpo, pero lo más presente era la sed.
- La anciana era una mujer muy **mala** y el único motivo que tuvo para recoger lo de la entrada era usarlo como plato principal en una cena que preparaba.

Lo que se hizo fue ajustar la longitud de las oraciones y obtener la posición en un rango entre 1 y 100; por ejemplo, para la oración “Le **dolía** cada fibra del cuerpo, pero lo más presente era la sed” que tiene trece palabras, y la palabra con polaridad relacionada con el dolor que

se encuentra en la posición dos, se calcula de la siguiente forma: 2 entre 13 por 100, dando como resultado 15 (2 / 13 x 100 = 15%).

A continuación, se muestra en la Figura 6 cómo se distribuyen todas las palabras con polaridad, en la parte superior el corpus de novelas tomando el léxico afectivo de (Baca Gómez, 2014) como base, y en la parte de abajo se muestra la distribución ahora utilizando el léxico de (Sidorov, 2013b).

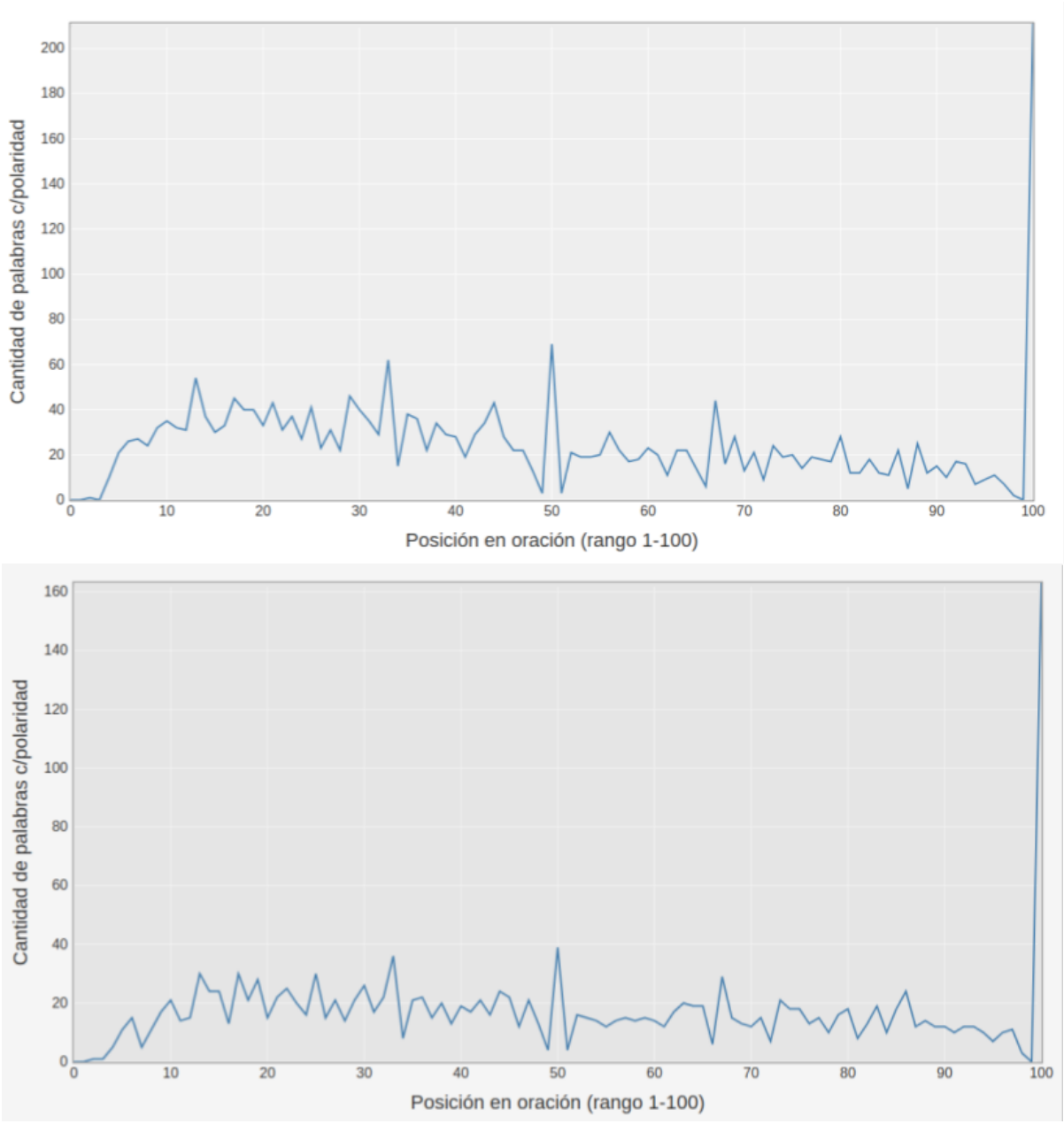

Figura 6 Distribución de palabras con polaridad en oración en novelas

Por otra parte, se realizó el experimento también para los cuentos infantiles, en la Figura 7 se muestra la distribución de las palabras con polaridad al usar los léxicos afectivos de (Baca Gómez, 2014) en la parte superior y (Sidorov, 2013b) en la parte inferior.

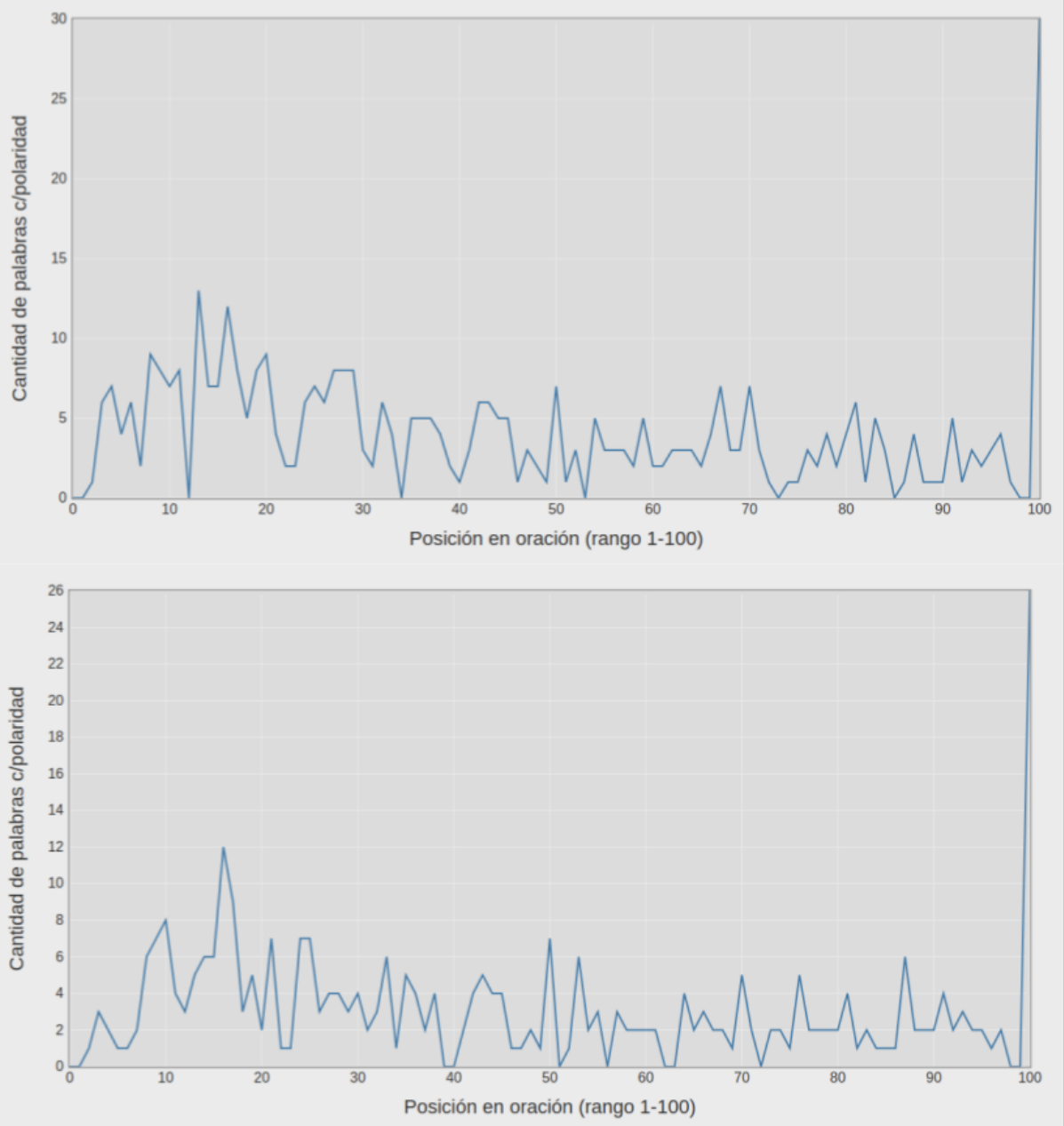

Figura 7 Distribución de palabras con polaridad en oración en cuentos

A simple vista las gráficas muestran que la mayoría de las palabras aparecen al final de la oración, pero para tener una mejor idea del sesgo de los vocablos se propuso agruparlos en tres rangos, si las palabras se presentan al inicio (1-33), en medio (34-66), o al final (67-100) en el rango de la oración; además, se propuso ya no colocar la cantidad de palabras sino la cantidad de oraciones en las que se presentan.

La Figura 8 muestran la cantidad de oraciones y su distribución en los tres rangos del corpus de la novela, se toma como base el léxico afectivo de (Baca Gómez, 2014) en la parte superior, y en la parte inferior se muestra la distribución al usar el léxico de (Sidorov, 2013b). Paralelamente, en la Figura 9 se muestran los resultados al procesar los cuentos infantiles. Las columnas muestran como las palabras tienen una frecuencia de aparición menor al inicio de una oración, mayor al final y en medio tienen una frecuencia de aparición intermedia.

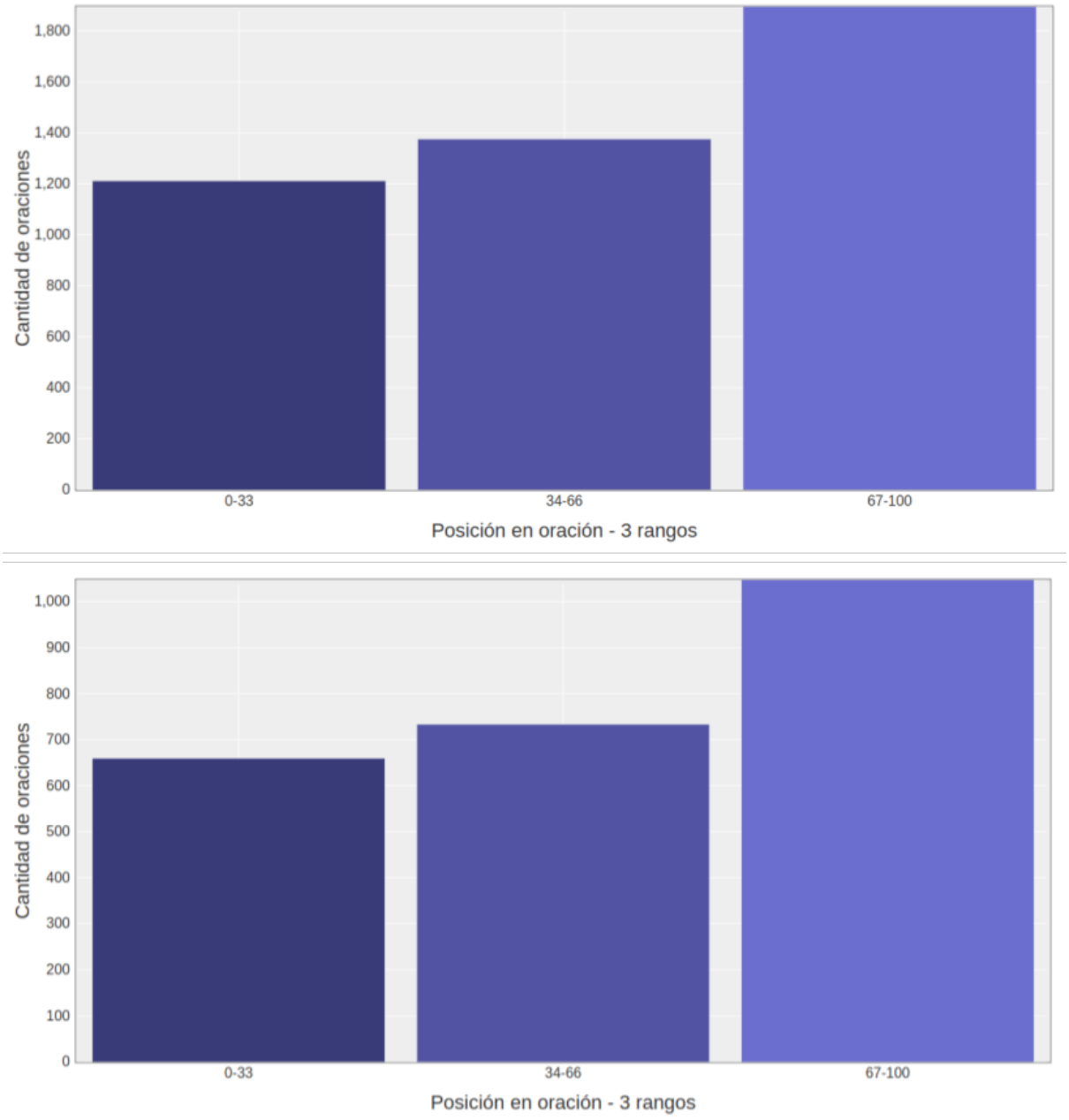

Figura 8 Distribución de palabras con polaridad (agrupación en 3 rangos) en novelas

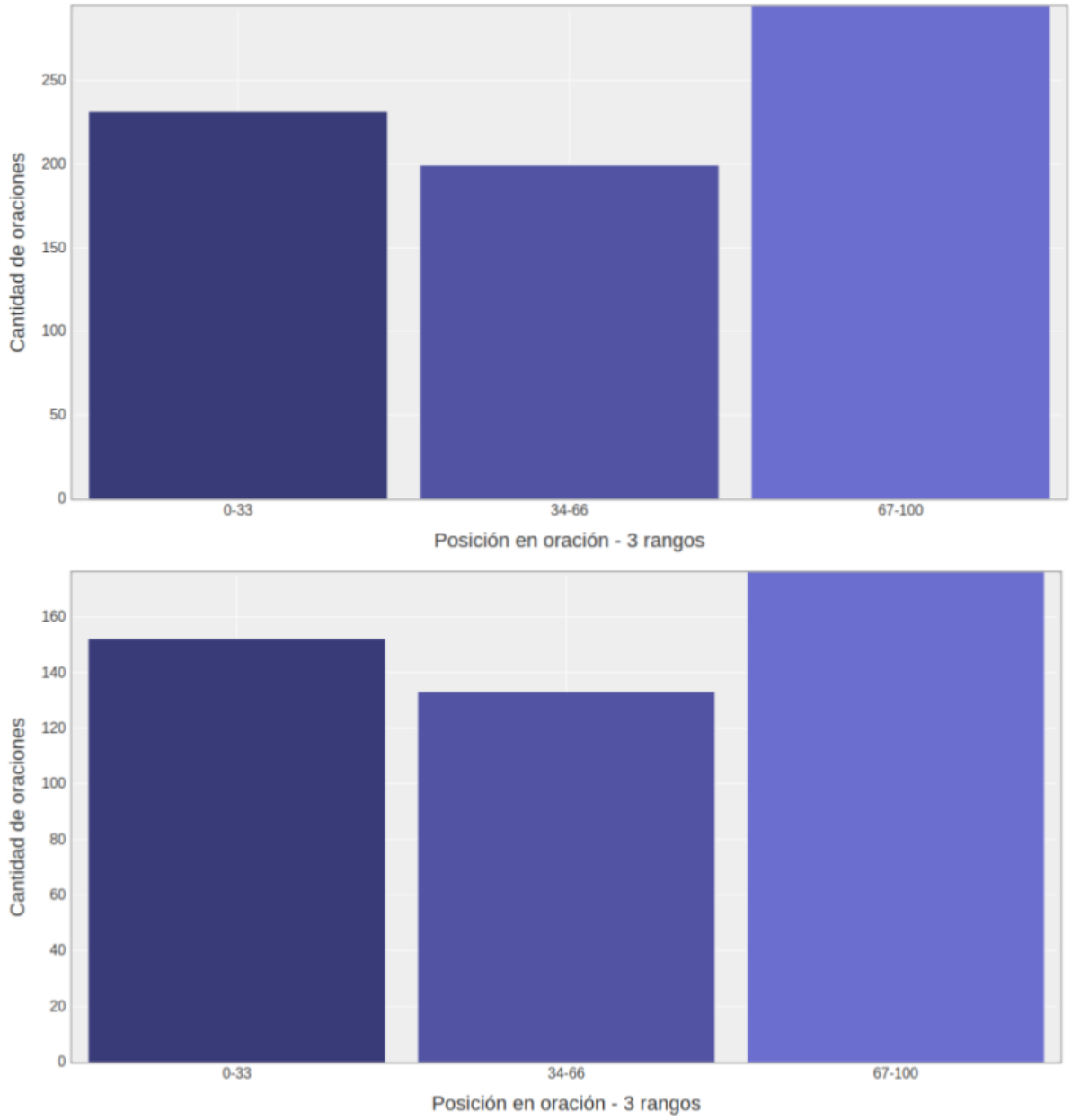

Figura 9 Distribución de palabras con polaridad (agrupación en 3 rangos) en cuentos

Hasta este punto se obtenían todas las palabras con polaridad de todas las categorías, por lo que se modificó con el fin de observar cómo se comporta el sesgo de acuerdo con una categoría gramatical, se presentan los resultados al buscar verbos únicamente, y después se muestra lo obtenido al identificar solo los adjetivos.

En esta parte las gráficas se agrupan por categoría gramatical, presentando en la Figura 10 siguiente el comportamiento de los verbos en el texto de novelas (lado superior) y en los cuentos infantiles (lado inferior). Se utilizaron ambos léxicos afectivos para realizar las pruebas, el de (Baca Gómez, 2014) para el resultado del lado izquierdo y de (Sidorov, 2013b) para el lado derecho.

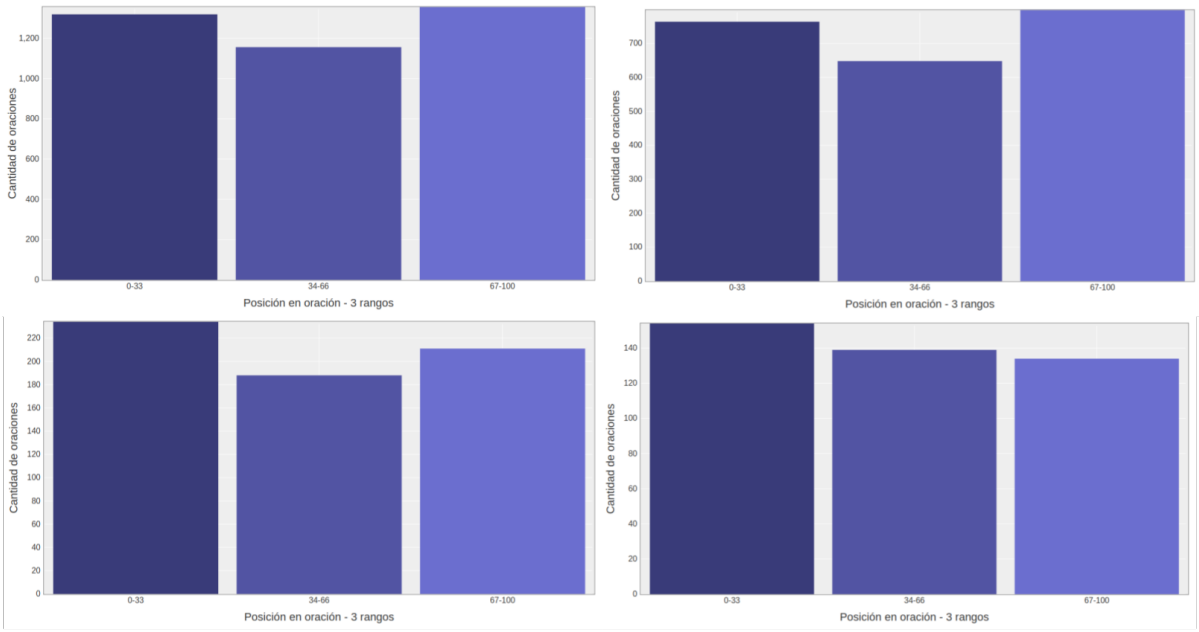

Figura 10 Distribución de verbos con polaridad en una novela (parte superior) y cuentos infantiles (parte inferior)

En la Figura 11 se muestra cómo se distribuyen los adjetivos con polaridad en los corpus.

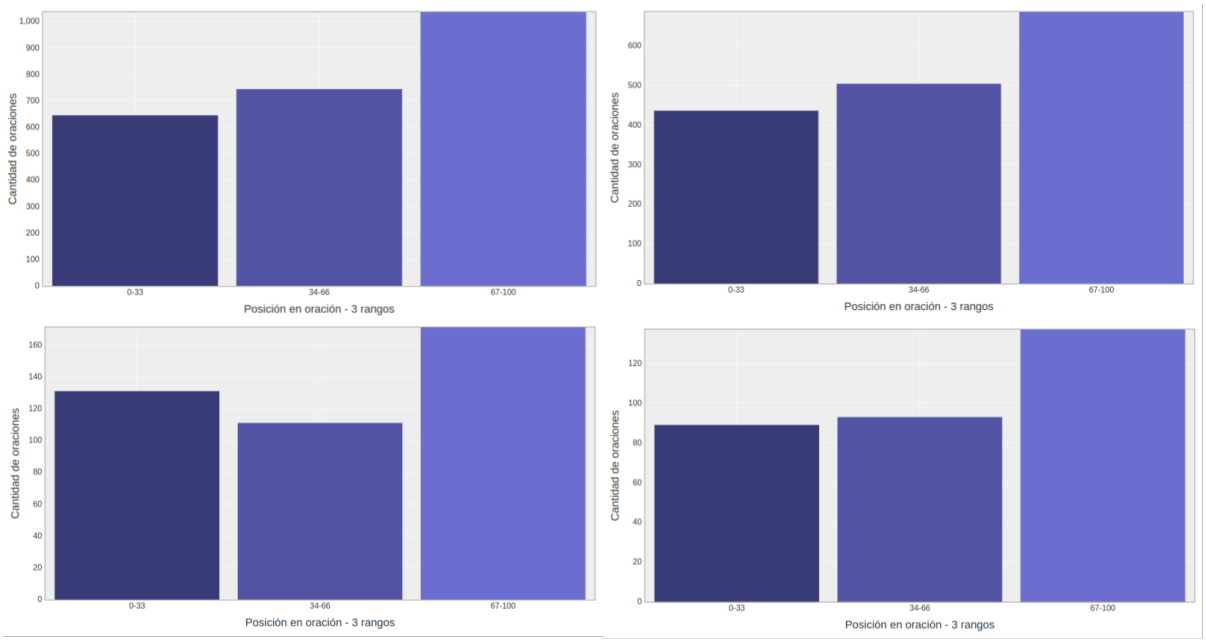

Figura 11 Distribución de adjetivos con polaridad en una novela (parte superior) y cuentos infantiles (parte inferior)

También, se presentan las palabras que no fueron identificadas como verbos o adjetivos, a continuación se muestra el resultado (ver figura siguiente).

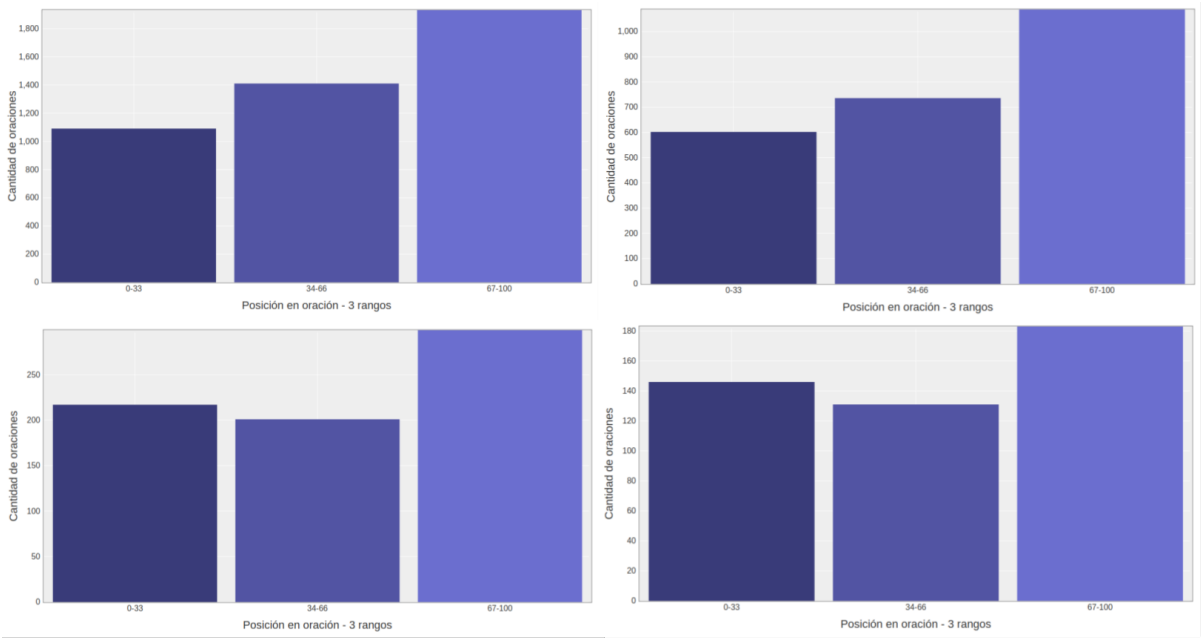

Figura 12 Distribución de otras categorías con polaridad en una novela (parte superior) y cuentos infantiles (parte inferior)

Derivado de los experimentos anteriores, se hicieron modificaciones al algoritmo propuesto con el objetivo de incrementar su rendimiento. Los principales cambios se enlistan a continuación:

- Cuando es una palabra de tipo verbo, únicamente se toma en cuenta las que aparecen al final o al inicio de la oración.
- Cuando la palabra identificada es un adjetivo, solo es agregado si se ubica al final o en medio de la frase.
- Para los demás tipos de palabras candidatas se toman como prioridad las palabras candidatas que aparecen al final de la oración.

Aplicando estos cambios al algoritmo, se obtuvieron los siguientes resultados (ver Tabla 22 y Tabla 23).

Tabla 22: Comportamiento al aplicar cambios a algoritmo, usando unigramas

Texto utilizado		Cuentos de 14,036 palabras	
Contexto		Unigramas a la izquierda	
Léxico afectivo	**Versión del algoritmo**	**Precisión**	**Cobertura (Recall)**
(Baca Gómez, 2014)	Anterior	0.18	0.49
	Nueva	0.13	0.55
(Sidorov, 2013b)	Anterior	0.12	0.57
	Nueva	0.08	0.58

Tabla 23: Resultado al aplicar cambios a algoritmo, usando bigramas

Texto utilizado		Cuentos de 14,036 palabras	
Contexto		Bigramas a la izquierda	
Léxico afectivo	**Versión del algoritmo**	**Precisión**	**Cobertura (Recall)**
(Baca Gómez, 2014)	Anterior	0.12	0.57
	Nueva	0.10	0.66
(Sidorov, 2013b)	Anterior	0.07	0.57
	Nueva	0.06	0.63

En las tablas se puede observar que el cambio fue mínimo; además, se puede notar que no mejora el resultado al utilizar el léxico afectivo de (Sidorov, 2013b), dando valores similares a los del punto 4.2.2.

Acto seguido, se propuso experimentar con la Ley de Zipf, haciendo un análisis en dónde podrían agruparse las palabras con polaridad según su frecuencia.

4.3.2 Aplicación de la Ley de Zipf

Las palabras en textos en lenguaje natural, sea en español u otra lengua, suelen tener coocurrencias determinadas, es decir, es más probable que ciertas palabras vayan detrás de otras y es más probable que existan más palabras detrás de unas que de otras. Esto tiene que ver con la Ley de Zipf que afirma que un pequeño número de palabras son utilizadas con mucha frecuencia, mientras que ocurre que un gran número de palabras son poco empleadas (Muñoz & Álvarez, 2014).

En general, la mayoría de las palabras frecuentes son también las más cortas y más fáciles de recordar, estas son las palabras funcionales (también llamadas palabras vacías o *stop words*, en inglés), tales como artículos, pronombres, preposiciones y conjunciones son las más frecuentes en el texto, mientras que las menos frecuentes son palabras que reflejan el estilo y riqueza del vocabulario del autor. Por lo tanto, las palabras que aparecen en la zona media de transición entre las de alta y baja frecuencia de ocurrencia son las que con frecuencia representan al documento (Moyotl-Hernández & Macías-Pérez, 2016).

Derivado de lo expuesto acerca de esta ley, se propuso realizar algunos experimentos con el objetivo de observar cual es la frecuencia de ocurrencia de las palabras con polaridad y así aumentar el rendimiento del método propuesto, basados en la idea de que las palabras con polaridad de un texto se agrupan en una zona específica, o como se menciona en el artículo (Moyotl-Hernández & Macías-Pérez, 2016), donde se encuentran las palabras de alta frecuencia con las de baja frecuencia. Se muestra en la Figura 13, dónde aparecen las palabras con polaridad en el texto de la novela. Para esta actividad se toman como base los léxicos afectivos de (Baca Gómez, 2014) en la parte superior de la imagen, y el de (Sidorov, 2013b) en la parte inferior.

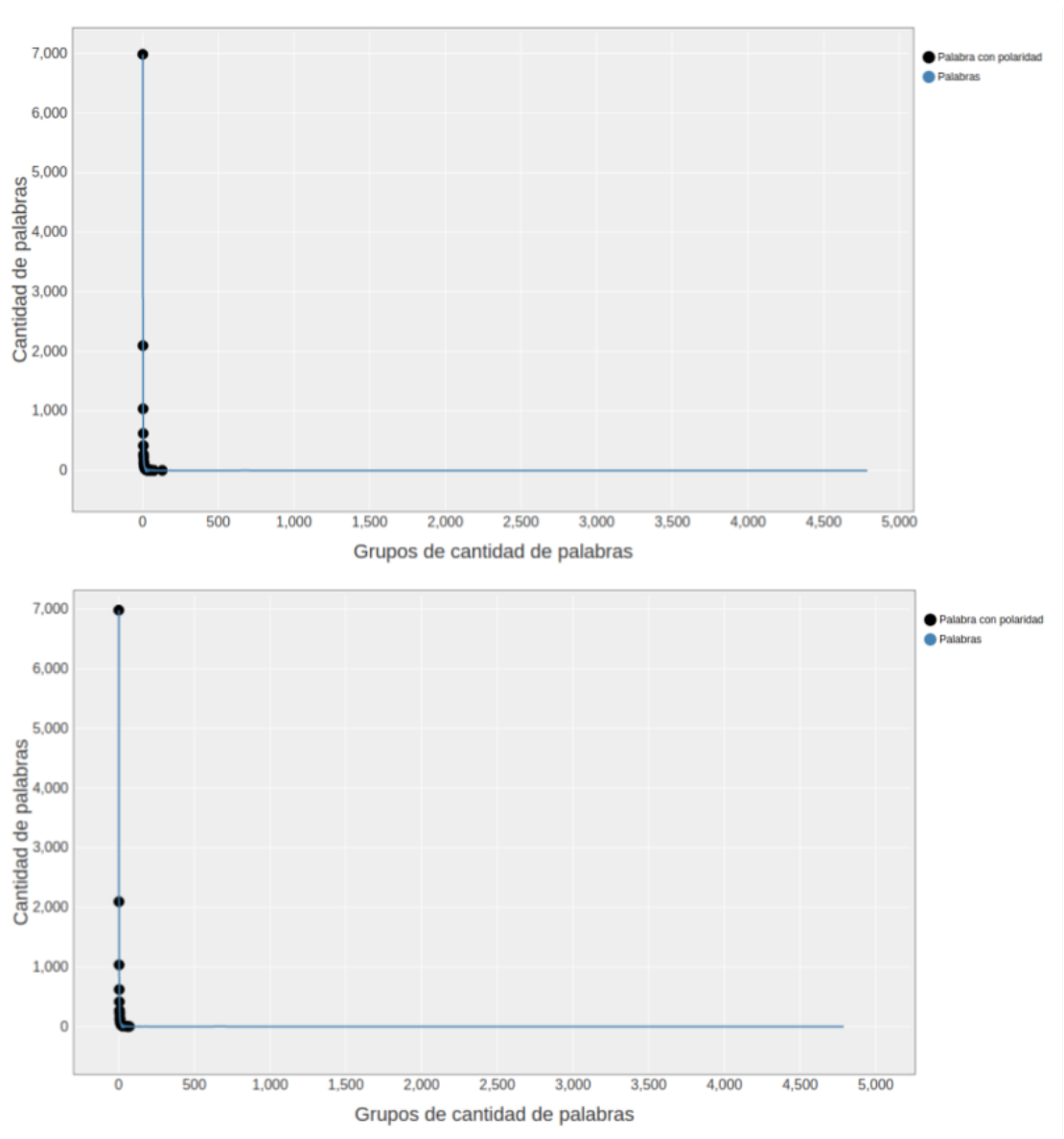

Figura 13 Agrupamiento de palabras según su aparición en novelas

Se aplicó el mismo procedimiento al grupo de cuentos infantiles, dando como resultado lo que se muestra en la Figura 14. En la parte superior se utiliza el léxico de (Baca Gómez, 2014) para extraer las palabras con polaridad, y en la parte inferior el de (Sidorov, 2013b).

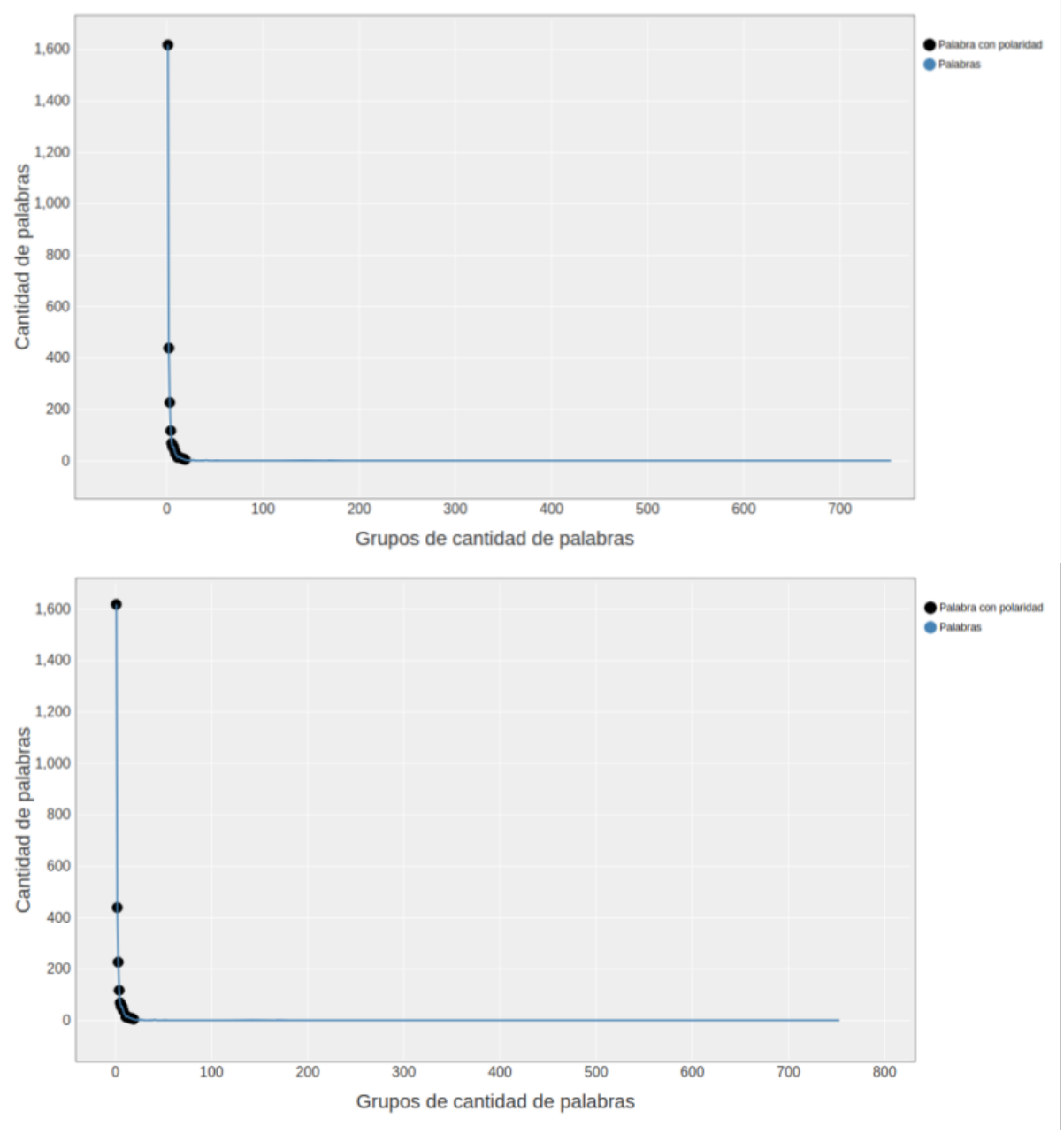

Figura 14 Agrupamiento de palabras en cuentos

El resultado de aplicar la Ley de Zipf proporcionó otras ideas de cómo una palabra puede estar relacionada dentro de un texto, una de las ideas que se tuvo fue, ya que una palabra puede aparecen en una zona específica relacionada con su frecuencia, es posible que al obtener su representación vectorial se pueda relacionar con otras palabras con un similar vector. Para eso se optó por evaluar la técnica llamada representación continua de palabras (en inglés, *word embedding*), esto se explica en la siguiente actividad.

4.3.3 Emplear la técnica representación vectorial de palabras (*word embedding*)

La motivación de probar esta técnica fue debido a los resultados obtenidos en la actividad

anterior, por otra parte, es muy utilizada en el estado del arte, ya que como se menciona, la información que proporcionan los *word embedding* sirven de base para el análisis de sentimientos, ampliación de lexicones, detección similitud de textos, o traducción de texto, entre otros.

Las técnicas de *word embedding* parten de representaciones bolsas de palabras (en inglés, *bag of words* o *BOW*) de los distintos contextos de las palabras para obtener representaciones vectoriales de las palabras de dimensiones mucho más reducidas que capturan el significado y las relaciones entre palabras. Hay diversas técnicas para calcular estas representaciones, una de las más empleadas se basa en redes neuronales de una sola capa oculta que predicen la palabra dado el contexto o viceversa, adaptando así una de las piezas básicas de los modelos de aprendizaje profundo, los auto-codificadores (López-Solaz, Troyano, Ortega, & Enríquez, 2016).

La herramienta que se implantó en esta actividad para obtener la representación vectorial de las palabras es Word2vec de (Mikolov, Le, & Sutskever, 2013), disponible para el lenguaje de programación Python que es parte de Gensim (Sojka, 2010). Al ingresar el corpus a la herramienta, está arrojo por cada palabra una matriz como la que se muestra en la Figura 15, aquí solo se muestra un ejemplo de los posibles valores que contiene un vector de este tipo.

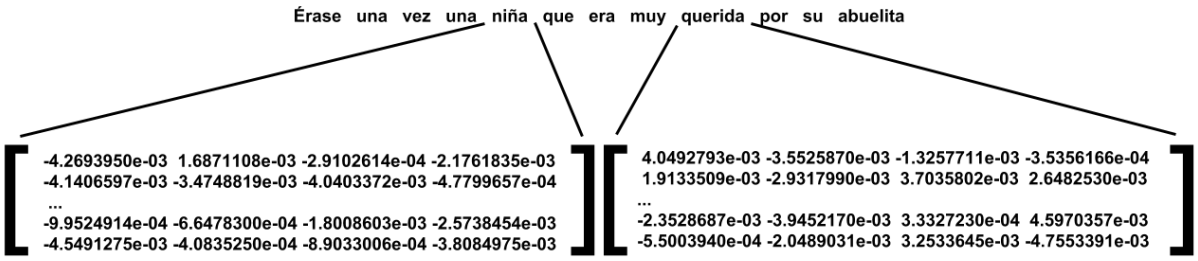

Figura 15 Ejemplo de la representación vectorial de palabras

La herramienta se utiliza para buscar palabras similares a las semillas que el algoritmo recibe como entrada, esto se obtiene con las mismas funciones que Word2vec ofrece. Al realizar el experimento se obtuvieron los resultados que se muestran en la Tabla 24.

Tabla 24: Palabras identificadas con la herramienta Word2vec

Texto	Corpus novelas		Corpus cuentos	
Léxico afectivo	(Baca Gómez, 2014)	(Sidorov, 2013b)	(Baca Gómez, 2014)	(Sidorov, 2013b)
Palabras obtenidas	190		120	
Palabras con polaridad	16	14	6	5

Conforme a esto, se optó por utilizar los *word embedding* de otra forma; buscando simular el algoritmo actual se propuso una modificación, que consiste en utilizar los vectores arrojados

por la herramienta de Word2Vec en lugar de la categoría gramatical anterior o posterior que aparece junto a una palabra con polaridad. En la Figura 16 se muestra el flujo del algoritmo modificado.

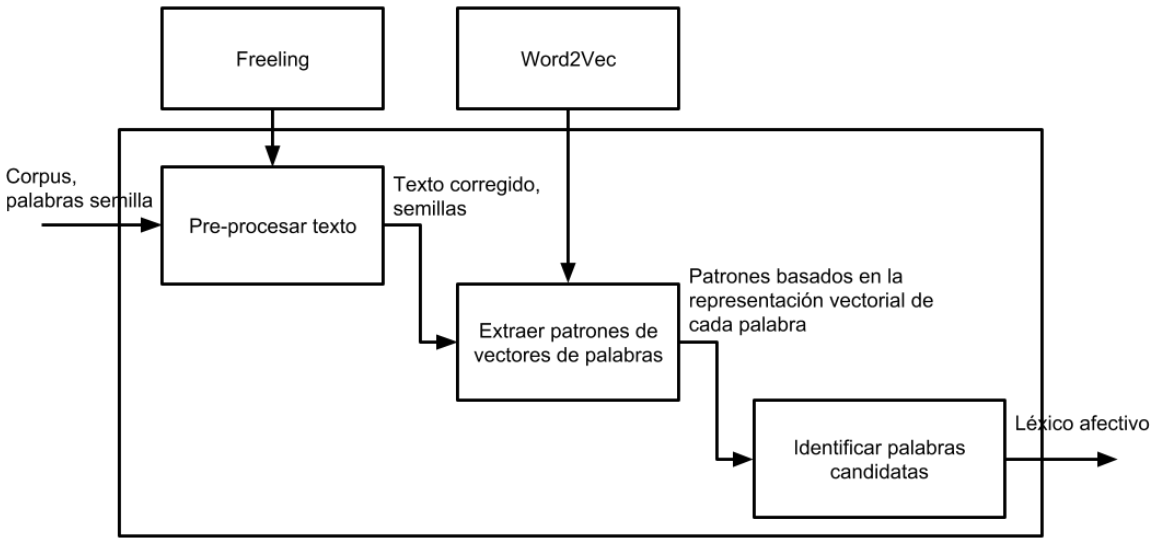

Figura 16 Algoritmo utilizando word embedding

Esta versión del algoritmo sigue utilizando la idea de los n-gramas a la izquierda y derecha de la semilla, pero ahora con los vectores de los contextos. Un ejemplo se muestra en la Figura 17, se obtienen los patrones, y con estos, el algoritmo compara si otras palabras aparecen rodeadas de vectores similares, es decir, que estén en un mismo contexto.

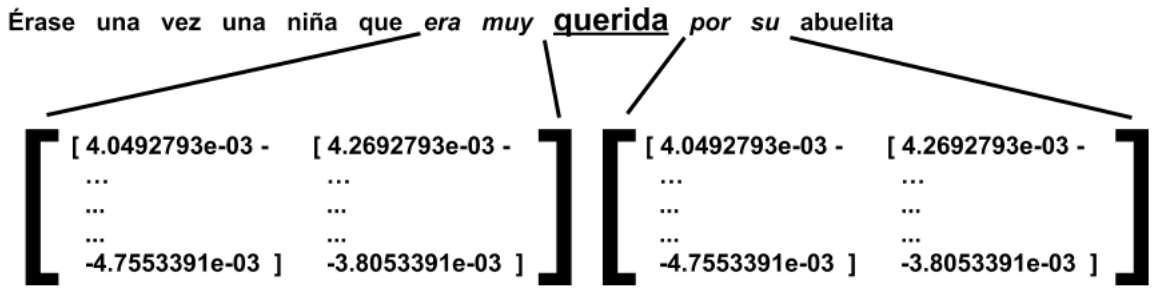

Figura 17 Extracción del contexto usando la representación vectorial de palabras

Se muestra el resultado obtenido de este experimento (ver Tabla 25); para este experimento se utilizó el léxico afectivo de (Baca Gómez, 2014).

Tabla 25: Resultados de algoritmo basado en la representación vectorial de palabras

Texto	Corpus novelas		Corpus cuentos	
N-gramas	Unigramas	Bigramas	Unigramas	Bigramas
Precisión	0.08	0.18	0.21	0.35
Cobertura	0.02	0.02	0.10	0.09

Como se puede observar los resultados no fueron óptimos, son bajos comparado con el rendimiento de las versiones anteriores del algoritmo. Conforme al estado del arte la explotación de este tipo de vectores puede dar buenos resultados, por lo que se concluyó de esta actividad es necesario una experimentación más profunda con ellos, podría considerarse como un trabajo futuro.

4.3.4 Modificación al algoritmo utilizando los lemas que componen una oración

De acuerdo con lo visto en el estado del arte, la mayoría de métodos no se basan únicamente en una sola palabra, sino que utilizan las características del texto completo o de sus oraciones, algunos ejemplos de ello se mencionan en (Tribhuvan, Bhirud, & Tribhuvan, 2014), (Fu, He, Song, & Wang, 2015), (Akkarapatty & Raj, 2016).

Partiendo de esta idea, se decidió desarrollar un programa ya no basándose en el contexto de las palabras, sino en la composición de la oración. Por otra parte, resultado de los experimentos anteriores ayudó a percibir que, aunque en su mayoría es correcta la asignación de la categoría gramatical de las palabras, Freeling tiene algunos desaciertos, por ejemplo, etiquetando a un adjetivo como verbo o a la inversa; debido a lo anterior se optó por utilizar lemas en lugar de la categoría gramatical, como se hacía anteriormente. A continuación, se describe el funcionamiento de esta versión del algoritmo.

1. Como entrada recibe el corpus y las palabras semillas.
2. Enseguida, el texto se procesa con Freeling y se divide en oraciones, obteniendo por cada oración la matriz de los lemas que la componen.
3. Después, se buscan los vectores en donde aparecen las palabras semillas.
4. Teniendo los vectores de las palabras semillas, se compara su similitud con otros.
5. Finalmente, como salida se obtiene una lista de oraciones con posible polaridad, es decir que pueden contener alguna palabra con polaridad.

Para tener una mejor idea de los pasos a seguir del programa, se plasma el proceso del algoritmo en el siguiente diagrama SADT (Figura 18).

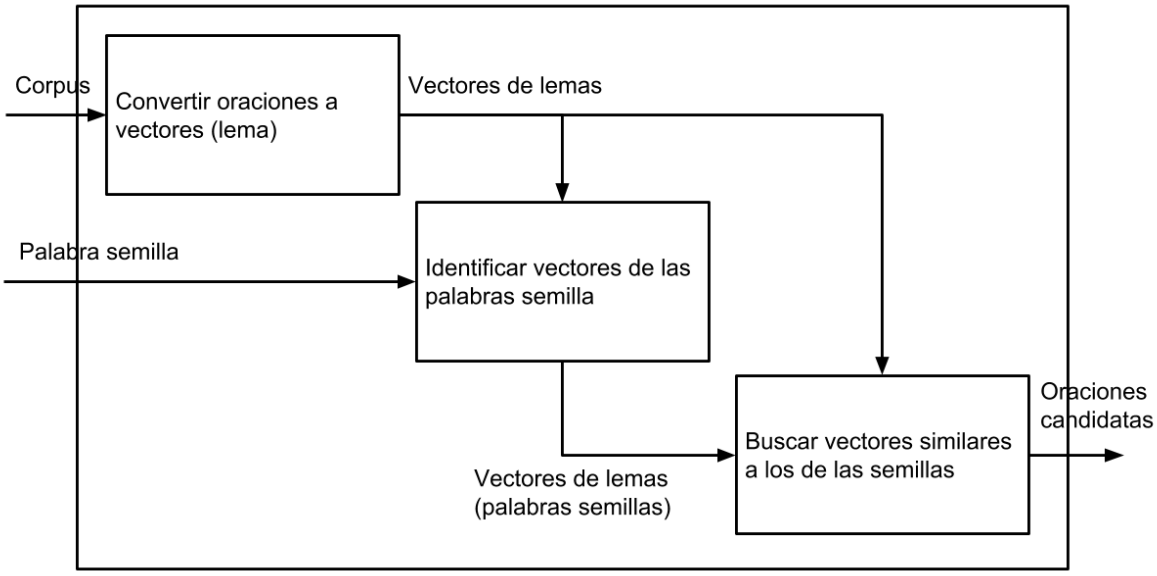

Figura 18 Algoritmo usando vectores de lemas de la oración completa

Como se observa, este algoritmo no detecta la palabra con polaridad, sino que detecta la posible oración que puede tener polaridad.

Para realizar la validación de las oraciones, se hizo una comparación de las palabras que estas contenían con las palabras con polaridad de los léxicos afectivos de (Baca Gómez, 2014) y (Sidorov, 2013b) obteniendo como resultado lo presente en la Tabla 26.

Tabla 26: Comportamiento de algoritmo basado en vector de lemas

Texto	Corpus novelas		Corpus cuentos	
Léxico afectivo	(Baca Gómez, 2014)	(Sidorov, 2013b)	(Baca Gómez, 2014)	(Sidorov, 2013b)
Precisión	0.49	0.30	0.54	0.35
Cobertura	0.44	0.51	0.41	0.42

En este experimento se observa un avance significativo en comparación con las versiones anteriores del algoritmo, se demuestra una mejora en la precisión y cobertura. Pero como ya se mencionó, el resultado es un conjunto de oraciones que pueden presentar palabras con polaridad; sin embargo, lo esperado sería obtener un conjunto de vocablos con polaridad que sirvan de apoyo para formar un léxico afectivo u otros métodos que utilicen este tipo de recursos, por lo cual fue necesario reconsiderar focalizar el trabajo en mejorar el método de identificación de palabras.

Capítulo 5

Evaluación

En este capítulo se presentan las pruebas realizadas y resultados obtenidos en el trabajo de tesis “Exploración para la identificación automática de palabras con polaridad en el texto escrito”.

Se realizaron diversas pruebas experimentando con los parámetros de entrada (palabras con polaridad -palabras semillas- y un corpus de texto sin anotar) con la intensión de identificar la combinación de variables que permitieran obtener la mejor precisión y cobertura.

La cronología de las pruebas se resume en los siguientes puntos:

- Se determinó el criterio que definiría qué tamaño de n-grama se considerarían como patrones contextuales y cuáles no (explicación en punto 5.1.2).
- A partir de las características de los patrones, se realizaron diversos experimentos haciendo variaciones al algoritmo de análisis usando los datos de entrada y se compararon los resultados con el propósito de elegir las dos mejores combinaciones en términos de precisión y cobertura (explicación en el capítulo 5.1.3.1).
- Se realizaron más experimentos, pero considerando únicamente las dos mejores versiones, manteniendo los parámetros de entrada y variando sus cantidades y tamaños para observar su comportamiento. De los experimentos con las dos mejores versiones, se eligió la mejor (explicación en el capítulo 5.1.3.2).
- Se realizaron los siguientes experimentos con la mejor variación (explicación en el capítulo 5.1.4 y 5.1.5):
    - Se incrementó el número de semillas y se agruparon por su frecuencia (alta, media, baja) de aparición en los corpus. Además, se hicieron pruebas considerando la categoría gramatical, agrupando las semillas en verbos y adjetivos, y otro más pequeño con las demás categorías.
    - Se probó usando lemas como patrones contextuales y, además, se evaluaron las diferencias de resultados al cambiar las entradas del programa, siendo: usar distintas cantidades de semillas (20, 50, 200 y 500), aplicar el corpus incrementado, emplear la fusión de los léxicos afectivos y agrupar semillas con una polaridad (negativa, positiva).

5.1 Ejecución de experimentos

Bajo la presunción de que el contexto de las palabras con polaridad puede aparecer con cierta regularidad y que esto permite predecir la ubicación de nuevas palabras (denominadas palabras candidatas), se desarrolló un algoritmo que analiza los n-gramas a la izquierda y a

la derecha de las palabras con polaridad con la intensión de identificar patrones en ellas. Con esto se podría partir de un número reducido de palabras con polaridad predefinidas manualmente (denominadas palabras semilla) para extraer, vía sus contextos, nuevas palabras a partir de los contextos de las semillas.

En este capítulo se describen los experimentos para encontrar los patrones contextuales que arrojaron los mejores resultados en términos de precisión y cobertura.

5.1.1 Recursos utilizados para la ejecución de experimentos

Para la ejecución de las pruebas se utilizaron dos léxicos afectivos y dos corpus. Los corpus no están etiquetados y son usados para extraer los patrones contextuales y las palabras candidatas. En la Tabla 27 se muestra su contenido.

Tabla 27 Características del corpus

Corpus completo		
Identificador	**Corpus**	**Cantidad de palabras**
C1	Corpus de novelas	89,281
C2	Corpus de cuentos	14,036
Cantidad total de palabras		103,317

Los léxicos afectivos se usaron para identificar en los corpus el universo de palabras con polaridad que el sistema debería de obtener, es decir, la línea tope. En la Tabla 28 se presentan las características de estos recursos.

Tabla 28 Características de léxicos afectivos

Identificador	**Léxico afectivo**	**Cantidad de palabras**
LA1	(Baca Gómez, 2014)	2,036
LA2	(Sidorov, 2013)	3,550

La línea tope de cada corpus se muestra en la Tabla 29, relacionada con su identificador, para cada uno de los léxicos afectivos.

Tabla 29 Cantidad de palabras con polaridad en cada corpus

	Cantidad de palabras con polaridad	
Corpus	**LA1**	**LA2**
C1	4,434	2,425
C2	717	460

Por otra parte, se hizo una unión y depuración de los léxicos afectivos y un incremento en el corpus (ver Tabla 30).

Tabla 30 Cambios en recursos aplicados a experimentos

Léxicos afectivos		**Corpus**	
Recurso	**Cantidad de palabras**	**Recurso**	**Cantidad de palabras**
Unión de (Baca Gómez, 2014) y (Sidorov, 2013)	4,080	Corpus de novelas y cuentos	373,792

Esta unión se realizó para definir la mejor frecuencia de aparición de patrones que sirve para la identificación de palabras candidatas (ver capítulo 5.1.2.3) y para los últimos experimentos utilizando lemas que hacen una variación con sus recursos (ver capítulo 5.1.5.3).

5.1.2 Criterio para considerar patrones

Por patrón se entiende a la serie de sucesos o elementos que ocurren de manera recurrente. Aunado a ello, se considera en este trabajo que un patrón es un n-grama que se repite con cierta frecuencia junto a una palabra semilla.

Para decidir que un n-grama es un patrón se requieren tomar tres consideraciones:

- La ubicación del patrón respecto a la palabra con polaridad, esto es, si el patrón se encuentra ubicado a la izquierda, a la derecha o en ambos lados de la palabra (combinación de n-gramas).
- La determinación del valor de n, es decir, si los patrones se conformarán por unigramas, bigramas, trigramas, etc.
- La frecuencia mínima de aparición, es decir, cuántas veces debe aparecer el n-grama

junto a la palabra semilla para ser considerado patrón.

A continuación, se explicará cómo se abordó cada punto.

5.1.2.1 Ubicación de los patrones

Para el primer caso se realizó un experimento que consistió en determinar la ubicación promedio de las palabras emocionales en una oración. Interesaba saber si tenían una distribución uniforme a lo largo de las oraciones o tendían a aparecer al inicio, en medio o al final de ellas. Se extrajeron 6645 oraciones del corpus de novelas que contenían palabras emocionales del léxico afectivo de (Baca Gómez, 2014). Dada la variabilidad de las longitudes de las oraciones, se decidió normalizar la posición de las palabras emocionales en las oraciones considerando una base de normalización de valor 100. Por ejemplo, en la oración:

*"Le **dolía** cada fibra del cuerpo, pero lo más presente era la sed"*

Se toma la posición de la palabra emocional "dolía", la cual es dos, se divide entre la longitud de la oración, que es 13, y se multiplica por el valor de 100, lo que da como resultado un valor de 15. De esta manera es posible establecer tres rangos: si el valor resultante tiene un valor entre 1 y 33, se considera que la palabra emocional aparece al inicio, si tiene un valor entre 34 y 66, se considera en medio, y por último, un valor entre 67 a 100, se considera al final de la oración.

En la Figura 19 se indican los resultados de la ubicación de las palabras con polaridad. Las columnas muestran que dichas palabras tienen una frecuencia de aparición menor al inicio de las oraciones, mayor al final y en medio tienen una frecuencia de aparición intermedia.

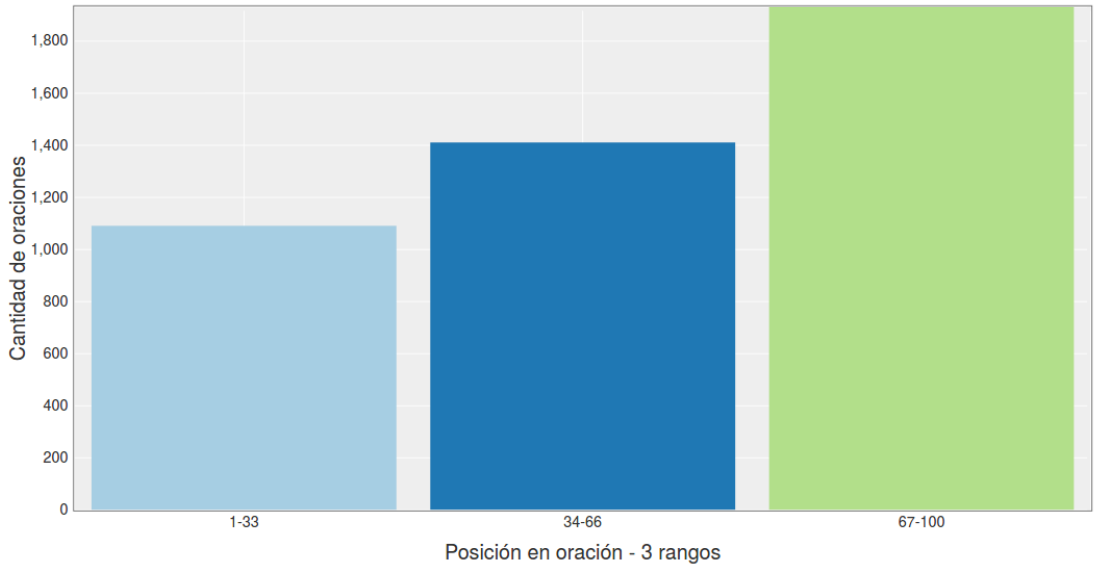

Figura 19 Distribución de la ubicación de palabras con polaridad

Dada esta distribución, se considera más probable que exista mayor cantidad de patrones ubicados a la izquierda de las palabras con polaridad, que a la derecha. Por tal motivo, se decidió hacer pruebas sólo con n-gramas ubicados a la izquierda de las palabras con polaridad.

5.1.2.2 Determinación del valor de *n*

Referente a la longitud que debe de tener el n-grama, se realizaron pruebas tomando desde una palabra hasta cinco. Se utilizaron los léxicos afectivos de la Tabla 28 y el corpus de cuentos. En la Tabla 31 se muestran los valores de precisión y cobertura para cada longitud asignada al n-grama.

Tabla 31 Selección de longitud de n-gramas para pruebas

Léxico afectivo	**Contexto**	**N-gramas**	**Precisión**	**Cobertura**
LA1	Izquierdo	1	0.14	0.39
		2	0.47	0.15
		3	0.09	0.13
		4	0.08	0.24
		5	0.11	0.28
LA2	Izquierdo	1	0.10	0.42
		2	0.42	0.19
		3	0.03	0.07
		4	0.08	0.33
		5	0.11	0.36

Se observa que los valores más altos se obtienen con bigramas, por lo cual, se consideraron éstos para el resto de experimentos.

5.1.2.3 Frecuencia de ocurrencia mínima de los n-gramas

Respecto a la frecuencia de ocurrencia, se consideró que los bigramas tuvieran una frecuencia mínima que permitiera obtener los valores más altos de precisión en la predicción de palabras con polaridad. Para identificar este valor mínimo, se consideraron los siguientes aspectos:

1. Se realizaron pruebas con frecuencias de 2, 5, 10 y 15 ocurrencias mínimas.

2. Se utilizaron como semillas palabras de categoría adjetivos, ya que es la categoría gramatical que más palabras emocionales tiene y por lo tanto se tendrían más muestras del corpus analizado.
3. Las pruebas se hicieron sobre 10, 20 y 50 semillas en el corpus expandido (ver Tabla 30). En la Tabla 32 se muestran los resultados obtenidos.

Tabla 32 Experimentos para determinar la frecuencia de aparición de patrones

Cantidad de semillas	**Frecuencia**	**Precisión**	**Cobertura**
10	2+	0.23	0.12
	5+	0.26	0.07
	10+	0.31	0.04
	15+	**0.37**	**0.03**
20	2+	0.23	0.14
	5+	0.26	0.08
	10+	0.26	0.06
	15+	**0.35**	**0.03**
50	2+	0.23	0.14
	5+	0.23	0.11
	10+	0.25	0.07
	15+	0.26	0.06

Los resultados obtenidos muestran que los mejores patrones en términos de precisión son aquellos que tienen una frecuencia mínima de 15 apariciones, tomando ya sea 10 semillas o 20. Por tal motivo, el resto de pruebas se realizarán considerando patrones con dicha frecuencia mínima.

Con estos experimentos, se logra precisar que los patrones tienen las siguientes características:

- Se ubican a la izquierda de las palabras con polaridad.
- Los que permiten obtener los mejores resultados son bigramas.
- Su frecuencia mínima de aparición es de 15.

Considerando estas características, se realizaron diversos experimentos para obtener las mejores combinaciones de variables que permitieran obtener las predicciones de palabras emocionales con los valores más altos de precisión y cobertura. Estos experimentos son explicados a continuación.

5.1.3 Experimentos

Se realizaron diferentes experimentos para encontrar la combinación de variables que permitieran obtener los valores más altos de precisión y cobertura. A continuación, se describen cada uno de los experimentos.

5.1.3.1 Generación de diversas versiones del algoritmo

Se realizaron diferentes pruebas para determinar la combinación de variables que permitieran obtener los valores más altos de precisión y cobertura. Obtener la mejor combinación depende de diversos criterios, por ejemplo:

a) ¿Cuántas semillas es mejor utilizar?
b) ¿Cómo patrones contextuales es mejor usar lemas o categorías gramaticales?
c) ¿Tomar en cuenta la frecuencia de aparición de las semillas en el corpus permitirá mejorar los resultados?
d) ¿Se desempeñará mejor el algoritmo si se considera la categoría gramatical de las semillas?

El algoritmo de análisis de contexto se modificó para adecuarlo al tipo de análisis que se implementó. Cada modificación generó una nueva versión. En la Tabla 33 se listan las versiones y el funcionamiento de los algoritmos que se desarrollaron.

Tabla 33 Versiones y funcionamiento del algoritmo

#	Versiones de algoritmo	Funcionamiento
1	Contextos obtenidos de forma manual	Esta versión no utiliza semillas, ya que estos contextos se obtuvieron de forma manual del corpus y se insertaron directamente al programa.
2	Categorías gramaticales de los bigramas	Esta versión recibe como entrada semillas, busca y toma como contexto la categoría gramatical de los bigramas y aplica comparación por similitud coseno para encontrar palabras candidatas.
3	Palabras candidatas de acuerdo con su posición dentro de la oración	Esta versión usa semillas, comparación por similitud y aplica el filtro a las palabras candidatas de acuerdo con como tienden a distribuirse con mayor frecuencia del medio hacia el final de la oración.

Como complemento y para una mejor comprensión del funcionamiento de los algoritmos, en el Anexo 8.1, se presenta el pseudocódigo de las diferentes versiones de estos. La comparación entre resultados se muestra en la Tabla 34.

Tabla 34 Resultados obtenidos de las versiones del algoritmo

#	Versiones del algoritmo	Precisión		Cobertura	
		LA1	LA2	LA1	LA2
1	Contextos obtenidos de forma manual	0.68	0.37	0.09	0.05
2	Categorías gramaticales de los bigramas	0.13	0.08	0.55	0.57
3	Palabras candidatas de acuerdo con su posición dentro de la oración	0.09	0.08	0.63	0.66

A partir de lo obtenido se optó por elegir los algoritmos que arrojaron los mejores resultados, en este caso fueron las versiones 1 y 3. Con estos dos algoritmos se continuó haciendo pruebas usando los mismos parámetros de entrada, esto con el objeto de analizar si cambiando los recursos se obtenía un resultado diferente.

5.1.3.2 Obtención del algoritmo más eficiente

Esta actividad tuvo por objetivo observar el comportamiento de los dos algoritmos con mejor rendimiento y verificar si al aplicar el mismo corpus como recursos de entrada, la misma línea tope y un mismo número de posiciones tomadas alrededor de una palabra semilla, se obtiene

un resultado similar en ambos. Se aplican los mismos recursos a ambos métodos, pero de estos se utilizan diferentes tamaños y cantidades, por ejemplo diferente tamaño de corpus, diferente cantidad de n-gramas, se aplican distintos léxicos afectivos para generar la línea tope.

Se hicieron pruebas con bigramas de una palabra semilla. Los resultados más altos se muestran en la Tabla 35.

Tabla 35 Comportamiento con mismos parámetros de entrada

Versión de algoritmo	**ID de corpus**	C1		C2	
	ID de léxico afectivo	LA1	LA2	LA1	LA2
Algoritmo 1	**Precisión**	0.12	0.07	0.12	0.08
	Cobertura	0.58	0.62	0.59	0.64
Algoritmo 3	**Precisión**	0.10	0.05	0.11	0.06
	Cobertura	0.96	0.97	0.98	0.98

De acuerdo con lo obtenido en esta actividad se prefirió a partir de este punto continuar probando con el algoritmo en su versión 3, dado que su diferencia en cobertura es mucho mayor que el Algoritmo 1, y en precisión poca su desventaja (en C1, aunque un poco mayor en C2), y del cual se obtiene una mayor cantidad de resultados que si únicamente se analizan contextos ya definidos.

5.1.4 Pruebas con algoritmo más eficiente

Esta actividad tuvo como objeto mejorar la precisión del algoritmo en su versión 3, mientras se mantenía su cobertura, se experimenta con las entradas del algoritmo.

5.1.4.1 Semillas agrupadas por frecuencia de aparición

Anteriormente, se habían utilizado las palabras con polaridad que aparecen con más frecuencia en los textos, sin embargo, para esta prueba se trabaja con semillas que se agrupan por su frecuencia de aparición, esto es, se hacen pruebas utilizando las semillas más frecuentes, las que aparecen con frecuencia media y las menos frecuentes. Se hicieron estos experimentos para saber si la frecuencia de aparición de las palabras semilla afectan de alguna forma los resultados.

Los recursos utilizados se presentan en la Tabla 27 y Tabla 28.

Las pruebas que se hicieron fueron con cantidades de 20 semillas aplicándolas a los 2 diferentes corpus y de los cuales la línea tope era extraída usando los léxicos afectivos. En la Tabla 36, se muestran los resultados obtenidos.

Tabla 36 Resultado aplicando diferente frecuencia de aparición de semillas

Frecuencia de semillas	**ID de corpus**	C1		C2	
	ID de léxico afectivo	LA1	LA2	LA1	LA2
Alta	**Precisión**	0.14	0.07	0.12	0.08
	Cobertura	0.47	0.47	0.36	0.40
Media	**Precisión**	0.15	0.07	0.12	0.07
	Cobertura	0.29	0.29	0.08	0.08
Baja	**Precisión**	0.14	0.07	0.07	0.05
	Cobertura	0.34	0.35	0.09	0.11

Se observó que al utilizar semillas que tienen una frecuencia de aparición media/baja, disminuye tanto la calidad como la cantidad de palabras candidatas, en consecuencia se eligió seguir probando únicamente con las semillas más frecuentes.

Así mismo, se hizo una prueba con las 40 semillas más frecuentes, el propósito de esto fue comparar los resultados al incrementar la cantidad de semillas al doble, en el trabajo presentado por (Turney & Littman, 2003) utilizan cantidades de 2, 4 y 14 semillas, mientras que otros métodos aplican cantidades de hasta 1187 palabras (Takamura, Inui, & Okumura, 2004). En la Tabla 37, se presentan los resultados obtenidos en este experimento.

Tabla 37 Prueba utilizando 40 semillas

ID de corpus	C1		C2	
ID de léxico afectivo	LA1	LA2	LA1	LA2
Precisión	0.14	0.07	0.13	0.08
Cobertura	0.52	0.53	0.39	0.42

Se percibe que, al incrementar la cantidad de semillas, incrementa la cobertura, mas es poca la mejora en precisión.

Por otro lado, hasta este punto, el algoritmo tenía un comportamiento el cual procesaba todos los contextos sintácticos obtenidos de las palabras semilla obteniendo así resultados duplicados, lo que se advirtió con esta prueba es que existen contextos repetidos, es decir,

hay semillas que se presentan en mismos/similares contextos y de los cuales algunos eran irrelevantes, ya que con ellos no se obtenía ninguna palabra candidata o las palabras candidatas obtenidas no se asociaban a polaridad alguna (ejemplo en la Figura 20).

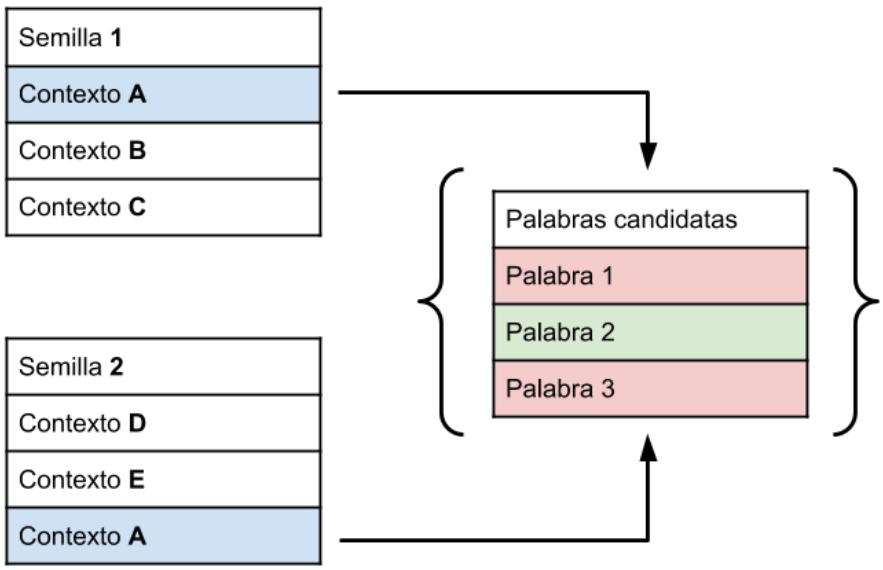

Figura 20 Ejemplo de contextos repetidos

Es por esta razón que se aplicó un filtro a los patrones repetidos buscando con esto mejorar la precisión del algoritmo. Se realizaron pruebas con cantidades de 20 y 40 semillas, en la Tabla 38, se presentan los resultados obtenidos.

Tabla 38 Prueba con filtro usando 20 y 40 semillas

Cantidad de semillas	**ID de corpus**	C1		C2	
	ID de léxico afectivo	LA1	LA2	LA1	LA2
20	**Precisión**	0.13	0.07	0.13	0.08
	Cobertura	0.28	0.28	0.36	0.39
40	**Precisión**	0.14	0.07	0.13	0.08
	Cobertura	0.47	0.47	0.36	0.39

Este experimento no arrojó los resultados que se buscaban, incluso disminuyó un poco la cobertura, lo que se pudo notar es que según la categoría gramatical de las semillas se obtenían menor o mayor cantidad de palabras candidatas, por lo tanto, se propuso experimentar ahora con semillas agrupadas de acuerdo con su categoría gramatical.

5.1.4.2 Semillas agrupadas por categoría gramatical

Este experimento se hizo con el propósito de exponer si la categoría gramatical de las semillas afectaba al resultado obtenido. Anteriormente, se habían realizado pruebas enfocadas en obtener la posición en la que aparecen las palabras con polaridad dentro de una oración según su categoría gramatical; dando como producto que, efectivamente, siendo verbos, adjetivos u otro tipo de categorías (adverbios, sustantivos, nombres comunes), suelen distribuirse hacia el final de la oración.

Teniendo en cuenta lo mencionado, para este ejercicio se utilizaron 20 palabras semilla de una categoría gramatical en específico; se agruparon en verbos, adjetivos y un último grupo conformado por las demás categorías (adverbios y sustantivos), esta agrupación se hizo, ya que las palabras candidatas y semillas etiquetadas con estas categorías, aún en conjunto, llegan a ser mucho menores en comparación de los grupos de verbos y adjetivos. Los resultados para cada conjunto se muestran en la Tabla 39.

Tabla 39 Resultados de semillas por agrupación gramatical

Agrupación por categoría gramatical	**ID de corpus**	C1		C2	
	ID de léxico afectivo	LA1	LA2	LA1	LA2
Verbos	**Precisión**	0.16	0.08	0.10	0.08
	Cobertura	0.15	0.14	0.12	0.16
Adjetivos	**Precisión**	0.14	0.07	0.15	0.09
	Cobertura	0.49	0.43	0.42	0.46
Adverbios, sustantivos	**Precisión**	0.14	0.08	0.13	0.08
	Cobertura	0.48	0.49	0.38	0.41

Así pues, los resultados obtenidos fueron similares a lo obtenido en experimentos anteriores, en este caso solo se percibe que siendo un conjunto de verbos se obtiene menor cobertura en comparación de los otros dos grupos; en consecuencia, se propuso seguir experimentando para mejorar la precisión y cobertura mediante grupos de semillas formados por diferentes categorías gramaticales.

5.1.5 Pruebas utilizando lemas de los bigramas

Anterior a estos experimentos solamente se había utilizado patrones de contexto formados por la categoría gramatical de las palabras; sin embargo, en la literatura existe un gran número de métodos que aplican modelos construidos a partir de lemas del texto. Por ello, en

esta prueba se propuso utilizar los contextos de lemas de las semillas en el algoritmo.

Los recursos utilizados son los que se presentan en la Tabla 27 y Tabla 28, estos sirvieron tanto como corpus a ser procesado como para generar la línea tope.

5.1.5.1 Semillas más frecuentes

Fueron ingresadas al método, cantidades de 20 y 40 de semillas más frecuentes que aparecen con mayor frecuencia dentro del corpus; en la Tabla 40, se observa el resultado.

Tabla 40 Experimento con 20 y 40 semillas usando lemas

Cantidad de semillas	**ID de corpus**	C1		C2	
	ID de léxico afectivo	LA1	LA2	LA1	LA2
20	**Precisión**	0.65	0.37	0.94	0.62
	Cobertura	0.13	0.14	0.13	0.12
40	**Precisión**	0.62	0.38	0.89	0.64
	Cobertura	0.16	0.18	0.16	0.18

5.1.5.2 Semillas agrupadas por categoría gramatical

Así mismo, en esta prueba se hizo lo mismo que en pruebas anteriores de este documento, es decir, utilizar semillas agrupadas de acuerdo con una categoría gramatical específica. En la Tabla 41, se reflejan los resultados al aplicar cantidades de 20 semillas de cada conjunto de categorías gramaticales.

Tabla 41 Prueba con semillas agrupadas por categoría gramatical

Agrupación por categoría gramatical	**ID de corpus**	C1		C2	
	ID de léxico afectivo	LA1	LA2	LA1	LA2
Verbos	**Precisión**	0.65	0.56	0.67	0.57
	Cobertura	0.05	0.09	0.06	0.08
Adjetivos	**Precisión**	0.64	0.37	0.93	0.62
	Cobertura	0.05	0.05	0.12	0.13
Adverbios, sustantivos.	**Precisión**	0.59	0.37	0.90	0.67
	Cobertura	0.12	0.15	0.09	0.10

Al margen de lo anterior, observamos que este método dio mejores resultados de precisión en comparación con utilizar contextos conformados por las categorías gramaticales de las palabras; no obstante, influyó al disminuir la cobertura.

5.1.5.3 Agrupación de palabras candidatas por categoría gramatical

Ahora bien, considerando estos resultados, se planteó repetir el experimento, pero ahora agrupando los resultados de acuerdo con su categoría gramatical. Es decir, ahora ya no se extrae la precisión y cobertura sobre todas las palabras candidatas devueltas por el programa, sino el valor porcentual de precisión y cobertura correspondiente para cada una de las categorías gramaticales presentes en lo obtenido por el algoritmo. Por otra parte, se propuso experimentar con el tamaño del corpus aumentado y utilizar la fusión de los léxicos afectivos, por eso, los recursos utilizados en este y los siguientes experimentos fueron los presentes en la Tabla 30.

En la Tabla 42, se identifican las cantidades de palabras candidatas, precisión, y cobertura para los grupos de categorías gramaticales devueltas, al usar 20 adjetivos como semillas. Igualmente, se muestra el resultado de probar el algoritmo ingresando 20 verbos (ver Tabla 43).

Tabla 42 Grupos de palabras candidatas por categoría gramatical usando adjetivos

20 semillas adjetivos			
Categoría gramatical	**Cantidad de palabras candidatas**	**Precisión**	**Cobertura**
Adjetivos	2697	0.60	0.40
Verbos	3690	0.32	0.22
Nombres	9894	0.19	0.23
Adverbios	984	0.08	0.07
Preposiciones	1867	0.01	0.07
Global	19132	0.25	0.25

Tabla 43 Palabras candidatas agrupadas por categoría gramatical utilizando verbos

20 semillas verbos			
Categoría gramatical	**Cantidad de palabras candidatas**	**Precisión**	**Cobertura**
Adjetivos	616	0.54	0.08
Verbos	9347	0.38	0.65
Nombres	2143	0.21	0.05
Adverbios	701	0.06	0.04
Preposiciones	1370	0.01	0.03
Global	14177	0.31	0.23

Se puede identificar que dependiendo de la categoría gramatical de las semillas que se

ingresan puede recaer su precisión o cobertura sobre esa misma categoría, se puede notar al momento de ingresar adjetivos; sin embargo, también se obtienen otras categorías que podrían estar afectando los resultados. Ahora, este trabajo propone generar un método para lograr obtener la mayoría de palabras con polaridad, sean adjetivos, verbos, o cualquier otra categoría gramatical, al margen de esto, se propuso continuar con pruebas a fin de buscar cómo mejorar la cobertura y alcanzar una mejor precisión.

5.1.5.4 Incremento de semillas

Esta prueba se enfocó en exponer el comportamiento de los patrones de una semilla, es decir cuántos y cuáles de los patrones obtenidos tienen mejor precisión y cobertura. Se utilizan para esta prueba los recursos de la Tabla 30.

En la Figura 21 se muestran los ejemplos de los contextos obtenidos por cada una de las 20 semillas; se incluyen la precisión, cobertura y medida-F de cada uno. Se muestran las leyendas repetidas ya que son varios patrones los que se extraen de las mismas.

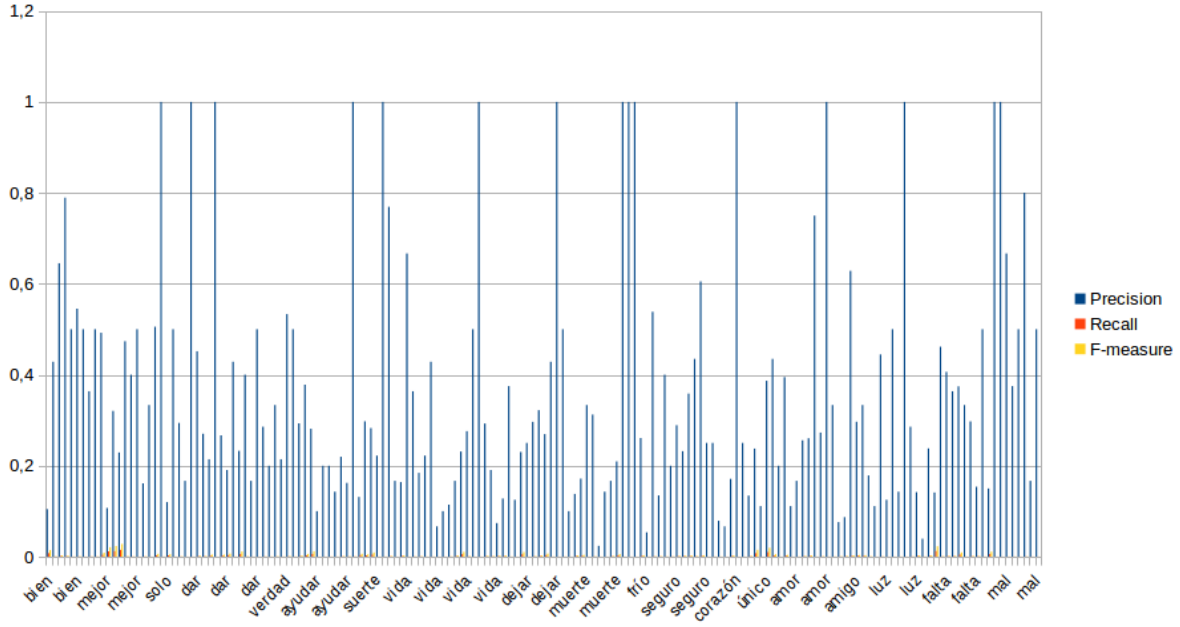

Figura 21 Comportamiento de prueba con 20 semillas

Se observa que existen patrones que tienen muy buena precisión, aunque son pocos realmente, la mayoría no van más allá de una precisión de 0.40. Posterior a esto, se realizó un experimento incrementando el número de palabras semilla utilizadas como entrada del método.

En las Figura 22, Figura 23, Figura 24 se muestran los experimentos para las cantidades de 50, 200 y 500 respectivamente.

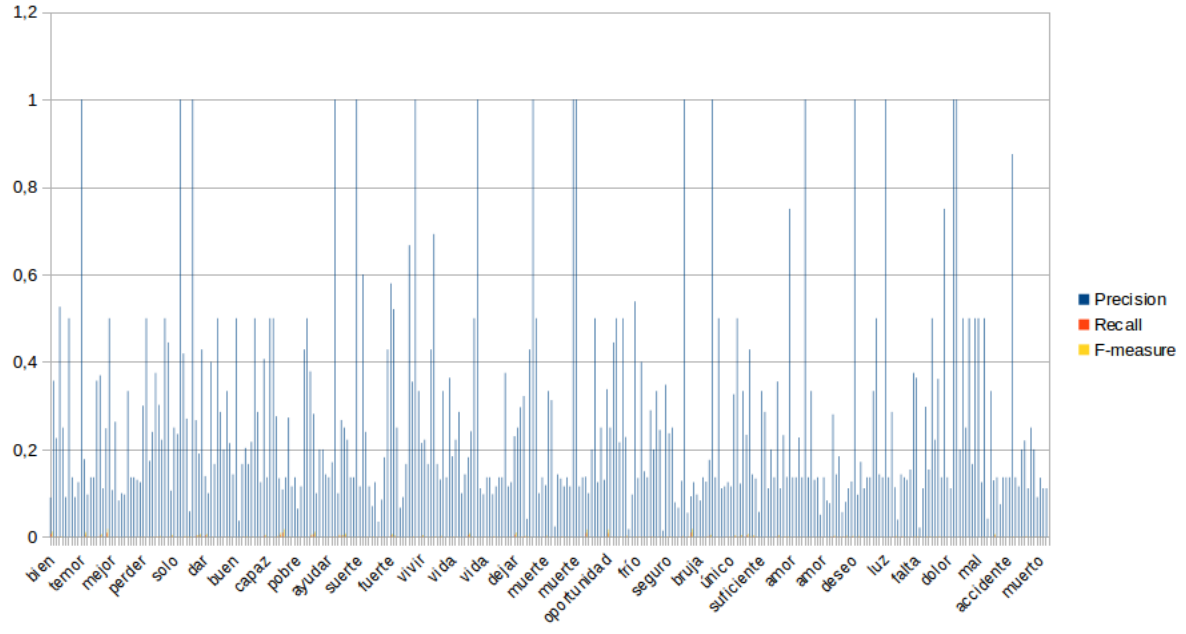

Figura 22 Muestra de comportamiento para 50 semillas

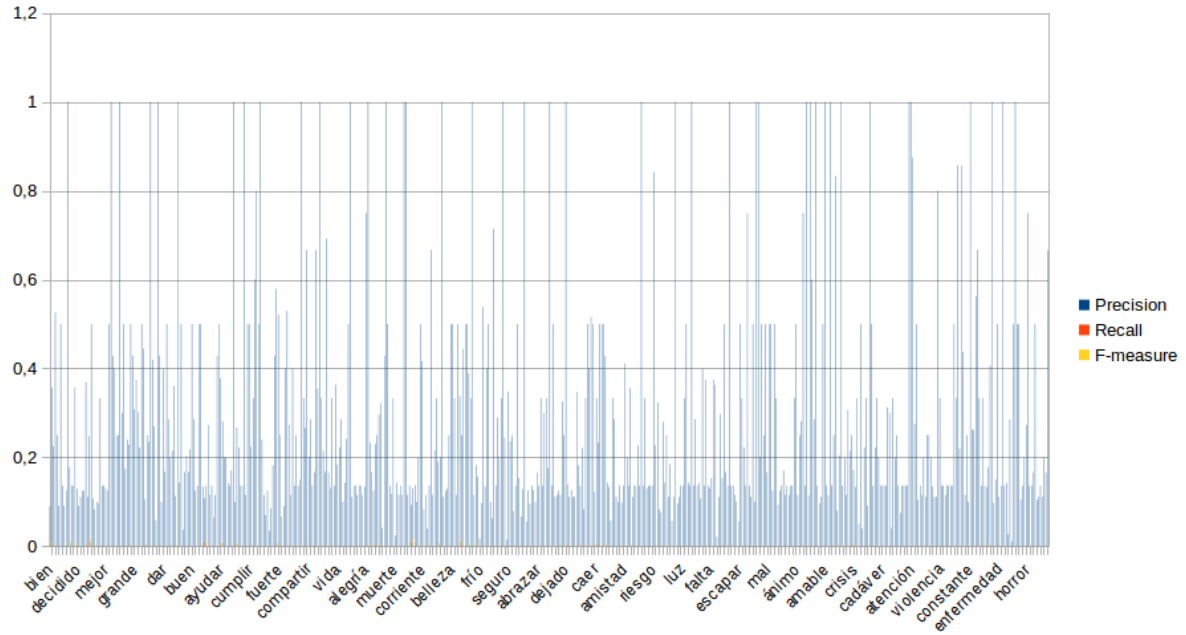

Figura 23 Experimento con 200 semillas

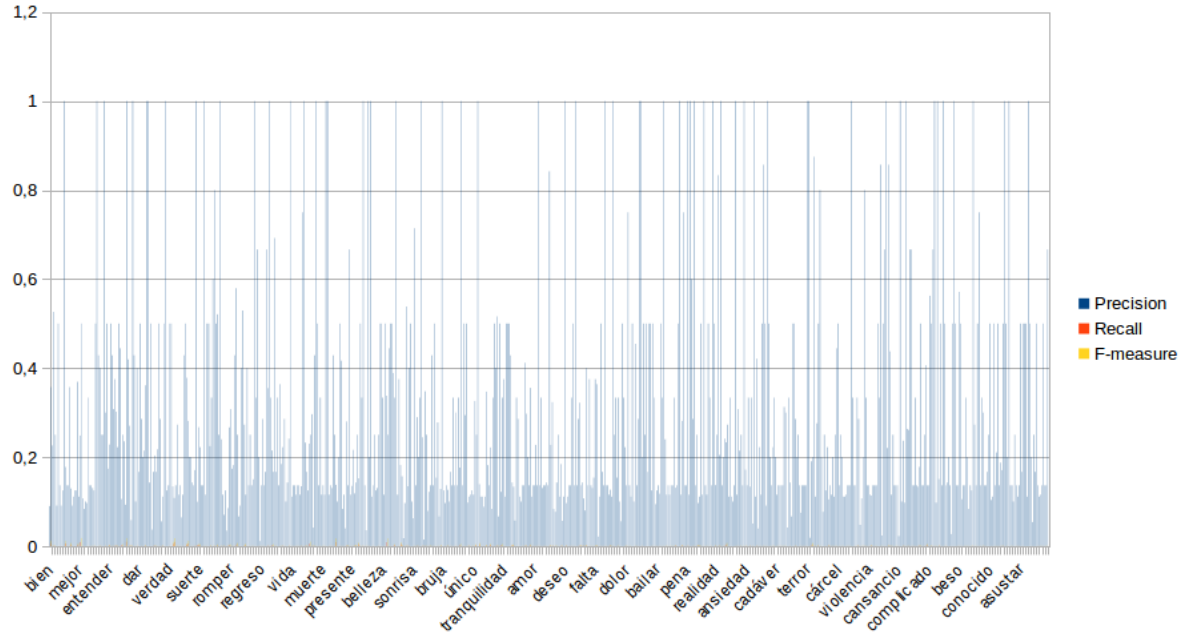

Figura 24 Comportamiento de prueba con 500 semillas

Analizando los datos obtenidos con estos experimentos se expone lo siguiente:

- No todos los patrones tienen mismas cualidades para identificar palabras candidatas de manera correcta, siendo la mayoría los que identifican palabras sin polaridad alguna.
- Entre más alta precisión tenga el patrón, se obtienen menos palabras candidatas.
- La cantidad de semillas solo incrementa la cobertura, pero patrones con menor precisión ocasionan una disminución de la precisión general del método.
- El incremento en el corpus procesado no afecta el resultado, caso contrario al disminuirlo.

5.1.5.5 Semillas por polaridad específica

En esta actividad se aborda el procesamiento de cada uno de los contextos obtenidos por cada una de las palabras semilla, esta vez para observar si al agrupar las palabras semilla de acuerdo con una polaridad en específico se obtenían mejores y/o diferentes resultados.

Se realizaron pruebas para cantidades de 10, 20 y 30 de las semillas más frecuentes, procesando por cada cantidad ambas polaridades (positiva, negativa). Como muestra de los datos arrojados por este experimento se presentan las siguientes ilustraciones.

- Figura 25 y Figura 26 para semillas positivas y negativas, respectivamente.
- Figura 27 y Figura 28 para cantidades de 20 semillas.
- Figura 29 y Figura 30 para 30 semillas, positivas y negativas.

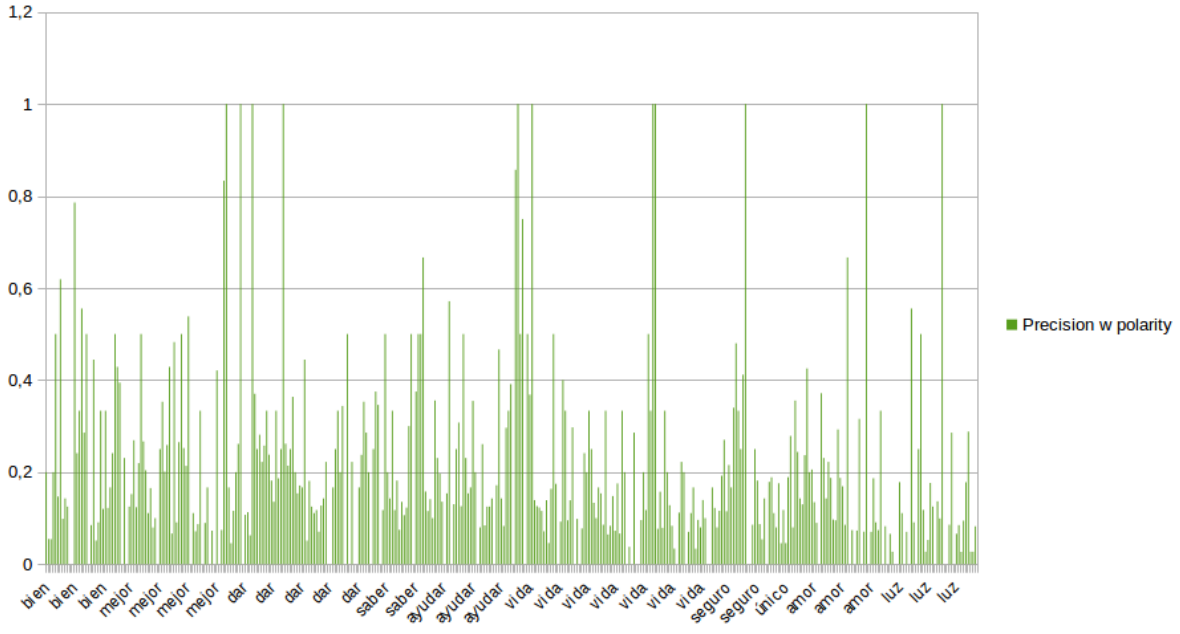

Figura 25 Utilización de 10 semillas positivas

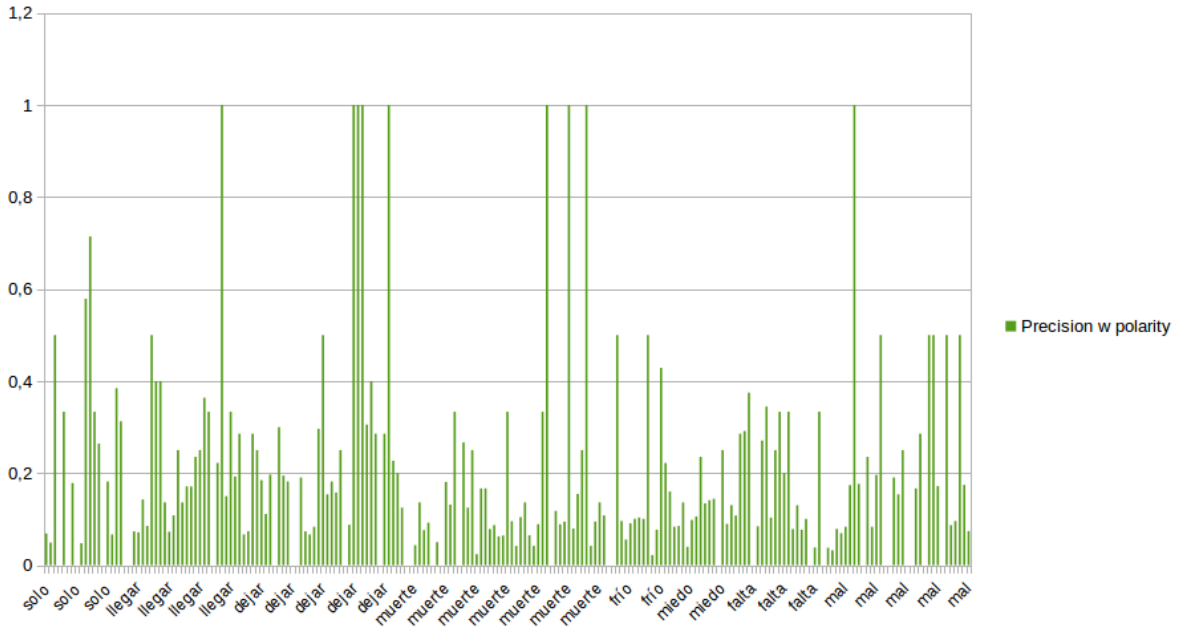

Figura 26 Uso de 10 semillas negativas

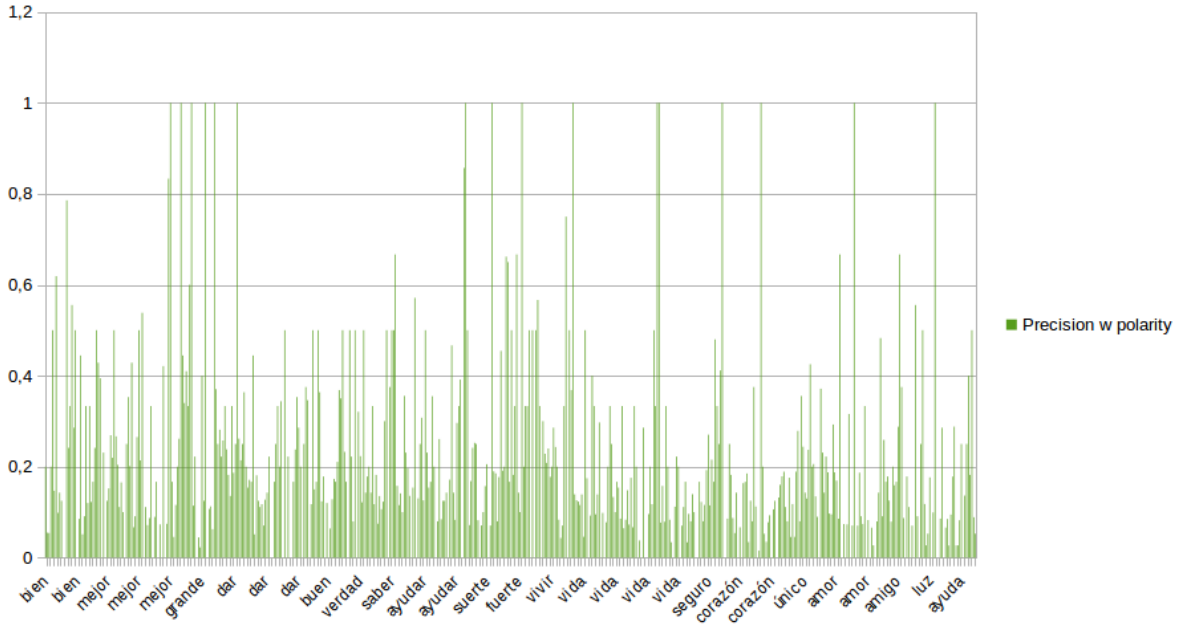

Figura 27 Uso de 20 semillas positivas

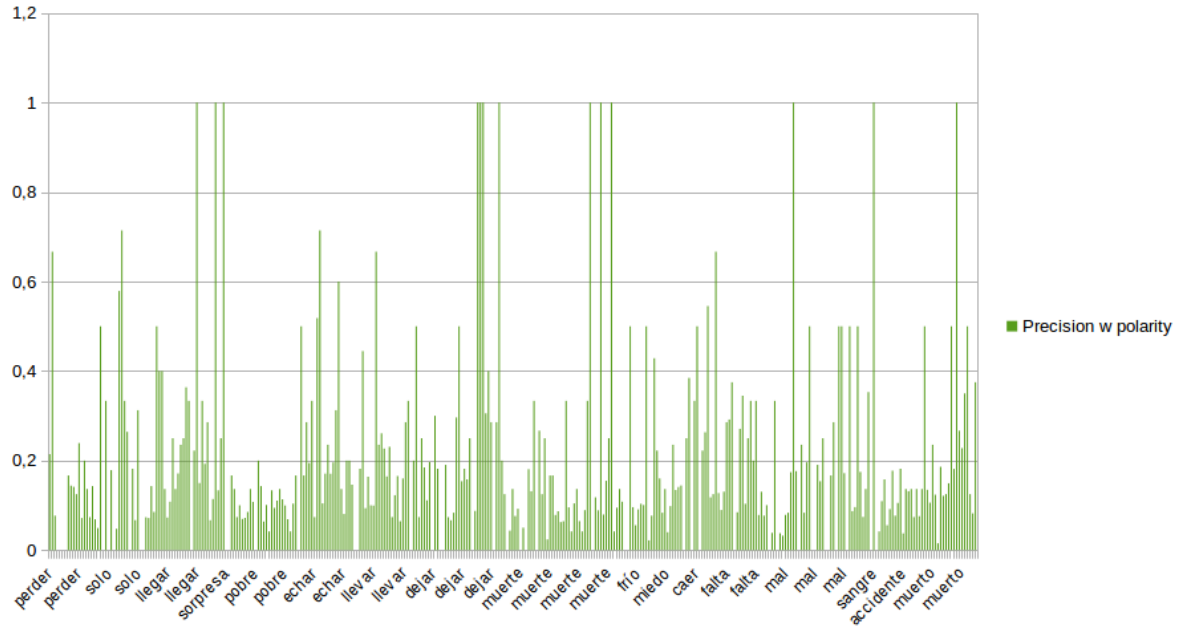

Figura 28 Experimento con 20 semillas negativas

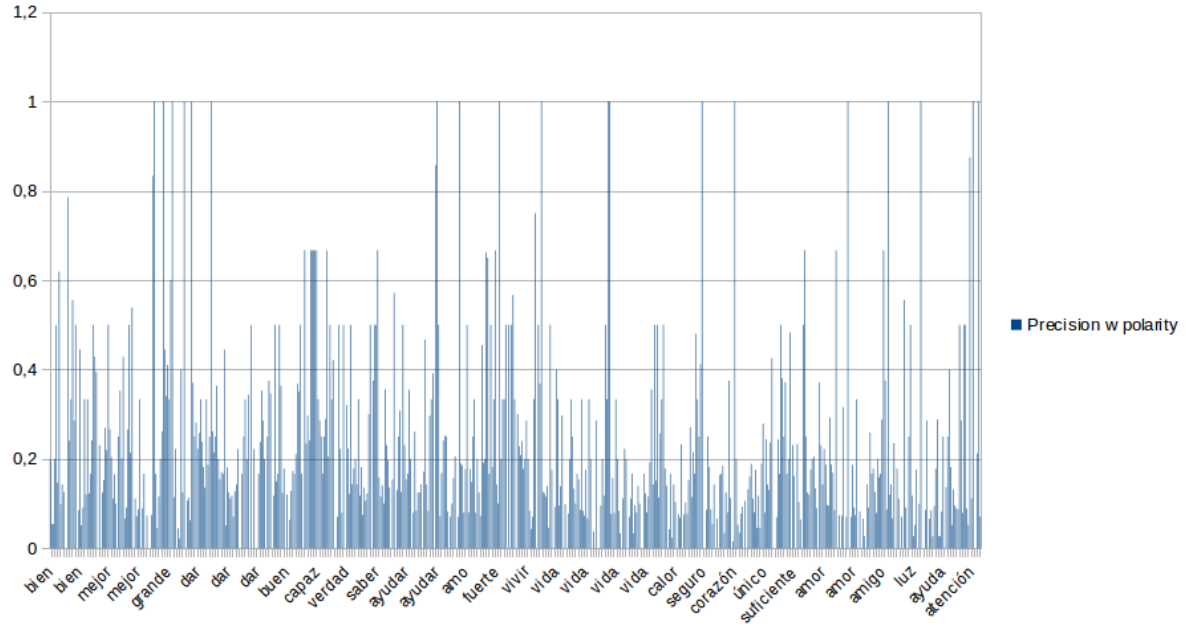

Figura 29 Prueba con 30 semillas positivas

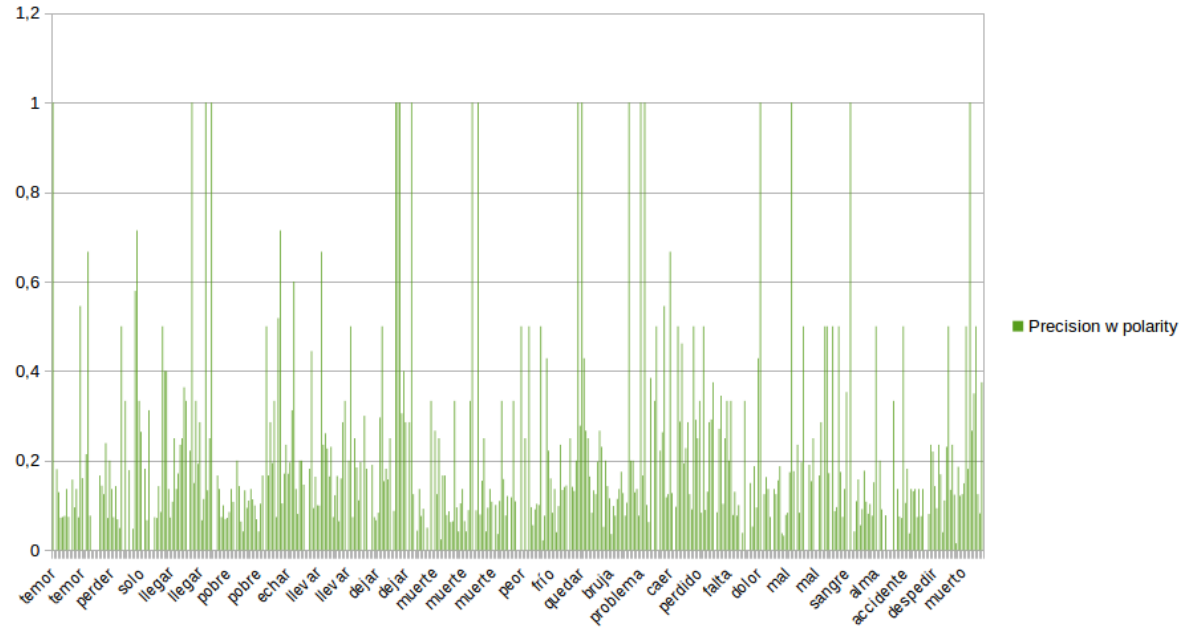

Figura 30 Prueba con 30 semillas negativas

Sobre lo que se ha expuesto anteriormente se infiere lo siguiente:

- En definitiva, existen algunos patrones que solo identifican palabras con misma polaridad que la de la semilla de la cual fueron extraídos; sin embargo, también identifican palabras con polaridad contraria.
- Al agrupar las semillas de acuerdo con una polaridad negativa, se obtiene menor precisión.

- Un conjunto de patrones extraídos de una semilla con una polaridad específica puede ser más preciso al momento de identificar palabras candidatas con misma polaridad.

5.2 Resultados

A continuación, se hace una recapitulación referente a la versión que indica los resultados más altos.

1. Los resultados del uso de patrones contextuales formados lemas en el algoritmo mostraron una gran mejora llegando a obtener una precisión de 0.94; sin embargo, en comparación con otras pruebas se disminuyó a una cobertura de 0.18, esto debido a que las palabras candidatas aparecen en patrones precisos y a veces únicos, en comparación de cuando se utiliza categorías gramaticales.
2. De los experimentos al aplicar diferentes cantidades de semillas como entrada para el algoritmo, aumentar el corpus y unir los léxicos afectivos usados, se concluye que:
   a) Entre más alta precisión tenga un patrón, se obtendrán menos palabras candidatas
   b) Aumentar la cantidad de semillas incrementa la cobertura, pero patrones con menor precisión ocasionan una disminución de la precisión general del método.
   c) Agrupar las palabras candidatas de acuerdo con la misma categoría gramatical que la semilla puede aumentar la precisión considerablemente.
3. Finalmente, al agrupar las semillas de entrada con una polaridad (negativa, positiva), derivó en lo siguiente:
   a) Existen patrones que solo identifican palabras con misma polaridad que la de la semilla de la cual fueron extraídos y de igual forma, hay modelos que identifican palabras con polaridad contraria.
   b) Agrupar semillas con polaridad negativa puede generar una menor precisión.
   c) Patrones extraídos de una semilla con una polaridad específica pueden ser más precisos al momento de identificar palabras candidatas con misma polaridad.

5.2.1 Recursos generados

Como resultado de este trabajo de investigación se generaron dos corpus y los siguientes algoritmos, también se integran los léxicos afectivos con el formato necesario para los algoritmos.

a) Corpus de cuentos y novelas.
b) Algoritmo para conteo de palabras y conteo de palabras con polaridad usando un léxico afectivo.
c) Algoritmo para identificación de palabras con polaridad.
d) Algoritmo para identificación de oraciones con palabras con polaridad.
e) Léxicos afectivos y su fusión.

Los recursos pueden obtenerse en el departamento de ciencias computacionales o biblioteca del Centro Nacional de Investigación y Desarrollo Tecnológico.

Capítulo 6

Conclusiones

Se realizaron diversas pruebas para intentar demostrar que a partir de un conjunto reducido de palabras con polaridad (semillas) es posible obtener nuevas palabras (palabras candidatas), suponiendo que los contextos que las semillas tienen en un corpus, se replican en las palabras candidatas. Las pruebas consistieron en lo siguiente:

- Generar diversas variaciones del algoritmo de extracción de palabras candidatas para seleccionar la más eficiente.

    - La primera versión del algoritmo (ver Tabla 44) alcanzó la más alta precisión, mientras que la tercera versión obtuvo la mayor cobertura, por lo que se decidió dar continuidad a los experimentos con estas dos versiones. Se utiliza una 'P' para identificar la precisión y 'C' para la cobertura.
    - Se aplicaron los mismos parámetros de entrada en ambas versiones, utilizando las mismas semillas en diferentes cantidades y el mismo corpus en diferentes tamaños. En promedio se obtuvo una baja precisión, pero, como se observa en el segundo experimento de la misma tabla, el algoritmo en su tercera versión generó una alta cobertura. Se optó por tomar la última versión, con la intensión de mejorar la precisión intentando mantener la cobertura.

Tabla 44 Resultados base para selección de algoritmo mas eficiente

Versión del algoritmo	1	2	3
Experimento1	**P: 0.68** C: 0.09	P: 0.13 C: 0.57	P: 0.09 **C: 0.66**
Experimento 2	**P: 0.37** C: 0.66	**-**	P: 0.11 **C: 0.98**

- Experimentar con los recursos utilizados como entrada.

    - Se duplicó la cantidad de las semillas a la que se había estado utilizando, tal como hacen algunos métodos presentados en el estado del arte; también, se formaron grupos de palabras semillas con base a su frecuencia de aparición dentro de los corpus (las más frecuentes, con frecuencia promedio y menos

frecuentes).

- Adicionalmente se usaron palabras semilla con una categoría gramatical específica. Eso es, aplicar como entrada semillas agrupadas por verbos, adjetivos y otro conjunto conformado por demás categorías gramaticales para observar el comportamiento de los resultados.

- Experimentar lematizando las palabras que se procesaron.

  - Se observó el comportamiento del método utilizando patrones contextuales formados por lemas.
  - El tamaño del corpus parece estar relacionado con los valores de precisión y cobertura que pueden encontrarse. En los experimentos se mostró que, al menos con pruebas en dos corpus, en el corpus de menor tamaño se obtuvo la mejor precisión de 0.94, 0.62 y cobertura 0.13, 0.12, a diferencia de los valores de 0.65, 0.37 y 0.13, 0.14 que respectivamente se obtuvieron en el corpus de mayor tamaño.
  - Se experimentó cambiando a diferentes cantidades de semillas como entrada para el algoritmo, siendo: 20, 50, 200, 500; también, con el aumento del corpus y la unión de los léxicos afectivos.
  - Por último, se probó agrupando las semillas de entrada con una polaridad (negativa, positiva).

Esta investigación demostró que es posible extraer palabras con polaridad de un corpus sin anotar utilizando patrones obtenidos a partir de un conjunto de palabras inicial (denominadas “semillas”), estos modelos pueden formarse tanto por sus categorías gramaticales como por lemas de su contexto sintáctico.

Se considera, que los mejores resultados del método en lo que se refiere a la precisión se dieron en los siguientes casos:

- Sustituyendo las categorías gramaticales por lemas para formar los patrones;
- utilizando patrones seleccionados de forma manual para ser agregados al algoritmo;
- cuando al extraer palabras candidatas se filtran únicamente aquellas que tienen misma categoría gramatical que las semillas;
- buscando las oraciones que contengan palabras con polaridad, al contrario de querer encontrar únicamente las palabras;
- agrupando las palabras semilla de acuerdo con una polaridad y se hiciera una depuración manual de las candidatas.

Por otra parte, los mejores resultados obtenidos en cobertura se presentaron cuando:

- Se incrementó el número de semillas;
- se utilizaron categorías gramaticales para formar patrones, esto debido a que varias palabras con polaridad pueden presentarse en un contexto similar, al contrario de que aparezca rodeado por mismos lemas.

La identificación de la polaridad, sea de un texto, oración o palabra en específico, es bastante compleja, varios trabajos han abordado este tema de diferente forma, enfocándose en resolver el problema aplicándolo a un dominio en concreto, a un idioma específico; además, suelen utilizar una gran cantidad de recursos para hacerlo, sean económicos, personas o tiempo, lo hagan de forma automática o en definitiva lo realizan de forma manual.

Se propuso un método independiente del dominio, que utilizando un pequeño número de palabras semillas extrajera de un corpus sin anotar palabras con polaridad. Aun cuando la precisión y cobertura no son tan altos como en otros trabajos del estado del arte, se considera que hubo éxito en identificar palabras candidatas, logrando obtener una **precisión** de hasta **0.98** y una **cobertura** de **0.18**. Por otra parte, es necesario contrastar este método con otros disponibles, haciendo énfasis en la diferencia de que este no necesita recursos tales como bases de datos léxicas, corpus anotados, léxicos afectivos e incluso, no requiere de grandes recursos computacionales. Se utilizó la herramienta Freeling (Padró & Stanilovsky, 2017) para procesar el texto; sin embargo, es posible dirigir el trabajo para que en un futuro se omita la utilización de esta.

6.1 Trabajos futuros

El examinar los resultados de los últimos experimentos, se considera como trabajo futuro que puede afectar de manera positiva el método, al llevar a cabo las siguientes actividades:

- Realizar una depuración de los patrones contextuales de acuerdo con su contenido, antes de comenzar a identificar palabras candidatas. En los experimentos se pudo observar que existen patrones que rara vez logran identificar palabras con polaridad, podría aplicarse un análisis para reconocer en qué momento y cómo se presentan dentro de los corpus.
- Experimentar con tamaños diferentes de corpus que se usarán para extraer las palabras candidatas, para determinar si la frecuencia mínima asociada al patrón está relacionada con el tamaño del corpus.
- Darle un enfoque al método para que a partir de las semillas (agrupadas por categoría gramatical) se obtengan exclusivamente palabras candidatas del mismo tipo.
- Agrupar las palabras semillas de entrada conforme a una polaridad. A partir de las últimas pruebas, se pudo observar que existen patrones contextuales que extraen más palabras candidatas con misma polaridad que la semilla. En este caso, se podría examinar en qué puntos ocurre este fenómeno y trabajar en una mejora para el

método.

- Omitir la utilización de la herramienta Freeling. En este punto puede considerarse reemplazar esta herramienta con algún método disponible en el estado del arte que ayude en el etiquetado de categorías gramaticales y lematización del texto.

Capítulo 7

Bibliografía

Akkarapatty, N., & Raj, N. S. (2016). A Machine Learning approach for classification of sentence polarity. In *2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN)* (pp. 316–321). IEEE. https://doi.org/10.1109/SPIN.2016.7566711

Alm, C. O., Roth, D., & Sproat, R. (2005). Emotions from text: machine learning for text-based emotion prediction. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05*, (October), 579–586. https://doi.org/10.3115/1220575.1220648

Aman, S., & Szpakowicz, S. (2007). Identifying Expressions of Emotion in Text. *Text, Speech and Dialogue*, *4629*, 196–205. https://doi.org/10.1007/978-3-540-74628-7_27

Baca Gómez, Y. R. (2014). *Desarrollo de un Servicio Web para Determinar la Polaridad de Textos de Redes Sociales en Español*. Centro Nacional de Investigación y Desarrollo Tecnológico.

Balahur, A., & Montoyo, A. (2008). Applying a culture dependent emotion triggers database for text valence and emotion classification. *Procesamiento Del Lenguaje Natural*, *40*, 107–114. Retrieved from http://sepln.org/revistaSEPLN/revista/40/16p21.pdf

Castellón, I., & Juarros, E. (2014). Corpus anotado. Retrieved from http://www.ub.edu/diccionarilinguistica/content/corpus-anotado

Castro-Sánchez, N. A., & López-Santiago, B. (2014). Enriquecimiento automático de un léxico afectivo basado en relaciones semánticas obtenidas de un diccionario explicativo en español. *Research in Computing Science*, *84*(2014), 113–121.

Cheng, J., & Lapata, M. (2016). Neural summarization by extracting sentences and words. *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, *1*, 484–494. https://doi.org/10.18653/v1/p16-1046

Engonopoulos, N., Lazaridou, A., Paliouras, G., & Chandrinos, K. (2011). ELS: a word-level method for entity-level sentiment analysis. *International Conference on Web Intelligence, Mining and Semantics*, 1–9. https://doi.org/10.1145/1988688.1988703

Fu, G., He, Y., Song, J., & Wang, C. (2015). Improving Chinese Sentence Polarity Classification via Opinion Paraphrasing, (October), 35–42. https://doi.org/10.3115/v1/w14-6807

García-Pablos, A., Cuadros, M., & Rigau, G. (2015). Unsupervised word polarity tagging by exploiting continuous word representations. *Procesamiento de Lenguaje Natural*, *55*, 127–134.

Geman, S., & Johnson, M. (2004). Probability and statistics in computational linguistics, a brief review. *The IMA Volumes in Mathematics and Its Applications*, *138*, 1–26.

https://doi.org/10.1007/978-1-4419-9017-4_1

Huang, A. (2008). Similarity measures for text document clustering. *Proceedings of the Sixth New Zealand*, (April), 49–56. Retrieved from http://nzcsrsc08.canterbury.ac.nz/site/proceedings/Individual_Papers/pg049_Similarity_Measures_for_Text_Document_Clustering.pdf

Jorge-botana, G., Olmos, R., & León, J. A. (2007). Análisis de la Semántica Latente (LSA) y estimación automática de las intenciones del usuario en diálogos de telefonía (call routing). *Faz Revista de Diseño de Interacción*, *1*, 53–66.

Kamp, H., & Reyle, U. (1993). *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Springer Science & Business Media.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

López-Solaz, T., Troyano, J. A., Ortega, F. J., & Enríquez, F. (2016). Una aproximación al uso de word embeddings en una tarea de similitud de textos en español. *Procesamiento de Lenguaje Natural*, *57*, 67–74.

López, T., Cruz, F., & Enríquez, F. (2016). Ampliación de lexicones de opinión específicos de dominio usando representaciones continuas de palabras. *Procesamiento Del Lenguaje Natural*, *57*(2016), 49–56.

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Mikolov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting Similarities among Languages for Machine Translation. Retrieved from http://arxiv.org/abs/1309.4168

Mohammad, S. (2011). From One Upon a Time to Happily Ever After: Tracking Emotions in Novels and Fairy Tales. *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, (October), 105–114. https://doi.org/10.1016/j.dss.2012.05.030

Moyotl-Hernández, E., & Macías-Pérez, M. (2016). Método para autocompletar consultas basado en cadenas de Markov y la ley de Zipf. *Research in Computing Science*, *115*(1), 157–170. https://doi.org/10.13053/rcs-115-1-13

Muñoz, A. M., & Álvarez, I. A. (2014). Esteganografía lingüística en lengua española basada en modelo N-gram y ley de Zipf, *190*.

Padró, L., & Stanilovsky, E. (2017). FreeLing. Retrieved from http://nlp.lsi.upc.edu/freeling/

Pang, B., & Lee, L. (2008). *Opinion Mining and Sentiment Analysis*. *Opinion mining and sentiment analysis*. Hanover, MA: now Publishers Inc.

Sharifi, M., & Cohen, W. (2008). Finding Domain Specific Polar Words for Sentiment Classification. *Language*. Retrieved from http://www.cs.cmu.edu/~mehrbod/polarity_08.pdf%5Cnhttp://www.cs.cmu.edu/%7B~%7Dmehrbod/polarity%7B_%7D08.pdf

Sidorov, G. (2013a). *Construcción no lineal de n-gramas en la lingüística computacional*.

Sociedad Mexicana de Inteligencia Artificial.

Sidorov, G. (2013b). Empirical study of machine learning based approach for opinion mining in tweets. *Advances in Artificial …*, 1–14. https://doi.org/10.1007/978-3-642-37807-2_1

Sinclair, J. (1996). Preliminary Recommendations on Corpus Typology. *EAGLES (Expert Advisory Group on Language Engineering Standards) EAG-TCWG- CTYP/P(May)*, 1–13. Retrieved from http://www.ilc.cnr.it/EAGLES/corpustyp/%5Cncorpustyp.html

Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpor, (May). https://doi.org/10.13140/2.1.2393.1847

Takamura, H., Inui, T., & Okumura, M. (2004). Extracting Emotional Polarity of Words using Spin Model. *Proceedings of the Joint Workshop of Vietnamese Society of AI SIGKBSJSAI ICSIPSJ and IEICESIGAI on Active Mining AM2004*, 1–6. Retrieved from http://www.lr.pi.titech.ac.jp/~takamura/pubs/SpinPN_AM.pdf

Tomar, D. S., & Sharma, P. (2016). A Text Polarity Analysis Using Sentiwordnet Based an Algorithm. *International Journal of Computer Science and Information Technologies*, *7*(1), 190–193.

Tribhuvan, P. P., Bhirud, S. G., & Tribhuvan, A. P. (2014). A Peer Review of Feature Based Opinion Mining and Summarization. *International Journal of Computer Science and Information Technologies (IJCSIT)*, *5*(1), 247–250.

Turney, P. D., & Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, *21*(4), 315–346. https://doi.org/10.1145/944012.944013

Vadivukarassi, M., Puviarasan, N., & Aruna, P. (2017). Identification of Opinion Words and Polarity of Reviews in Tweets using Aspect Based Opinion Mining. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *2*(5), 282–289.

Vilares Calvo, D., Alonso Pardo, M. Á., & Gómez Rodríguez, C. (2013). Clasificación de polaridad en textos con opiniones en español mediante análisis sintáctico de dependencias. *Procesamiento Del Lenguaje Natural*, *50*(2013), 13–20.

Villarejo Martínez, R. (2016). *Método para la Identificación Automática de Albures Cortos en Textos*. Centro Nacional de Investigación y Desarrollo Tecnológico.

Vincze, N., & Bestgen, Y. (2011). Identification de mots germes pour la construction d'un lexique de valence au moyen d'une procédure supervisée. In *Actes de TALN11 : Traitement automatique des langues naturelles* (Vol. 1, pp. 223–234).

Xu, G., & Huang, C.-R. (2016). Extracting Chinese polarity shifting patterns from massive text corpora. *Lingua Sinica*, *2*(1), 5. https://doi.org/10.1186/s40655-016-0014-z

Yazidi, A., Bai, A., Hammer, H., & Engelstad, P. (2015). A Simple and Efficient Algorithm for Lexicon Generation Inspired by Structural Balance Theory. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 9101, pp. 336–347). https://doi.org/10.1007/978-3-319-19066-2

Yu, H., Deng, Z.-H., & Li, S. (2013). Identifying Sentiment Words Using an Optimization-based Model without Seed Words. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 855–859. Retrieved from http://www.aclweb.org/anthology/P13-2148

Capítulo 8

Anexos

8.1 Pseudocódigo de las versiones de los algoritmos

Versión #1

Obtener palabras con polaridad más frecuentes (servirán como semillas), ***forma manual***

Identificar contextos de categorías gramaticales de las semillas, ***forma manual***

Agrupar contextos de acuerdo con frecuencia de aparición, ***forma manual***

Agregar contextos a algoritmo, ***forma manual***

Procesar corpus para obtener lemas y categorías gramaticales

Recorrer corpus

    Si gramas, bigramas o trigramas de categorías gramaticales anteriores de una palabra son idénticos a algún contexto de semilla Entonces

        Seleccionar palabra como ***candidata***

Versión #2

Seleccionar semillas, ***forma manual***

Asignar valor de para medir similitud, ***forma manual***

Asignar contexto a utilizar, sea anterior, posterior o ambos, ***forma manual***

Recorrer corpus

    Identificar contextos de categorías gramaticales de las semillas

    Recorrer corpus

        Calcular valor de similitud entre gramas, bigramas o trigramas de categorías gramaticales basado en contexto a utilizar de una palabra y los contextos de semilla

            Si valor similitud calculado >= valor similitud asignado Entonces

                Seleccionar palabra como ***candidata***

Versión #3

Seleccionar semillas, ***forma manual***

Asignar valor para similitud, ***forma manual***

Asignar contexto a utilizar, sea anterior, posterior o ambos, ***forma manual***

Asignar anotación a utilizar, sea lema o categoría gramatical, ***forma manual***

Recorrer corpus

Identificar contextos de anotación asignada de las semillas

Recorrer corpus

Identificar posición de palabra en su oración

Calcular valor de similitud entre gramas, bigramas o trigramas de anotación asignada basado en contexto a utilizar de una palabra y los contextos de semilla

Si valor similitud calculado >= valor similitud asignado Y posición en oración = (medio o final) Entonces

Seleccionar palabra como ***candidata***