INSTITUTO POLITÉCNICO NACIONAL
CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN

ANÁLISIS DE OPINIÓN EN REDES SOCIALES UTILIZADO CONCEPTOS ANTÓNIMOS EN GRAFOS

T E S I S

QUE PARA OBTENER EL GRADO DE

MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:
ING. MARCO ANTONIO MARTÍNEZ BEDOLLA

DIRECTOR DE TESIS:
DR. FRANCISCO HIRAM CALVO CASTRO

“LA TÉCNICA AL SERVICIO DE LA PATRIA”
MEXICO D.F. MAYO 2014

SIP-14 bis

INSTITUTO POLITÉCNICO NACIONAL
SECRETARÍA DE INVESTIGACIÓN Y POSGRADO

ACTA DE REVISIÓN DE TESIS

En la Ciudad de México, D.F. siendo las 16:30 horas del día 22 del mes de mayo de 2014 se reunieron los miembros de la Comisión Revisora de la Tesis, designada por el Colegio de Profesores de Estudios de Posgrado e Investigación del:

Centro de Investigación en Computación

para examinar la tesis titulada:

"Análisis de opinión en redes sociales utilizando conceptos antónimos en grafos"

Presentada por el alumno:

MARTÍNEZ	**BEDOLLA**	**MARCO ANTONIO**
Apellido paterno	Apellido materno	Nombre(s)

Con registro:

A	1	2	0	4	0	7

aspirante de: **MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN**

Después de intercambiar opiniones los miembros de la Comisión manifestaron ***APROBAR LA TESIS***, en virtud de que satisface los requisitos señalados por las disposiciones reglamentarias vigentes.

LA COMISIÓN REVISORA

Director de Tesis

Dr. Francisco Hiram Calvo Castro

Dr. Edgardo Manuel Felipe Riverón

Dr. Sergio Suárez Guerra

Dr. Grigori Sidorov

Dra. Olga Kolesnikova

Dr. Salvador Godoy Calderón

PRESIDENTE DEL COLEGIO DE PROFESORES

Dr. Luis Alfonso Villa Vargas

INSTITUTO POLITECNICO NACIONAL
CENTRO DE INVESTIGACION
EN COMPUTACION
DIRECCION

INTITUTO POLITÉCNICO NACIONAL
SECRETARÍA DE INVESTIGACIÓN Y POSGRAGO

CARTA CESIÓN DE DERECHOS

En la Ciudad de México D.F. el día 10 del mes junio del año 2014 , el que suscribe Marco Antonio Martínez Bedolla alumno del Programa de Maestría en Ciencias de la Computación con número de registro A120407 , adscrito a Centro de Investigación en Computación , manifiesta que es autor intelectual del presente trabajo de Tesis bajo la dirección de Dr. Francisco Hiram Calvo Castro y cede los derechos del trabajo intitulado Análisis de opinión en redes sociales utilizando conceptos antónimos en grafos , al Instituto Politécnico Nacional para su difusión, con fines académicos y de investigación.

Los usuarios de la información no deben reproducir el contenido textual, graficas o datos del trabajo sin el permiso expreso del autor y/o director del trabajo. Este puede ser obtenido escribiendo a la siguiente dirección marcoambedolla@gmail.com . Si el permiso se otorga, el usuario deberá dar el agradecimiento correspondiente y citar la fuente del mismo.

Marco Antonio Martínez Bedolla

Nombre y firma

Resumen

En el presente trabajo se propone un método para obtener el grado de relación o semejanza que tiene una lista de pares de conceptos antónimos, con un conjunto de opiniones en inglés acerca de un determinado tema. Dado que en el área de análisis de opinión, sólo se clasifica un determinado texto u opinión considerando su polaridad positiva o negativa, se busca poder definir o clasificar dichas opiniones con este nuevo método, en uno de dos grupos representados por los pares de conceptos antónimos propuestos por el usuario. El conjunto de opiniones debe reflejar algunos de los aspectos importantes de un determinado producto, servicio o persona.

El método que se propone está basado en WordNet, el cual es una base de datos léxica que proporciona la relación que tienen dos palabras vinculadas al estar agrupadas en conjuntos de sinónimos llamados *synsets*. Estos grupos, están interconectados entre sí, por medio de relaciones conceptuales semánticas y léxicas. De igual forma, al proporcionar WordNet diferentes medidas de semejanza entre palabras, se realizaron los mismos experimentos con tres de estas medidas y los resultados se compararon a los obtenidos con el método propuesto, demostrando así que el desempeño de estas medidas es inferior a la medida que proponemos.

El desarrollo de un sistema que permite obtener el grado de relación o semejanza que tiene un conjunto de opiniones acerca de un determinado tema con una lista de conceptos, es de gran ayuda para el estudio de tendencias, ya que además de poder decir si dichas opiniones son buenas o malas, es capaz de utilizar muchos otros conceptos que reflejen otros aspectos tales como *honradez*, *limpieza*, *aburrido*, *barato*, etc. Esta relación puede ser considerada para clasificar las opiniones tomando en cuenta el par de conceptos evaluados, asignándolas al grupo representado por el concepto que tenga mayor relación. Esto es de gran utilidad a empresas, estudios de mercado o personas famosas interesadas en saber qué piensa la gente acerca de ellos, o los productos y servicios que ofrecen.

Abstract

In this paper, we propose a new method to obtain the degree of relatedness or similarity that a list of pairs of antonyms concepts has, with a set of English opinions about a certain subject. Since in the area of sentiment analysis , only certain text or opinion is classified considering its polarity as positive or negative, it seeks to be able to define or classify those opinions, with this new method, in one of two sets represented by the pairs of antonym concepts proposed by the user. The set of opinions must reflect some important aspects of a specific product, service or person.

The used method is based on the lexical database WordNet, which provides the relatedness of two words linked when they are grouped into sets of synonyms called *synsets*. Conceptual, semantic and lexical relations interconnect these groups with each other. Similarly, by providing WordNet different measures of similarity between words, the same experiments were performed with three of these measures and the results were compared to those obtained with the proposed method, showing that the performance of these measures is far less than the one we propose.

The development of a system to obtain the degree of relatedness or similarity that has a set of opinions about a specific topic with a list of concepts, it is helpful for trend analysis, because as well as being able to say if those opinions are good or bad, it is able to use many other concepts that reflect other aspects such as honesty, cleanliness, boringness, expensiveness, etc. Relatedness can be considered for classifying the opinions according to the pair of evaluated concepts, assigning them to the set represented by the concept with the highest relatedness. This is useful for several companies, marketing research or even famous people concerned about what other people think about them, or their products and services they offer.

Agradecimientos

A mis padres, por brindarme una vida llena de cariño, por todos sus consejos, apoyo y motivación que me dieron, porque gracias a su dedicación y esfuerzo, logré llegar tan lejos. Gracias por ser mis papás.

A mi hermano Carlos Emilio, porque siempre me apoyó y motivó a seguir adelante con mis estudios. A pesar de que ya no está con nosotros, siempre estuvo conmigo. Fuiste un gran hermano.

A mi esposa Lijie, por su paciencia, cariño, confianza y apoyo durante muchos momentos difíciles. Sin ti no hubiera sido posible este gran logro.

A mi hijo André Christopher, por brindarme una de las más grandes felicidades cuando naciste. Me diste el coraje y la fuerza de seguir adelante cuando parecía muy difícil.

A mis compañeros y amigos Ángel, Eric, Edgar, Julio, Javier y Norma, por muchos momentos agradables que pasamos.

A mi asesor, el Dr. Hiram Calvo, porque siempre me apoyó y brindó muchos consejos cuando los necesité.

Al CONACYT por el apoyo económico que me brindaron durante la maestría.

Al IPN por brindarme una educación de calidad durante estos nueve años.

Índice

Índice de figuras

Índice de tablas

CAPÍTULO 1

INTRODUCCIÓN

En el presente capítulo se realiza una breve introducción acerca de las redes sociales. Se presenta la motivación así como el problema a resolver. Además, se muestra el objetivo general y los objetivos específicos propuestos, así como la descripción del contenido de cada capítulo que conforman el presente documento.

1.1 Redes sociales

El avance de la tecnología permite que las personas estamos cada vez más en constante uso de ella. Muchas de las actividades de nuestra vida cotidiana requieren de la tecnología como herramienta principal, especialmente el uso de la computadora, celular e Internet, convirtiéndolos en prácticamente una necesidad diaria. Personas de diferentes edades utilizan estas tecnologías para una gran variedad de cosas, desde simplemente el entretenimiento, hasta como fuente de información o educación. Con el paso de los años, aparecen nuevos sitios web y aplicaciones llamadas *redes sociales*, capaces de facilitarnos la comunicación con otras personas que comparten alguna relación, principalmente de amistad y mantienen intereses o actividades en común.

La primera vez que aparece el concepto de *red social* es en 1995, cuando el estadounidense Randy Conrads crea el sitio web classmates.com, en el cual pretendía que la gente pudiera recuperar o seguir en contacto con compañeros de la universidad. En los siguientes años, algunos nuevos sitios web muy básicos son creados, permitiendo a los usuarios crear perfiles, listas de amigos y enviar mensajes.

En el año 1999 Microsoft lanza MSN Messenger, el cual fue un programa sumamente popular que realizaba mensajería instantánea y que permitía una comunicación básica con otras personas a través del Internet. Con el lanzamiento de nuevas versiones, también incluía un perfil si el usuario así lo deseaba. En el año 2003, aparecen dos de los sitios de redes sociales más populares hasta ese momento: Myspace y Hi5 y en el mismo año aparece Skype, el cual es un software que permite comunicaciones de texto, voz y video. Un año después aparece Facebook y dos años más tarde, Twitter, desplazando eventualmente a la mayoría de las demás redes sociales existentes, convirtiéndolas en las dos más populares hasta el momento.

Facebook fue creado por Mark Zuckerberg para fomentar las redes sociales dentro de universidades, y posteriormente se amplió a todos los usuarios potenciales de Internet. Superó a las otras redes porque ofrecía más servicios que las demás e incluso utilizaba una aplicación para teléfonos móviles para seguir siempre conectado.

A diferencia de Facebook, Twitter es un servicio de *microblogging*, esto es, que fomenta la transmisión de información en forma breve, ya que permite enviar mensajes de texto plano de corta longitud con un máximo de 140 caracteres llamados *tweets*, que se muestran en la página principal de los usuarios, quienes pueden suscribirse a los tweets de otros usuarios para saber novedades, noticias o cosas que realizan los demás. De igual manera, Twitter se puede utilizar en teléfonos móviles.

Los servicios de redes sociales son de suma importancia hoy en día, ya que son un medio de comunicación social. La población mundial asciende a más de 7 billones de personas, de los cuales 2.5 billones son usuarios de Internet y 1.8 billones son usuarios de por lo menos una red social. En la Figura 1.1 podemos observar las redes sociales más utilizadas en todo el mundo, en donde Facebook es la que cuenta con un mayor número de usuarios, seguido por Qzone de uso exclusivo en China, y por Google+, LinkedIn, Twitter y Tumblr.

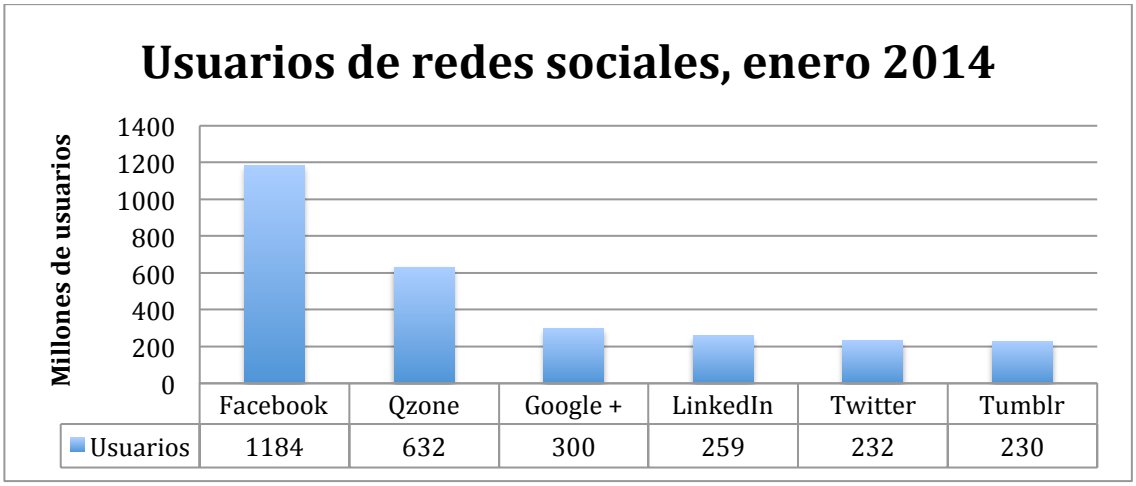

Figura 1. Usuarios de redes sociales en el mundo[1]

En los últimos años, el número de usuarios de Facebook en todo el mundo ha crecido de tal forma que es el líder con el mayor número de usuarios en la red. En la Tabla 1 podemos observar el porcentaje de usuarios de Internet de algunos países que hacen uso de redes sociales, enlistando las cuatro principales y de igual forma mostrando su respectivo porcentaje de uso.

Con los años, las empresas se han dado cuenta de que las redes sociales son un medio de comunicación en el que un gran porcentaje de personas en todo el mundo hace uso de ellas. Esto provoca que dichas empresas comiencen a utilizarlas, permitiéndoles hacer publicidad de sus productos o servicios en forma de anuncios de textos o fotos sin ningún costo alguno. Cuando a los usuarios les interesa algún tipo de publicidad, escriben toda clase de opiniones, ayudando a las empresas de manera indirecta a saber lo que piensan de sus productos.

Cabe mencionar que una opinión refleja la creencia o juicio personal del autor, resultado de una emoción o interpretación propia de un hecho. A diferencia de una opinión, un hecho es una acción ejecutada o acontecimiento ocurrido, presentado objetivamente sin creencias o juicios del autor, cuya principal característica es que es demostrable, pudiéndolo afirmar como verdadero o falso.

[1]Agencia de Marketing: http://wearesocial.es

Tabla 1. Algunos países y su porcentaje de uso de redes sociales

País	% de población total que utiliza redes sociales	Facebook	Twitter	Google+	LinkedIn
Alemania	74%	72%	32%	21%	-
Argentina	54%	90%	51%	67%	31%
Australia	73%	81%	42%	30%	23%
Brasil	48%	94%	56%	75%	54%
Canadá	82%	85%	46%	45%	30%
EUA	75%	85%	46%	44%	30%
Francia	68%	74%	24%	35%	23%
India	12%	94%	67%	78%	54%
Inglaterra	76%	79%	44%	33%	22%
Italia	54%	83%	41%	53%	24%
Japón	58%	38%	39%	17%	-
México	37%	94%	62%	74%	36%
Rusia	50%	68%	39%	52%	-
Sudáfrica	40%	92%	62%	69%	49%
Sur Corea	74%	75%	56%	38%	-

El presente trabajo se enfoca en las opiniones de la gente escritas en redes sociales, lo que facilita el proceso para realizar estudios de mercado y permite obtener tendencias o actitudes de un determinado producto o servicio que ofrece una empresa. Todo esto mediante la clasificación de opiniones utilizando aspectos que las caracterizan.

1.2 Motivación

Cuando la gente escribe opiniones acerca de un determinado producto, servicio o persona, lo hace tomando en cuenta diferentes características que lo describen, como tamaño, precio, forma de entretenimiento, velocidad de movimiento o procesamiento, etc. Todo esto lo realiza dependiendo de qué sea sobre lo que esté dando su opinión, por lo que en el trabajo que proponemos, nos gustaría ser capaces de poder representar cómo la gente expresa sus opiniones de productos, servicios o personas tomando en consideración los atributos que los describen de manera individual.

En este documento, proponemos la comparación entre diferentes conceptos organizados en pares de antónimos y textos. Con esto obtenemos más información de dichos textos, permitiéndonos cuantificar los atributos que los describen y así clasificarlo en el grupo representado por el atributo que tenga mayor relación.

1.3 Descripción del problema

Dentro del área de Análisis de Opinión, existen trabajos que clasifican distintos documentos o textos procedentes de diferentes fuentes, en donde la gran mayoría utiliza técnicas de clasificación y algún corpus previamente etiquetado. Estos trabajos sólo se enfocan en encontrar su polaridad y no en clasificar otros aspectos, además de que aquellos que utilizan opiniones dentro de redes sociales, lo hacen en Twitter y relativamente muy pocos en alguna otra. Ningún otro trabajo, hasta donde sabemos, hace uso de semejanza o distancias entre palabras para realizar una clasificación de textos.

En este trabajo, proponemos un método que, tomando opiniones obtenidas de alguna red social, no sólo entregue su polaridad, sino que se consideren otros aspectos para cuantificar con qué magnitud están relacionadas y así clasificarlas. Todo esto, mediante el uso de la base de datos léxica WordNet y sin necesidad de utilizar algún corpus.

1.4 Objetivos

1.4.1 Objetivo general

Identificar la relación de una opinión en inglés escrita en alguna red social, con cada uno de los pares de antónimos de una lista, utilizando la base de datos léxica WordNet.

1.4.2 Objetivos particulares

- Encontrar opiniones en redes sociales acerca de diferentes productos, personas o servicios, que reflejen características propias.
- Identificar algunos de los aspectos que reflejan las opiniones y determinar las listas de pares de antónimos para realizar pruebas posteriores.
- Realizar diferentes experimentos con las distintas medidas de semejanza de las palabras propuestas.
- Determinar el mejor método que refleje con certeza la relación de las opiniones con los pares de antónimos propuestos para cada tema.

1.5 Organización de la tesis

La presente tesis está organizada en seis capítulos.

En el capítulo uno, se habla de las redes sociales y por qué son de gran importancia en la actualidad. Se describe la motivación y el problema que se desea resolver en el presente documento. Se menciona el objetivo principal de la tesis así como los objetivos particulares.

En el capítulo dos se presentan algunos antecedentes acerca de la semejanza o relación entre palabras. Además se mencionan los trabajos más relevantes relacionados con este trabajo.

En el capítulo tres se presentan los conceptos, información y características de cómo está estructurada la base de datos léxica WordNet.

En el capítulo cuatro se muestra el desarrollo y características del método propuesto. Se menciona el contenido de las bases de datos creadas y se describe de manera detallada cada una de las variaciones del método propuesto.

En el capítulo cinco, se presentan los experimentos y resultados obtenidos con cada una de las variaciones del método propuesto.

En el capítulo seis, presentamos las conclusiones a las que se llegaron tomando en consideración los resultados obtenidos en el capítulo anterior. También se menciona el trabajo futuro.

CAPÍTULO 2

ANTECEDENTES Y ESTADO DEL ARTE

En este capítulo presentamos los antecedentes de la semejanza entre palabras. Se mencionan las diferencias entre semejanza y relación de palabras, así como los métodos más relevantes y que forman la base para una vasta cantidad de aplicaciones. También se hace mención de los trabajos relacionados con la semejanza de palabras, textos y clasificación de textos.

2.1 Semejanza entre palabras

Una definición para semejanza entre palabras, es que dos palabras son más semejantes que otras si comparten más características en su significado, sin necesidad de que sean sinónimos absolutos. La semejanza, igual que la sinonimia, no es una relación entre palabras sino entre conceptos a los cuales se refieren dichas palabras. A cada uno de los diversos conceptos a los que hace referencia una sola palabra, se le conoce como *sentido* de la palabra. Por ejemplo, no podemos decir que la palabra *bank* es semejante a la palabra *slope* ya que ambas pueden significar varias cosas. Si consultamos estas palabras en WordNet y obtenemos sus sentidos (*synsets*), podemos observar que el sentido uno para el sustantivo de *bank*, se refiere a la pendiente al lado de un cuerpo de agua, y el sentido uno de *slope* a una elevación geológica elevada, de esta forma, dichos conceptos expresados como sentidos de las palabras, son semejantes. Aunque la semejanza es técnicamente una propiedad entre sentidos, existen también diferentes maneras de encontrar la semejanza entre palabras, como por ejemplo tomando el máximo valor de semejanza entre los diversos sentidos de las palabras, o sumando los valores obtenidos para cada sentido, entre otros métodos.

El cómputo de la semejanza entre palabras es una tarea fundamental en el campo del Procesamiento de Lenguaje Natural ya que desempeña un papel muy importante en diversas aplicaciones que utilicen traducción automática, recuperación de información, modelado de lenguaje, generación de lenguaje natural, desambiguación de sentidos, entre otras [1][2]. El estudio de medidas de semejanza se remonta, por lo menos a la primer conferencia en lingüística computacional con el trabajo de Harper en 1965 [3]. En las aplicaciones que utilizan la semejanza entre palabras, usualmente se hace una distinción entre semejanza (*similarity*) y relación (*relatedness*). Se dice que dos palabras son semejantes si son sinónimos cercanos, lo cual es diferente a que estén relacionadas de algún modo. Por ejemplo, bicicleta y carro pueden ser semejantes pero no son sinónimos. Carro y automóvil son sinónimos mientras que carro y gasolina están claramente relacionados, ya que gasolina es algo que utilizan los carros, pero no son semejantes.

Una medida de semejanza toma como entrada dos conceptos y devuelve un valor numérico que cuantifica cuánto se parecen ambos conceptos. Esta medida usualmente está basada en relaciones "es-un" (*is-a*) dentro de la antología en la cual los conceptos residen. Generalmente se busca la semejanza, sin embargo, algunos algoritmos devuelven la relación de palabras, lo cual puede o no ser de utilidad, dependiendo de la aplicación.

Budanitsky [4] y Aguirre [5] presentan un extenso estudio y clasificación de la semejanza y relación semántica. Ellos explican las medidas desde aquellas que utilizan las estructuras

taxonómicas para tratar de cuantificar el contenido de información compartida entre dos conceptos, hasta aquellas medidas en las cuales se utiliza información contextual de los conceptos (vectores de contexto).

La mayoría de los métodos de semejanza entre palabras realizan un buen desempeño entre sinónimos, pero no así entre las palabras cuya semejanza es vaga. Existen dos tipos principales de algoritmos o métodos para realizar el cómputo de semejanza entre palabras, uno se basa en el uso de diccionarios tesauros y el segundo en algoritmos distribucionales.

2.2 Métodos basados en tesauro

En estos métodos se hace uso de un diccionario tesauro en donde se busca la distancia entre dos palabras en una jerarquía de hiperónimos o mediante el uso de sus glosas para definir la semejanza que existe entre ellas.

2.2.1 Trayectorias

El algoritmo más simple de este tipo utiliza trayectorias en una jerarquía. Se dice que dos conceptos o sentidos son semejantes si están cerca uno del otro en una jerarquía, esto es, mediante la trayectoria más corta que existe entre ellos. Un concepto tiene una longitud de distancia hacia sí mismo de uno y en cada camino hacia otro concepto se agrega uno más. En la Figura 2 se muestra una jerarquía de hiperónimos.

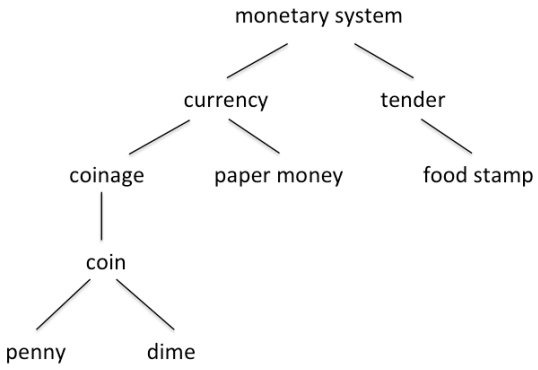

Figura 2. Ejemplo de jerarquía de WordNet

En el esquema, la palabra *penny* tiene una trayectoria de longitud uno hacia sí mismo, dos hacia *coin*, tres a *dime* o *coinage*, de cinco hasta *paper money* e incluso hasta *food stamp* tiene un camino de longitud siete. Formalmente, la longitud de la trayectoria (*pathlen*) es igual a uno más el número de aristas en la trayectoria más corta en la jerarquía entre los conceptos c_1 y c_2. Podemos transformar la longitud de una trayectoria en una distancia métrica invirtiendo la longitud del camino, obteniendo una métrica de semejanza.

$$simpath(c_1, c_2) = \frac{1}{pathlen(c_1, c_2)}$$

La formula *simpath* es una métrica de semejanza entre dos sentidos o conceptos c_1 y c_2, que también podemos transformar en una métrica entre dos palabras tomando la máxima semejanza entre todos los pares de sentidos que tienen ambas palabras w_1 y w_2.

$$wordsim(w_1, w_2) = \max_{c_1 \in sences(w_1), c_2 \in senses(w_2)} sim(c_1, c_2)$$

Existe un problema con este algoritmo, en el cual se asume que cada arista representa una distancia uniforme, sin embargo, al utilizar una jerarquía de hipónimos hay que recordar que los nodos que son más altos se vuelven muy abstractos. Por este motivo, es mejor utilizar una métrica que represente el costo de cada arista de manera independiente, ya que los conceptos conectados solamente a través de nodos abstractos son muy poco semejantes.

2.2.1.1 Método Leacock-Chodorow

La medida de Leacock y Chodorow [6] está basada en la longitud de trayectorias en una jerarquía. La semejanza entre dos conceptos c_1 y c_2 está dada por la longitud de la trayectoria más corta entre c_1 y c_2, dividida por el doble de la máxima profundidad (del nodo más profundo hasta la raíz) en la taxonomía en la que ocurren c_1 y c_2.

$$\mathrm{sim}_{\mathrm{LC}}(c_1, c_2) = -\log \frac{len(c_1, c_2)}{2D}$$

La medida Leacock-Chodorow asume un nodo superior virtual que domina todos los nodos, y para ello siempre devuelve un valor mayor que cero, siempre y cuando los dos conceptos se puedan encontrar en WordNet ya que siempre habrá una trayectoria entre ellos. Esta medida no requiere de ningún corpus, por lo que no existen problemas de datos dispersos.

2.2.1.2 Método Hirst-St-Onge

La medida Hirst-St-Onge propuesta en 1998 [7] detecta cadenas léxicas utilizando WordNet y cuantifica relaciones semánticas entre palabras de manera automática. Está basada en trayectorias extendiéndolas para todas las relaciones en WordNet (agrupándolas en horizontal, hacia arriba o hacia abajo) y penalizando los cambios de dirección. Por ejemplo, las relaciones *is-a* están hacia arriba, mientras que las relaciones *has-part* son horizontales. La medida establece la relación entre dos conceptos intentando encontrar una trayectoria entre ellos, que no sea demasiado larga ni que cambie de dirección con demasiada frecuencia.

2.2.1.3 Método Wu-Palmer

Zhibiao Wu y Martha Palmer [8] proponen otra medida de semejanza que fue originalmente utilizada en un sistema de traducción automática de verbos de inglés a chino mandarín. Esta medida al igual que la medida Leacock-Chodorow, está basada en la longitud de trayectorias. Se toma en cuenta la profundidad en la que se encuentra cada uno de los conceptos a calcular su semejanza y la de su nodo inmediato superior (LCS). La métrica esta definida por la siguiente expresión:

$$\mathrm{sim}_{\mathrm{WP}}(c_1, c_2) = \frac{2 * depth(LCS(c_1, c_2))}{depth(c_1) + depth(c_2)}$$

2.2.2 Contenido de información

2.2.2.1 Método Resnik

En el método de Resnik [9], se define un concepto P(c) como la probabilidad del concepto *c* de que una palabra seleccionada de manera aleatoria en un corpus, sea una instancia de ese concepto. Formalmente hay una variable distinta al azar, que va sobre cada palabra asociada a cada concepto en la jerarquía. Dado un concepto, cada sustantivo observado es, o bien miembro de ese concepto con una probabilidad P(c), o no es miembro, asignándole una probabilidad de 1 - P(c). Como cada palabra es un miembro del nodo raíz (*root*), el cual puede ser llamado "entidad" en WordNet, significa que la probabilidad de *root* es uno y la probabilidad de los nodos debajo de este son menores, disminuyendo conforme haya más profundidad. Para hacer uso de este método, se obtiene primero la probabilidad de c.

Dada una jerarquía de hipónimos y un corpus, para cada concepto se cuenta su frecuencia y la de sus padres hasta el concepto raíz en la jerarquía. Se define el concepto palabras de c *words(c)*, como el conjunto de palabras que tienen como padre al nodo *c* y al mismo c. Dados los datos anteriores, podemos obtener la probabilidad del concepto *c* como:

$$P(c) = \frac{\sum_{w \in words(c)} count(w)}{N}$$

Este valor nos devuelve la probabilidad del concepto *c*, por lo que la probabilidad de una palabra aleatoria, será una instancia de ese concepto. Ya que se tienen computadas las probabilidades de cada nodo, podemos asociarlas a la jerarquía. Una vez que tenemos dichas probabilidades es posible definir el contenido de información de un concepto IC(c) como el valor negativo del logaritmo de la probabilidad de ese concepto.

$$IC(c) = -\log P(c)$$

También definimos al nodo superior inmediato que incluye a c_1 y c_2 dentro de la jerarquía como LCS(c_1, c_2) de las siglas en inglés *least common subsumer*.

El método de Resnik dice que la semejanza entre dos palabras está relacionada con cuánta información tienen en común, por lo que cuanto más tienen en común, más semejantes serán. Si se tienen dos conceptos, lo que ambos tienen en común es lo que comparten heredado de su padre, esto es el nodo más bajo de la jerarquía que incluye a ambos. Para medir esto, se utiliza la siguiente formula.

$$\mathrm{sim}_{\mathrm{resnik}}(c_1,c_2) = \mathrm{IC}(\ \mathrm{LCS}(c_1,c_2)\ )$$

2.2.2.2 Método de Lin

Un método alternativa para tratar con la información teórica de semejanza es la métrica de Lin, propuesta en 1998 por Dekang Lin[10]. De igual forma que la métrica de Resnik, cuanto más tienen en común dos conceptos, más semejantes serán; sin embargo, en esta métrica también se toma en cuenta lo diferentes que son, es decir entre más diferencias tengan dos conceptos, menos semejantes serán. Para medir la cantidad de información en común dados dos conceptos c_1 y c_2, se utiliza el termino *common* indicando los puntos en común de c_1 y c_2: IC(common (c_1,c_2)). Para medir la cantidad de información que no tienen en común se utiliza toda la descripción de los conceptos c_1 y c_2 y restamos la información que tienen en común: IC(description (c_1,c_2) - IC(common (c_1,c_2))). En términos generales, c_1 y c_2 son más semejantes si IC(common (c_1,c_2)) es de un mayor valor, y IC(description (c_1,c_2) es menor, es decir entre más cosas tengan en común dos conceptos, mayor es su semejanza y menor será cuando tienen muchas más cosas que no tienen en común.

La métrica de Lin se diferencia de la de Resnik en que define las cosas que tienen en común dos conceptos como dos veces la información del nodo padre que tienen en común. Dados dos conceptos en una jerarquía, la métrica de Lin se define como dos veces el logaritmo de la probabilidad de su padre más cercano en común, entre la suma de los logaritmos de las probabilidades de cada concepto:

$$\mathrm{sim}_{\mathrm{Lin}}(c_1,c_2) = \frac{2 * IC(LCS(c_1,c_2))}{IC(c_1) + IC(c_2)}$$

2.2.2.3 Método Jiang-Conrath

El enfoque de Jiang y Conrath [11] también hace uso del concepto de contenido de información (IC). Este enfoque define la semejanza como:

$$\mathrm{sim}_{JCN}(c_1, c_2) = \frac{1}{IC(c_1) + IC(c_2) - 2 * IC(LCS(c_1, c_2))}$$

Existen dos casos especiales a los que hay que tener mucho cuidado cuando se computa la semejanza con este método; ambos involucran el caso cuando el denominador es igual a cero. El primer caso es cuando IC(c_1), IC(c_2) y IC(LCS(c_1,c_2)) son iguales a cero. En el mejor de los casos esto solamente puede pasar cuando los tres conceptos son el nodo raíz, sin embargo, cuando un concepto tiene una frecuencia de cero, se utiliza el valor de cero para IC. Cuando sucede esto, el resultado es cero por falta de información. El segundo caso sucede cuando los conceptos c_1 y c_2 son iguales. Intuitivamente esto es en el caso de máxima semejanza, en donde el resultado seria infinito. Ya que devolver esto es imposible, se encuentra la distancia más pequeña posible mayor a cero y se devuelve el inverso multiplicativo de esa distancia.

2.2.3 Glosas

2.2.3.1 Método Lesk

Otro método importante para medir la semejanza basada en el uso de un tesauro se llama *método de Lesk*, inventado por Michael Lesk en 1986 [12]. En esta métrica en lugar de usar la jerarquía de conceptos, se utilizan sus glosas. La idea de este método es que dos conceptos son semejantes si sus glosas contienen palabras semejantes o incluso las mismas palabras.

Para cada *n*-palabras que este en ambas glosas de los conceptos, el algoritmo de Lesk agrega un valor de n^2 para obtener la semejanza. Por ejemplo, si las glosas de dos conceptos tienen 3 palabras en común, entonces el valor sería $1+2^2+3^2 = 15$ siendo el valor total de la semejanza de Lesk.

Lamentablemente, el enfoque de Lesk es muy sensible a la redacción exacta de las glosas, por lo que la ausencia de una determinada palabra puede cambiar radicalmente los resultados. Además, el algoritmo determina superposiciones sólo entre las glosas de los sentidos que se están considerando. Esta es una limitación significativa en que las glosas del diccionario tienden a ser bastante cortas y no proporcionan suficiente vocabulario para relacionar las distinciones de sentido.

Existen otras versiones más avanzadas del algoritmo de Lesk, en las que no solamente se observa el contenido de las glosas de los conceptos para los cuales se desea medir su semejanza, sino que también se hace uso de una jerarquía observando a los hipónimos e hiperónimos de cada concepto y tomando en cuenta sus glosas.

Los métodos basados en el uso de tesauros tienen algunos problemas para el cálculo de semejanza. Muchas veces no se tiene un diccionario de sinónimos para un idioma en particular, e incluso cuando si se cuenta con el recurso, los tesauros presentan problemas con la precisión. Así, las palabras, las frases y las conexiones entre los sentidos pueden faltar, además de que en general los tesauros no funcionan para los verbos o adjetivos, ya que no tienen relaciones de hiponimia estructuradas como los sustantivos.

2.2.3.2 Método Vector

La medida *vector* propuesta por Patwardham y Pedersen [14], crea una matriz de ocurrencias de un corpus compuesto de glosas obtenidas de WordNet. Cada palabra de contenido utilizada en una glosa de WordNet tiene un vector de contexto asociado, y a su vez cada glosa está representada por un vector glosa el cual es el promedio de todos los vectores de contexto de las palabras encontradas en la glosa. La relación entre dos conceptos se mide encontrando el coseno entre un par de vectores glosa.

En el trabajo de Banerjee y Pedersen [15], se propone un enfoque llamado *Extended Gloss Overlap*. Es una medida de relación semántica entre conceptos que está basada en el número de palabras que comparten sus definiciones (glosas). Esta medida extiende las glosas de ciertos conceptos para incluir las glosas de otros conceptos con los cuales están relacionados de acuerdo a una jerarquía determinada. Ellos demuestran que los resultados obtenidos con la nueva medida de correlación es mejor que el criterio humano.

En otro artículo propuesto por Pedersen [16], se muestra que la calidad de la evaluación depende de la naturaleza y tamaño del corpus desde el cual los vectores de contexto fueron creados. Dentro del campo de la biomédica, los mismos autores adaptaron su método mediante el uso del corpus “notas clínicas” y el tesauro “clínica Mayo” para extraer palabras de contexto y términos de descripción respectivamente. El tesauro es una fuente de descripciones de los problemas clínicos que se han recogido de dicha clínica, el cual es el equivalente a las glosas que utiliza WordNet.

2.3 Métodos basados en semejanza distribucional

En el segundo conjunto de métodos, llamados algoritmos distribucionales, se analiza si las palabras suceden en contextos distribucionales semejantes mediante modelos de significado vector-espacio, y cuya idea viene desde los primeros trabajos de lingüística. En este modelo, se establece que si dos palabras se relacionan con palabras semejantes, entonces dichas palabras también son semejantes.

Mediante el uso de matrices término-documento, se puede saber si dos documentos son semejantes mediante el conteo de palabras con las cuales están formados. La Tabla 2 muestra un ejemplo de matriz término-documento.

Tabla 2. Representación de matriz término-documento

Palabra	Documento 1	Documento 2	Documento 3	Documento 4
batalla	1	1	8	15
soldado	2	2	12	36
tonto	37	58	1	5
payaso	6	117	0	0

Cada documento está representado por un vector de conteo, que a su vez, representa la lista de la frecuencia de las palabras que aparecen en él. Al hacer dicha matriz, se puede comparar con qué frecuencia aparecen todas y cada una de las palabras en un determinado documento. Se dice que dos documentos son semejantes si sus vectores de conteo son semejantes. Si consideramos que la Tabla 2 representa las palabras que conforman los documentos, podemos observar que los documentos 3 y 4 son semejantes, ya que las palabras batalla y soldado aparecen un número mayor de veces en comparación con las palabras tonto y payaso. De igual forma, los documentos 1 y 2 se considerarían semejantes.

2.3.1 Método de espacio de vectores de significado

Para saber si dos palabras son semejantes utilizando el mismo procedimiento que para los documentos, se consideran las palabras ahora como vectores de conteo en vez de documentos, conformado por el número de veces que aparece una palabra en todos y cada uno de los documentos tomados en cuenta. Dos palabras son semejantes si sus vectores de conteo lo son. En la Tabla 2, podemos observar que tanto las palabras batalla y soldado son semejantes, ya que ambas aparecen pocas veces en los documentos 1 y 2, pero muchas veces en los documentos 3 y 4. De igual forma, las palabras tonto y payaso, aparecen varias veces en los documentos 1 y 2, pero pocas veces en los documentos 3 y 4, entonces decimos que ambas palabras también son semejantes.

Para saber si dos palabras son semejantes mediante semejanza distribucional, se utiliza una cantidad relativamente pequeña de texto en vez de uno o varios documentos. La razón de esto es para saber el contexto en el cual ocurren las palabras que nos interesan, tomando en cuenta un determinado número de palabras con las cuales ocurren, como por ejemplo en un párrafo. Con estas palabras, se define una matriz de término-contexto, con la cual se realizan vectores de contexto.

Por ejemplo, supongamos que se tienen las palabras: piña, naranja, digital e información. Con ellas, se quiere saber si son semejantes tomando una lista de palabras que ocurren en distintos

párrafos para comparar sus contextos. La Tabla 3 muestra el ejemplo de la matriz ejemplo si suponemos que sólo se toman cuatro palabras de los párrafos para conocer su contexto.

Tabla 3. Ejemplo de matriz término-contexto

palabra	computadora	ácida	resultado	fruta
piña	0	1	0	2
naranja	0	1	0	2
digital	2	0	1	0
información	1	0	4	0

Los números muestran la frecuencia con la que aparecen las nuevas palabras junto con las que se quiere medir la semejanza. Siendo los vectores de contexto de las palabras piña y naranja además de los de digital e información semejantes, nos dice que probablemente dichos pares de palabras son semejantes.

Una vez obtenida una matriz de término-contexto, es muy común utilizar *Positive Pointwise Mutual Information*(PPMI), un método que permite decir que tanto dos palabras aparecen en el mismo contexto en vez de aparecer en forma independiente. Para ello se utiliza la siguiente fórmula que es igual al logaritmo del cociente de la probabilidad de que ambas palabras ocurran juntas, entre la probabilidad de cada palabra por separado. Se remplazan todos los valores negativos con cero.

$$\text{PPMI}(word_1, word_2) = \log_2 \frac{P(word_1, word_2)}{P(word_1)P(word_2)}$$

El problema con este método es la ponderación de palabras con poca frecuencia, ya que les asigna un valor alto, mientras que a las más frecuentes les asigna un valor menor. Sin embargo, se pueden utilizar métodos de suavizado para resolver este problema, y que afectan más a las bajas frecuencias de palabras que a las altas, lo que permite obtener valores mucho más razonables para PPMI.

Una vez obtenidos los valores de PPMI para la matriz de término-contexto, es posible obtener la semejanza mediante una variedad de métodos utilizando estos valores o incluso los mismos valores de la matriz. Dentro de estos métodos se encuentran Jaccard, Dice, Jensen-Shen o Coseno, siendo este último el método más popular para medir semejanza entre dos palabras. La siguiente formula, muestra cómo calcular la semejanza mediante el método del coseno.

$$\text{sim}_{\text{cosine}}(\vec{v}, \vec{w}) = \frac{\vec{v} \bullet \vec{w}}{|\vec{v}||\vec{w}|} = \frac{\sum_{i=1}^{N} v_i \times w_i}{\sqrt{\sum_{i=1}^{N} v_i^2} \sqrt{\sum_{i=1}^{N} w_i^2}}$$

2.4 Trabajos relacionados

En muchas de las áreas del Procesamiento de Lenguaje Natural, es muy común hacer uso de la semejanza entre palabras o conceptos para una gran variedad de aplicaciones. Algunos sistemas están enfocados solamente en calcular y obtener dicha semejanza, mientras que otros hacen uso de ella para realizar una tarea en específico tales como en el área de recuperación de información o desambiguación de sentidos. De igual forma, también existen nuevos trabajos en los que se proponen nuevos métodos para medir la semejanza o relación entre palabras, utilizando distintas fuentes de información. Muchos trabajos utilizan WordNet para obtener o proponer medidas de semejanza, como Meng y Gu[17], que en su trabajo proponen un nuevo método para medir la semejanza entre sentidos de palabra utilizando WordNet y basado en el contenido de información (IC) utilizando el método de Lin mediante la siguiente expresión:

$$\mathrm{sim}_{new}(c_1, c_2) = e^{sim_{lin}} - 1 = e^{\frac{2*IC(LCS(c_1,c_2))}{IC(c_1)+IC(c_2)}} - 1$$

Meng y Gu evalúan los resultados de su nuevo método utilizando un conjunto de datos proporcionado por Rubenstein y Goodenough[18], y comparando con los resultados de otros cinco métodos de semejanza: Wu-Palmer, Leacock-Chodorow, Resnik, Lin y Jiang-Conrath.

En [19] Qin, Yu, Lang y Wu, proponen un nuevo enfoque para medir la semejanza entre palabras dentro de una jerarquía de WordNet. Su enfoque considera no sólo la distancia entre dos palabras, sino también las características de la información obtenidas a partir de un grafo llamado DAG (Directed Acyclic Graph) el cual es creado a partir de las jerarquías de ambas palabras a analizar. La ecuación de semejanza propuesta considera la probabilidad del nodo superior inmediato, así como el contenido de información IC del mismo nodo y distintos coeficientes que son ajustables y que son obtenidos a partir del grafo. El método es evaluado utilizando treinta diferentes sustantivos y comparando los resultados con los demás métodos de semejanza y el criterio humano.

Por otro lado, Kamps, Marx, Mokken y Rijke proponen en su trabajo [20] medir la orientación semántica de adjetivos a través de tres medidas de factores que utilizan antónimos. Basados en el trabajo de Osgood de 1957, donde se menciona la teoría de la diferenciación semántica proponiendo diversos pares bipolares de adjetivos para cuantificar las respuestas de palabras, frases cortas y textos, demuestran que sólo son necesarios tres: *evaluativo* (bueno - malo), *potencia* (fuerte - débil) y *actividad* (activo - pasivo). Este trabajo es de gran relevancia, ya que Kamps et al. también utilizan el grafo de WordNet, pero solamente el que contiene adjetivos. Se toma la distancia entre los miembros de cada factor para normalizar la distancia obtenida a través de cada palabra a medir. Los resultados obtenidos están en el rango de -1 a 1, siendo evaluados mediante comparaciones con la lista construida manualmente. *General Inquirer*, es un clásico sistema de

análisis de contenido e incluye conjuntos de palabras. Ellos obtienen resultados correctos para el factor evaluativo del 68.19%, potencia del 71.36% y actividad del 61.35%.

En [21], Taieb, Aouicha y Hamadou, presentan un nuevo método que utiliza el contenido de información (IC) y características taxonómicas extraídas de una ontología para un concepto en particular. Este enfoque cuantifica el subgrafo formado por los conceptos superiores inmediatos utilizando la profundidad y el conteo de los descendientes como parámetros taxonómicos. Después, Taieb et al. integran la métrica del contenido de información a un enfoque multiestratégico con parámetros para medir la relación semántica entre palabras. Su método fue evaluado comparándolo con otros trabajos relacionados con un amplio conjunto de referencias pensadas para tareas de obtención de relación o semejanza semántica entre palabras.

Existe un paquete de software de libre acceso llamado *WordNet::Similarity* [22] basado en la estructura y contenido de WordNet e implementado como módulos de Perl. En este paquete es posible medir la semejanza semántica o la relación entre un par de conceptos (sentidos de palabra), utilizando seis medidas de semejanza y tres de relación, devolviendo un valor numérico que representa el grado en el que son semejantes o están relacionados. Las medidas de semejanza utilizan la jerarquía de hiperónimos y cuantifican cuan parecido o semejante puede ser un concepto A con respecto a un concepto B. Por ejemplo, una medida de semejanza debería de mostrar que *automobile* (automóvil) es más parecido a *boat* (barco) que a *tree* (árbol), debido a que *automobile* y *boat* comparten a *vehicle* (vehículo) como padre en la jerarquía de sustantivos de WordNet. Esta base de datos es especialmente adecuada para medidas de semejanza, ya que organiza los sustantivos y los verbos en jerarquías de relaciones *is-a*, sin embargo, a pesar de que también contiene adjetivos y adverbios, éstos no se encuentran organizados en dicha jerarquía, por lo que el uso de *WordNet::Similarity* no puede ser aplicado a estas categorías gramaticales. Tres de las seis medidas de semejanza están basadas en longitudes de trayectoria dentro de la jerarquía entre dos conceptos: *lch* (Leacock, Chodorow 1998), *wup* (Wu, Palmer 1994), y *path*. Las otras tres medidas de semejanza se basan en contenido de información (IC): *res* (Resnik 1995), *lin* (Lin 1998), y *jcn* (Jiang, Conrath 1997). Por otro lado, hay tres medidas de relación soportadas WordNet::Similarity: *hso* (Hirst, St Onge 1998), *lesk* (Banerjee, Pedersen 2003) y *vector* (Patwardhan 2003).

Muchos otros trabajos como [23][24][25][26], utilizan WordNet para calcular y proponer nuevas medidas de semejanza de sentidos, la cual es una de las fuentes léxicas más utilizadas para este propósito.

Otra base de datos utilizada para medir la semejanza entre sentidos de palabras para inglés o chino mandarín es HowNet [27]. HowNet es una base de datos en línea que, al igual que WordNet, denota las relaciones conceptuales así como propiedades que existen entre palabras chinas y su equivalente en inglés. Estos trabajos [28][29][30] utilizan las relaciones jerárquicas entre *sememes*,

unidades básicas usadas para describir cada uno de los conceptos, para definir distintas funciones de semejanza.

Otros trabajos no utilizan diccionarios, sino motores de búsqueda en línea [31] o páginas web. Tal es el caso del trabajo propuesto por Zhiquang et al. [32], en donde proponen un método para medir la semejanza semántica entre palabras utilizando la medida de cosenos y pequeños fragmentos de texto obtenidos de Wikipedia. De igual forma, en el trabajo de Milne y Witten [33] también utilizan este sitio web para medir la relación semántica entre palabras.

También existen muchos trabajos que proponen medidas de semejanza o relación entre textos. Algunos miden la semejanza entre documentos o textos de gran amplitud como en [34] o [35] que utilizan el método que se conoce como "bolsa de palabras" así como el método de vectores de significado. Otros trabajos calculan la semejanza en textos muy cortos como en el trabajo de Li et al.[36]. En él, proponen un algoritmo que toma en cuenta la información semántica y el orden de las palabras dentro de la oración. Para medir la semejanza de dos oraciones, utilizan información de WordNet y estadísticas del corpus Brown [37], para obtener la semejanza entre pequeños textos.

En el trabajo de Okazaki et al. [38], se propone una medida de semejanza entre dos oraciones, simplemente obteniendo los valores de semejanza de todos los pares de palabras que conforman ambas oraciones mediante el uso de un diccionario léxico japonés. Ellos utilizan sinónimos y otras relaciones.

En el trabajo de Tsatsaronis et al.[39], se presenta un método para medir la relación semántica entre palabras basándose en las relaciones semánticas que tienen mediante el uso de WordNet. Con esto, proponen una medida para obtener la relación semántica entre textos.

Ya que en el trabajo que proponemos, queremos poder clasificar las opiniones en uno de dos grupos considerando los pares de conceptos antónimos, es necesario ver algunos trabajos relacionados a esto. El análisis de opinión, también conocido como minería de opinión, se refiere al uso del procesamiento de lenguaje natural, análisis de texto y lingüística computacional para identificar y extraer información subjetiva de distintas fuentes. Generalmente hablando, tiene como objetivo determinar la actitud de un hablante o escritor respecto a un determinado tema, en donde su actitud puede ser su opinión, estado afectivo o sentimiento emocional expresado. En otras palabras, el análisis de opinión es el estudio computacional de opiniones, sentimientos y emociones expresadas en texto, pudiéndolo aplicar a una oración o a un documento completo que contenga información subjetiva. Su tarea más sencilla es la detección de la polaridad de un texto, mientras que más avanzadas serian dar una calificación a un texto en un determinado rango o incluso poder detectar otro tipo de actitudes. La gran cantidad de trabajos que se han llevado a cabo, están

basados en simplemente detectar la polaridad de un determinado texto, en donde puede ser desde reseñas de películas hasta opiniones en redes sociales.

En el trabajo de Pang y Lee [40], clasifican reseñas de películas mediante técnicas de aprendizaje automático, de acuerdo a su polaridad. Tomando en cuenta las palabras que conforman las reseñas y que denotan una polaridad, realizan pruebas con Naive Bayes, máxima entropía y máquinas de soporte vectorial.

Con el auge de las redes sociales, muchos trabajos utilizan las opiniones en redes sociales, sobre todo de Twiiter [41][42][43]. Todos estos trabajos consideran las características propias de Twitter y utilizan técnicas de aprendizaje automático como Tree Kernel o máquinas de soporte vectorial, para clasificar los tweets.

CAPÍTULO 3

MARCO TEÓRICO

En este capítulo presentamos algunas nociones básicas del Procesamiento del Lenguaje Natural y se realiza una descripción detallada de cómo está organizada la base de datos léxica en Inglés WordNet además de las relaciones que presenta para cada categoría gramatical.

3.1 Procesamiento de Lenguaje Natural

El lenguaje se define como el conjunto de sonidos articulados con los que el hombre manifiesta lo que piensa o siente. El Procesamiento de Lenguaje Natural (PLN) es una gama de técnicas computacionales para el análisis y la representación de textos que ocurren de forma natural en uno o más niveles del análisis lingüístico, con el propósito de lograr el procesamiento del lenguaje similar al humano para una serie de tareas o aplicaciones.

La lingüística puede aportar conocimiento lingüístico a diferentes campos. Este conocimiento dentro de un sistema de PLN, puede ser dividido en niveles como se muestra en la Tabla 4. El conocimiento lingüístico se puede organizar en componentes o niveles, ya que la estructura de cualquier lenguaje humano se puede dividir de forma natural en dichos niveles.

Tabla 4. Niveles del conocimiento en el Procesamiento de Lenguaje Natural

Nivel	**Características**	
Fonológico	Sonidos hablados	Formar morfemas
Morfológico	Composición de las palabras	Formar palabras, derivar unidades de significado
Léxico	Palabras	Diferenciar usos de las palabras
Sintáctico	Roles estructurales de palabras (o colección de palabras)	Formar oraciones
Semántico	Significado independiente del contexto	Derivar significado de oraciones
Discursivo	Roles estructurales de oraciones (o colección de oraciones)	Formar diálogos
Pragmático	Significado dependiente del contexto	Derivar significado de oraciones relativo al discurso circundante

3.1.1 Niveles de la lingüística

El método más explicativo para representar lo que realmente sucede dentro de un sistema de Procesamiento de Lenguaje Natural, es por medio del enfoque de niveles de lengua. Esto también se conoce como el modelo sincrónico del lenguaje. Investigaciones psicolingüísticas sugieren que el proceso del lenguaje es dinámico, ya que los niveles pueden interactuar su orden y no seguirlo de manera secuencial.

- *Fonológico*: Estudia cómo los sonidos (hablados) son usados en el lenguaje. Cada lenguaje tiene un alfabeto de sonidos que se pueden distinguir entre ellos llamados fonemas. En este nivel se estudia las realizaciones acústicas, por lo que sólo aparece en los sistemas de reconocimiento del habla.
- *Morfológico*: Estudia la descripción de la estructura de las palabras y sus procesos de formación. Los morfemas son las unidades más pequeñas con significado y pueden ser

combinados para formar palabras. Dado que el significado de cada morfema sigue siendo el mismo a través de las palabras, los seres humanos pueden descomponer una palabra desconocida en los morfemas que la forman con el fin de entender su significado. Del mismo modo, un sistema de PNL puede reconocer el significado transmitido por cada morfema con el fin de obtener y representar algún significado.

- *Léxico*: Se estudia el significado de cada palabra individualmente. Existen diversos tipos de procesamiento que contribuyen al entendimiento a nivel palabra. Uno de ellos, asigna una etiqueta gramatical a cada palabra. En este proceso, a las palabras que pueden funcionar como más de una categoría gramatical, se les asigna la etiqueta más probable de acuerdo al contexto en el que hayan ocurrido.
- *Sintáctico*: Se analizan las palabras en una oración con el fin de descubrir la estructura gramatical de la frase. Esto requiere una gramática y un analizador. La salida en este nivel de procesamiento es una representación de la oración que revela las relaciones de dependencia estructural entre cada palabra. La sintaxis transmite un significado en la mayoría de los idiomas, ya que el orden y las dependencias, contribuyen al significado.
- *Semántico*: Se determinan los posibles significados de una frase, centrándose en las interacciones entre los significados a nivel palabra de dicha oración. En este nivel de proceso se puede incluir la desambiguación semántica, la cual permite uno y sólo un sentido de las palabras polisémicas a ser seleccionadas e incluidas en la representación semántica de la oración.
- *Discursivo*: Estudia las propiedades del texto formado por más de una oración y transmiten un significado al hacer conexiones entre los componentes de las oraciones. Existen diversos tipos de procesamiento en este nivel, donde los dos más comunes son la resolución de anáforas y el reconocimiento de la estructura de un texto. Resolución de anáfora es el remplazo de palabras, como los pronombres que son semánticamente vacíos, con la entidad apropiada a la que se refieren. El reconocimiento de la estructura de un texto determina las funciones de sus oraciones, lo que a su vez, se suma a su representación significativa.
- *Pragmático*: Se explica la forma de cómo algún significado extra se lee dentro de textos sin llegar a ser codificado en ellos. Esto requiere mucho conocimiento del mundo, incluyendo la comprensión de interacciones, planes u objetivos. Algunas aplicaciones del PNL pueden utilizar bases de conocimiento y módulos de inferencia.

Los actuales sistemas de PNL, tienden a implementar módulos para llevar a cabo principalmente los niveles más bajos de procesamiento, ya que es posible que dichos sistemas no requieran los niveles superiores. Los niveles más bajos se han investigado y aplicado más a fondo, y manejan unidades más pequeñas de análisis en comparación con los niveles más altos.

3.1.2 Aplicaciones del Procesamiento del Lenguaje Natural

El PNL, proporciona tanto la teoría como implementaciones para un amplio rango de aplicaciones. De hecho, cualquier aplicación que utiliza texto es un candidato para PNL. Las aplicaciones que más utiliza el PNL son las siguientes:

- Recuperación de información
- Extracción de información
- Contestación de preguntas
- Creación de resúmenes
- Traducción automática
- Sistemas de diálogo

3.1.3 Breve historia del Procesamiento de Lenguaje Natural

El PNL es una de las primeras áreas de la Inteligencia Artificial. Su investigación se ha realizado durante décadas, empezando a finales de los años 40's. La Traducción Automática (TA) fue la primera aplicación basada en computadora relacionada con el lenguaje natural. Weaver y Booth [45] comenzaron uno de los primeros proyectos de TA en traducción por computadora basándose en la habilidad de descifrar códigos enemigos durante la Segunda Guerra Mundial. Surge la idea de utilizar la teoría de la criptografía e información para la traducción de diversos idiomas, iniciando la investigación en diferentes instituciones de investigación de los Estados Unidos unos pocos años después.

Los primeros trabajos de TA traducían los vocablos entre lenguas y posteriormente las colocaban en el orden apropiado de acuerdo al lenguaje destino. Los sistemas que utilizaban esta perspectiva simplemente utilizaban diccionarios de búsqueda y no tomaban en cuenta la ambigüedad léxica del lenguaje natural, produciendo malos resultados. No fue hasta 1957 cuando Noam Chomsky publica las estructuras sintácticas, introduciendo la idea de la gramática generativa y mejorando los algoritmos de TA mediante el uso de lingüísticas convencionales.

En 1962, se funda la Association for Computational Linguistics (ACL), con el fin de desarrollar diferentes técnicas, métodos y aplicaciones, estableciendo límites con otras áreas del conocimiento. Durante este periodo, otras áreas del PLN comenzaron a surgir, tales como el reconocimiento de voz. En 1965, Chomsky introduce el modelo de transformación de la competencia lingüística. Junto con el desarrollo teórico, se lograron implementar muchos sistemas prototipo para demostrar la efectividad de algunos principios particulares. En 1966, surge un proyecto llamado SMART creado por Gerard Salton [46], el cual era un generador de índices automático que utilizaba técnicas de análisis de texto en recuperación de información y varios diccionarios.

Al inicio de la década de 1970, aparecen los primeros sistemas funcionales, como SHRDLU desarrollado por Terry Winograd [47] el primero de ellos. Éste podía interpretar preguntas y órdenes sencillas, así como realizar inferencias, explicar sus acciones y aprender nuevas palabras, demostrando que era posible que una computadora entendiera una lengua natural en un dominio restringido. Durante esta década, se desarrollan muchos más proyectos como ELIZA [48], el cual era un sistema que imitaba a un psiquiatra o PARRY [49] que imitaba el comportamiento de un paciente con esquizofrenia.

Para los años 1980, se produce un cambio de tendencia, en donde se pretende modelar el lenguaje con modelos puramente estadísticos. Los primeros sistemas que aplicaban modelos probabilísticos tenían como objetivo estudiar la variación lingüística con el fin de determinar regularidades estadísticas.

En los años 1990, el campo estaba creciendo rápidamente. Esto pudo atribuirse al aumento de la disponibilidad de grandes cantidades de texto electrónico y de computadoras con mayor velocidad y memoria, además de la llegada del Internet. Los enfoques estadísticos lograron hacer frente a muchos problemas genéricos en lingüística computacional, tales como identificación de etiquetas gramaticales, desambiguación de los sentidos de las palabras, etc.

En los últimos años, las aportaciones que se han hecho han mejorado sustancialmente, a tal punto, que muchas tareas de PLN ya se consideran completamente resueltas y se han convertido en parte de nuestra vida diaria, tales como corrección de ortografía o gramática en los procesadores de texto, traducción de lenguajes de forma automática en la web, detección de *spam* en correo electrónico, detección de opiniones o incluso extracción de citas desde nuestro correo electrónico. Además, existe un claro giro hacia una parte más aplicada y comercial, lo que hace surgir las técnicas probabilísticas basadas en corpus de datos.

3.2 WordNet

La base de datos léxica en Inglés WordNet [50], es quizás una de las fuentes léxicas más importantes y ampliamente usadas para aplicaciones de PNL por su diseño y formato electrónico. WordNet es una red semántica, en la cual los significados de cada sustantivo, verbo, adjetivo y adverbio son agrupados en conjuntos de sinónimos llamados *synsets*. Los Synsets son representados en términos de sus vínculos con otras palabras a través de relaciones léxicas y semánticas-conceptuales. Cada categoría gramatical es tratada de diferente forma, ya que reflejan diferentes propiedades semánticas.

WordNet puede ser libremente consultada en Internet, además de estar disponible para su descarga. Desde el sitio web, es posible obtener el contenido de muchas palabras, incluyendo

definición, sentidos y las relaciones que tienen con otras palabras. Superficialmente WordNet es parecido a un tesauro, ya que agrupa las palabras de acuerdo a sus significados; sin embargo, existen importantes diferencias. WordNet interconecta no solamente la forma de las palabras, sino sentidos específicos, dando como resultado que las palabras que se encuentran en estrecha proximidad entre sí en la red, sean semánticamente desambiguadas. Además, WordNet etiqueta las relaciones semánticas entre palabras, en tanto que los grupos de palabras en un tesauro no siguen ningún patrón explícito que no sea el sentido de semejanza.

3.2.1 Origen de WordNet

WordNet se originó en el año 1986 en la Universidad de Princeton, en donde continúa su desarrollo y mantenimiento. George A. Miller, un psicolingüista, se inspiró en experimentos de Inteligencia Artificial que trataban de entender la memoria semántica humana [51]. Dado el hecho de que los hablantes posen conocimientos sobre decenas de miles de palabras y conceptos expresados por éstas, parecía razonable suponer un almacenamiento que fuera eficiente y económico, así como mecanismos de acceso para dichas palabras y conceptos. El modelo de Collins y Quillian proponía una estructura jerárquica de conceptos en donde los conceptos más específicos heredan la información de sus conceptos superiores (conceptos más generales); sólo la información concreta que pertenece a los conceptos específicos tiene que ser almacenada con dichos conceptos. Aunque tales teorías parecían ser confirmadas por evidencia experimental basándose en un número muy limitado de conceptos, Miller y su equipo se preguntaban si la gran cantidad de los conceptos lexicalizados de una lengua podía ser representada mediante relaciones jerárquicas en una estructura tipo red. El resultado fue WordNet, una gran red semántica construida de forma manual donde las palabras con un significado similar están relacionadas entre sí. A pesar de que WordNet ya no tiene como objetivo modelar la organización semántica humana, se ha convertido en una herramienta importante para el PNL y ayudó al origen de la investigación de la semántica léxica y la ontología.

3.2.2 Diseño y contenido

Dada la forma en que WordNet organiza las palabras e interconecta unas con otras mediante arcos etiquetados, los cuales representan relaciones de significado, WordNet tiene la forma de un grafo. Relaciones léxicas conectan cada palabra mientras que relaciones semánticas-conceptuales conectan conceptos que pueden ser expresados por más de una palabra.

WordNet se compone de cuatro diferentes partes, cada una de las cuales contiene synsets con palabras de las principales categorías gramaticales: sustantivos, verbos, adjetivos y adverbios. La Tabla 5 muestra el total de palabras y synsets que existen para cada categoría gramatical, en la versión 3.0 de WordNet.

Tabla 5. Número de palabras y synsets por categoría gramatical

Categoría Gramatical	Palabras	Synsets
Sustantivos	117,798	82,115
Verbos	11,529	13,767
Adjetivos	21,479	18,156
Adverbios	4,481	3,621
Total	155,287	117,659

Al igual que un diccionario estándar, WordNet incluye no sólo palabras individuales, sino también palabras compuestas y colocaciones, pero a diferencia del diccionario, no toma la palabra o lexema como su bloque elemental de construcción. WordNet se asemeja a un tesauro en el que sus unidades son conceptos, lexicalizados por una o más cadenas de palabras o formas de palabra. Un grupo de palabras que pueden referirse al mismo concepto (sentido) se denomina un conjunto de sinónimos *synset*. Un usuario de WordNet puede encontrar el significado de una palabra específica, no sólo en términos de los miembros que se encuentran en el mismo synset, sino también a través de sus relaciones con otras palabras, es decir, tomando en cuenta su ubicación dentro del grafo. Por ejemplo, el significado de *dog* está dado en parte por todos los términos que refieren a tipos de perro, como *corgi*, *poodle* y *dalmatian*. Los significados de *corgi*, *poodle* y *dalmatian* a su vez, se dan en parte en términos de su concepto de orden superior: *dog*. En la Figura 3, se puede observar esta jerarquía de las palabras que refieren a tipos de perro y *dog*, donde cada una de las palabras agrupadas, se refiere al mismo concepto.

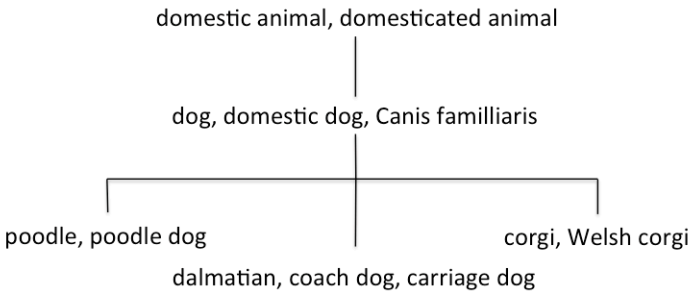

Figura 3. Jerarquía para tipos de perro

La relación principal entre las palabras de WordNet es la sinonimia, como entre las palabras *shut* y *close* o *car* y *automobile*. Las palabras sinónimas que denotan el mismo concepto y son intercambiables en muchos contextos, son agrupadas en conjuntos no ordenados (synsets). La substitución de un miembro de un synset por otro no cambia el valor de verdad del contexto; sin embargo, un sinónimo puede ser estilísticamente mejor que otro en algunos contextos. La sinonimia es la relación de muchos a uno entre las formas de las palabras y los conceptos. Por ejemplo, las palabras *boot* o *trunk* pueden referirse al mismo concepto: el compartimiento para equipaje de un automóvil.

Cada uno de los synsets de WordNet, están vinculados a otros synsets por medio de un pequeño número de relaciones conceptuales. Adicionalmente, un synset contiene una breve definición (glosa) y en la mayoría de los casos, una o más oraciones ilustrando el uso de las palabras miembro de dicho synset. Las palabras con un número determinado se significados, son representadas en el mismo número de synsets. Así, cada par de palabra con su significado en WordNet, es único. La Tabla 6 muestra el número total de pares que existen para cada categoría gramatical.

Tabla 6. Cantidad de pares palabra-sentido para cada categoría gramatical

Categoría gramatical	Par palabra-significado
Sustantivos	146,312
Verbos	25,047
Adjetivos	30,002
Adverbios	5,580
Total	206,941

La polisemia es la relación de muchos a uno, de los significados a las palabras. Así, la palabra *trunk* puede referirse a una parte de un carro, árbol, torso o trompa de elefante. En WordNet, la pertenencia de una palabra a varios synsets refleja la polisemia de dicha palabra o multiplicidad de significados. Por lo tanto, *trunk* aparece en varios synsets diferentes, cada uno con sus propias definiciones. Del mismo modo, por ejemplo la palabra polisémica *boot* aparece en varios synsets, en uno junto con *trunk*, en otro junto con *iron boot* y *iron heel*, etc. En la Figura 4 se muestran los cuatro synsets en los que aparece la palabra *trunk*, así como las palabras que comparten el sentido y la respectiva definición para cada uno.

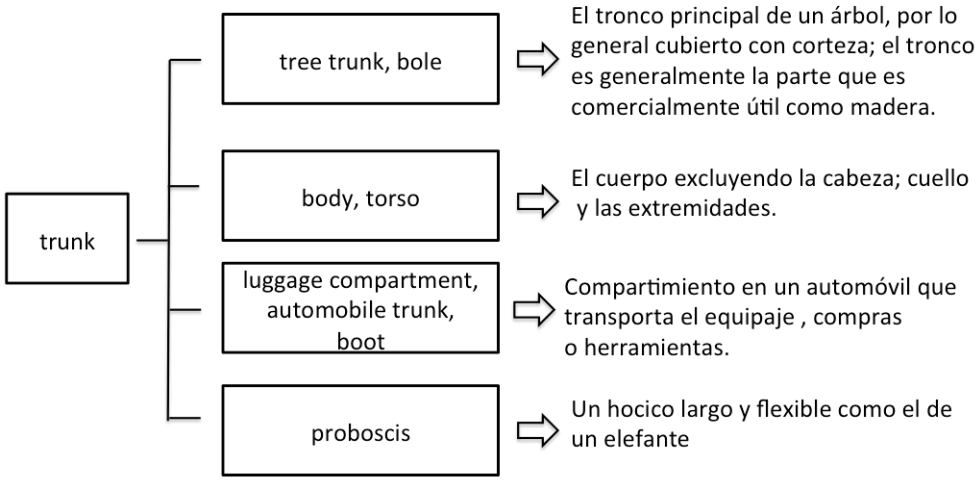

Figura 4. Synsets de la palabra trunk

Los synsets son los nodos o bloques de construcción de WordNet. Como resultado de la interconexión de los synsets a través de las relaciones basadas en significado, surge una estructura de red.

3.2.3 Relaciones

Además de la sinonimia, WordNet tiene otra relación léxica de palabra a palabra: antonimia o más general, el contraste semántico u oposición. La antonimia es particular entre pares de adjetivos como seco–húmedo o largo–corto, pero también está codificada para algunos pares de verbos y sustantivos como subir–bajar o ir–regresar o líder–seguidor.

Otro tipo de relación léxica, llamada morfosemántica, es la única que une palabras de las cuatro categorías gramaticales. Conecta palabras que están morfológicamente y semánticamente relacionadas, como por ejemplo los sentidos relacionados semánticamente de las palabras: interrogatorio, interrogador, interrogar e interrogativa están conectados.

Todas las demás relaciones en WordNet son relaciones semántico-conceptuales y conectan no sólo palabras individuales, sino synsets completos. Para cada etiqueta gramatical, podemos encontrar diferentes relaciones.

3.2.4 Sustantivos en WordNet

Los sustantivos constituyen la mayoría del léxico en inglés. Los sustantivos son relativamente fáciles de organizar en una red semántica.

3.2.4.1 Hiponimia

Los synsets de sustantivos están interconectados principalmente por la relación hiponimia o hiperonimia, la cual vincula conceptos específicos a los conceptos más generales o viceversa. Esta relación construye jerarquías con conceptos cada vez más específicos, en donde el más general es un concepto raíz. Las jerarquías de sustantivos pueden ser profundas y comprender hasta quince niveles, particularmente las categorías biológicas, donde WordNet incluye tanto términos expertos como populares.

Todos los synsets de sustantivos descienden de una sola raíz: root. El siguiente nivel comprende tres synsets: *physical entity*, *abstract entity*, y *thing*. Debajo de éste, se encuentra el tercer nivel, en el que encontramos los synsets: *object, living thing, causal agent, matter, physical process, substance, psychological feature, attribute, group, relation, communication, measure, quantity, amount*, y *otherworld*. La selección de esta gran variedad de categorías fue de alguna forma subjetiva y ha generado diversas discusiones. A un nivel empírico, queda por ver si las versiones para otros idiomas de WordNet, contienen las mismas distinciones fundamentales.

En la Figura 5 podemos ver un ejemplo de una jerarquía de hiponimia. Por ejemplo, el synset que contiene las palabras *gym shoe, sneaker* y *tennis shoe* es un hipónimo o subordinado de *shoe*, el

cual a su vez es hipónimo de *footwear, covering*, etc. De igual forma, *footwear y footgear* son hiperónimos de *shoe* y *boot,* y que a su vez son hiperónimos de *gym shoe, sneaker, tennis shoe, cowboy boot y hip boot*. La relación de hiponimia es bidireccional y transitiva.

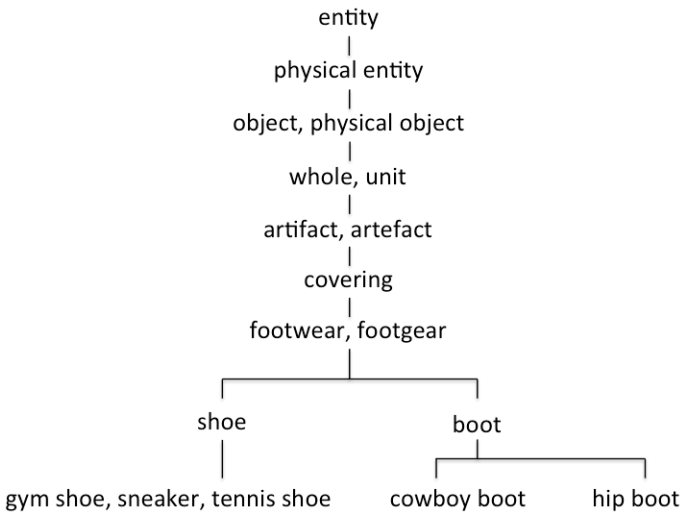

Figura 5. Ejemplo de jerarquía de hiponimia

3.2.4.2 Meronimia

Otra relación importante entre synsets de sustantivos es la meronimia ("parte-todo"). Esta relación vincula synsets mediante partes, componentes o miembros de otros synsets que denotan el conjunto. Así, por ejemplo *toe* es un merónimo de *foot*, y que a su vez, es merónimo de *leg* y así sucesivamente. Como la hiponimia, la meronimia es bidireccional. WordNet nos dice que *foot* tiene *toe* y que *toe* es una parte de *foot*. Los hipónimos heredan la relación meronimia hacia sus subordinados: Si *car* tiene *wheels*, entonces muchos tipos de *cars* tienen *wheels*. La meronimia en WordNet en realidad abarca tres relaciones "parte-todo" semánticamente distintas. Una se refiere a partes propias o componentes, como *feather* y *wing*, las cuales son partes de *bird*. Otra vincula substancias que constituyen otras substancias: *oxygen* es una parte que constituye a *water* y *air*. Y por último, miembros como *tree* y *student* son partes de grupos como *forest* y *class* respectivamente.

3.2.5 Verbos en WordNet

Los verbos son fundamentalmente diferentes a los sustantivos en que ellos se refieren a eventos y estados que involucran participantes (sustantivos), además de que los verbos tienen extensiones temporales. Los synsets de verbos están organizados por varias relaciones de vinculación léxica. La más frecuente relación codificada es la "troponimia" (troponymy), la cual relaciona dos verbos tal que uno de ellos especifica una determinada forma de llevar a cabo la acción mencionada por el

otro. Por ejemplo, *mumble* es un tropónimo de *talk* y *scribble* es un tropónimo de *write*. Para los verbos que denotan eventos, la troponimia similarmente relaciona de un concepto más general a uno semánticamente más elaborado.

Como la hiponimia, la troponimia construye jerarquías con diversos niveles de especificidad, pero las jerarquías de verbos son más superficiales que las jerarquías de sustantivos y raramente exceden los cuatro niveles. La troponimia es una relación muy polisémica cuya semántica son dominios dependientes, es decir se toman en cuenta una variedad de formas, dependiendo del campo semántico. Los verbos de movimiento son semánticamente elaborados a lo largo de dimensiones tales como velocidad (*walk, run*), dirección (*go up*, *go down*), y medios de desplazamiento (*walk, drive*). Verbos de comunicación como hablar, tienen tropónimos que especifican el volumen (*whisper, murmur*, *shout*, *yell*). No se hace distinción en WordNet entre diferentes tipos de troponimia. Algunas relaciones entre los verbos incluyen oposición semántica, la cual como la troponimia es polisémica; verbos de movimiento a menudo que forman pares de opuestos en función de la dirección del movimiento (*ascend - descend, come-go*).

La troponimia es un tipo de vinculación, que mantiene sólo una dirección. Esta relación unilateral es similar a algunos tipos de meronimia. Algunos sustantivos se refieren a grupos, colecciones, o sustancias que existen sólo en virtud de sus partes, miembros o ingredientes. Al mismo tiempo, las entidades que constituyen las partes, miembros o ingredientes pueden existir fuera de estos grupos, colecciones o sustancias. Por ejemplo, una biblioteca no es una biblioteca a menos que contenga libros como su parte más importante. Pero un libro no se define necesariamente como parte de una biblioteca. Del mismo modo, un bosque debe contener árboles, pero un árbol no es necesariamente una parte de un bosque.

Otra relación entre verbos, es la vinculación hacia atrás, donde el evento codificado en uno de los verbos implica necesariamente que haya ocurrido un evento antes de él, expresado por el segundo verbo. Esta relación en WordNet se llama *entailment*. Por ejemplo los verbos: *divorce* y *marry*, *untie* y *tie, walk* y *step.* Estos tipos de vinculación difieren entre sí con respecto a las relaciones temporales entre las dos actividades denotadas por los verbos.

3.2.6 Adjetivos en WordNet

A diferencia de los sustantivos y los verbos, los adjetivos no se prestan a una organización jerárquica. En cambio, caen en grupos centrados en torno a dos adjetivos antónimos. Antonimia es la relación que prevalece entre los adjetivos. La mayoría de los adjetivos se organizan en pares de antónimos "directos", así como *dry* y *wet* o *large* y *short*. Cada miembro de un par antónimos directos se asocia con un número de adjetivos "semánticamente similares", ya sean sinónimos cercanos o de diferentes valores de una propiedad escalar dada. Por ejemplo, *damp* y *drenched* son

semánticamente similares a *wet*, mientras *arid* y *parched* son similares a *dry*. Estos adjetivos semánticamente similares se dice que son antónimos "indirectos" del antónimo directo de sus miembros centrales, es decir, *drenched* es un antónimo indirecto de *dry* y *arid* es un antónimo indirecto de *wet*.

También WordNet contiene adjetivos "relacionales", que son morfológicamente derivados y vinculados a los sustantivos a los que pertenecen. Adjetivos relacionales, también son menos polisémicos que los adjetivos centrales como *big* o *old*. A veces pueden ser reemplazados por el sustantivo del que se derivan. Por lo tanto, un sustantivo modificado por un adjetivo relacional de alguna forma se asemeja a un sustantivo compuesto. La inclusión de esta clase de adjetivos en una base de datos que se destina en gran parte para la desambiguación de sentidos y recuperación de información podría ser rentable.

3.2.7 Adverbios en WordNet

La mayoría de los adverbios se derivan a través del sufijo *-ly* de los adjetivos a los que están semánticamente relacionados. Siempre que sea posible, los adverbios están vinculados a adverbios antónimos, siguiendo las relaciones de los adjetivos de los que derivan. Para los adverbios léxicos como *hard* o *even* entre otros, ninguna relación particular se ha implementado.

CAPÍTULO 4

METODOLOGÍA

En este capítulo se describe detalladamente el proceso y cada una de las etapas de la metodología propuesta para obtener el grado de relación que tiene un conjunto de opiniones con una lista de conceptos.

4.1 Descripción del método propuesto

En el sistema que se ha implementado, proponemos un método para determinar la relación que existe en distintas opiniones escritas en redes sociales, con un conjunto de pares de conceptos antónimos definidos por el usuario. A través del desarrollo de este sistema, los pasos que se llevaron a cabo así como el procedimiento para determinar la relación de las opiniones con los conceptos, fueron cambiando gradualmente a lo largo de cinco formas diferentes de obtener dicho resultado. Al momento de comparar los resultados con cada una de ellas, la quinta variación arroja los mejores resultados.

Dada la forma del método así como los resultados obtenidos, es posible afirmar que este método sirve para clasificar textos en dos posibles grupos conformados por los conceptos a medir su relación. La clasificación de textos es la tarea de asignar cualquier tipo de categoría a un texto. La mayoría de los métodos de clasificación son métodos supervisados ya que hacen uso de algún corpus. Los algoritmos más utilizados son algoritmos de aprendizaje automático como máquinas de soporte vectorial (SVM), regresión logística o Naive Bayes. El método propuesto es considerado como método no supervisado, ya que no hace uso de ningún corpus etiquetado o fuente de información externa aparte de WordNet que contenga características propias que ayuden o intervengan en el proceso de asignar algún valor de relación entre los textos y los conceptos.

La base de desarrollo del sistema fue el uso de WordNet 3.1 [50] y la relación de sinonimia entre conceptos. Para definir una medida de distancia se utilizaron nociones básicas de teoría de grafos, tomando en cuenta la trayectoria más corta entre dos nodos. Se recolectaron todos los conceptos de WordNet, relacionando todos aquellos que son sinónimos (conceptos agrupados en los mismos synsets) para formar un grafo principal. Esta medida de distancia está basada bajo la hipótesis de que la relación de sinonimia entre dos conceptos es más fuerte y por lo tanto son más semejantes, si la distancia de su trayectoria que existe entre ellos dentro del grafo es menor, mientras que si su trayectoria es mayor, la relación es menor y de igual forma su semejanza. Por ejemplo, utilizando el grafo principal de WordNet, la distancia de la trayectoria del concepto *angry* a *good* es 6, mientras que de *angry* a *bad* es 3. Por lo que *angry* se considera malo en vez de bueno, ya que su distancia es menor con *bad*. Esto siguiere que es posible utilizar la distancia entre dos conceptos para medir la relación entre ellos, a pesar de no ser una escala precisa.

El sistema implementado está conformado por tres partes principales, en donde en cada una de ellas se realizan tareas más específicas. El lenguaje utilizado para el desarrollo del programa principal fue Perl; también se desarrollaron pequeños programas en el lenguaje de programación Phyton para utilizar paquetes ya desarrollados e implementar su uso. La Figura 6 muestra las etapas principales del proceso que realiza el sistema.

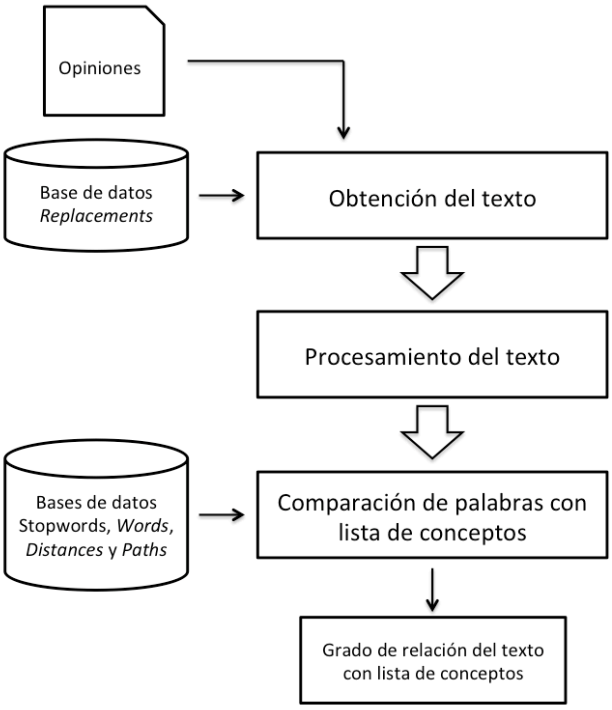

Figura 6. Diagrama de metodología

Dentro de la etapa de obtención del texto, se dividen las opiniones en oraciones; se buscan errores de escritura, acrónimos, abreviaciones y emoticones para ser remplazados; se realiza un etiquetado gramatical y se buscan negaciones. En la etapa de Procesamiento del Texto se remueven palabras sin ningún significado gramatical (*stopwords*) y se aplica el proceso de lematización a las palabras restantes. En la última etapa de comparación de palabras con lista de conceptos, se realiza el proceso para determinar la distancia que existe entre cada una de las palabras del texto y los pares de antónimos propuestos por el usuario.

Se cuenta con repositorios o bases de datos, los cuales almacenan la información necesaria para que cada etapa lleve a cabo su tarea.

4.2 Bases de datos

Para implementar las bases de datos se utilizó el modulo del lenguaje de programación Perl *BerkeleyDB*. Este módulo permite crear archivos de almacenamiento de gran tamaño de una manera muy sencilla y de rápido acceso mediante el uso de cuatro tipos de base de datos (hash, btree, queue y recno). Es de libre acceso y puede ser descargado en Internet[2].

[2]Web de distribución: http://search.cpan.org

Las bases de datos utilizadas en el sistema fueron implementadas con el tipo *Btree* enlazándolas a objetos tipo *hash* dentro del código de Perl. Para crear o leer los archivos, simplemente hay que escribir o leer en los objetos tipo hash. Esto permite que el acceso al contenido de las bases de datos sea muy sencillo, ya que al estar enlazado a objetos tipo hash, su lectura es muy rápida. Al usar objetos de tipo hash, la información de las bases de datos está almacenada como tablas hash, en donde existe una llave y un valor para cada una de ellas. El contenido de las llaves debe ser único, mientras que el de su respectivo valor, puede ser repetido en el valor de otra llave. En la Figura 7 se muestra la forma de crear una base de datos y enlazar una tabla hash a ella.

```
use BerkeleyDB;

my %hash;
my $filename = "archivo.db";

my $dbase = tie %hash, "BerkeleyDB::Btree",-Filename =>$filename,-flags =>
DB_Create or die "Cannot open file $filename: $! $BerkeleyDB::Error\n";
```

Figura 7. Uso del módulo BerkeleyDB para escribir en base de datos

La primera línea de código utiliza *use BerkeleyDB* para hacer uso de la librería. A continuación, se define un objeto tipo hash mediante *%* y otro tipo escalar con *$*, el cual contendrá el nombre de la base de datos *"archivo.db".* La siguiente línea de código, crea otro escalar el cual vinculará el contenido de la base de datos al objeto tipo hash, permitiendo su creación, o escritura.

En todo el sistema, se crearon un total de cinco bases de datos, en las cuales se almacenan diferentes tipos de datos.

4.2.1 Base de datos de *Replacements*

La gente que publica o da su opinión en redes sociales, generalmente lo hace de forma particular, deformando el lenguaje considerado correcto. Esto tiene muchas razones, como por ejemplo la cantidad de caracteres que Twitter permite escribir a la gente en sus tweets (nombre de los mensajes), con un máximo de 140, provocando que se deforme el lenguaje acortando palabras mediante el uso de una gran cantidad de abreviaturas o acrónimos; también porque muchos publican a través de dispositivos móviles queriendo escribir de forma rápida, ocasionando que los mensajes en redes sociales tengan muchas faltas de ortografía, abreviaciones o eliminación de apóstrofos. De igual forma, dada la naturaleza de las redes sociales, la gente utiliza emoticonos que son una secuencia de caracteres y que en un principio representaron una cara humana y expresan algún tipo de emoción. Posteriormente se crearon emoticonos más complejos con muchos otros significados. Ya que los mensajes que ocupa el sistema provienen de redes sociales, se utiliza una base de datos que contiene correcciones usuales del lenguaje, contracciones escritas de forma correcta y emoticones almacenados con su respectivo significado.

En la Tabla 7, se muestran algunos ejemplos de lo que está almacenado en esta base de datos. En las llaves de la tabla hash se encuentra la abreviatura, contracción o emoticono que puede ser encontrado en el texto original, y en valor de cada llave, se encuentra el significado, el cual será remplazado en el texto. Cada emoticono que se encuentra en la base de datos representa una llave, y es posible que más de un emoticono represente lo mismo. La lista de abreviaciones[3] así como la de emoticonos[4], se obtuvieron de fuentes en Internet. Existen 170 emoticones y 5,184 abreviaciones.

Tabla 7. Ejemplo de contenido de BD *Replacements*

Abreviación, contracción, emoticono	Significado
aykm	are you kidding me
b4	before
bff	best friend forever
gr8	great
lol	laughing out loud
lv	love
lil	little
sth	something
thx	thank you
arent	aren't
dont	don't
Im	I am
cannot	can not
:) , :-) , :] , :3	happy
XD , :D , =D	laughing
:*	kiss
:O, O.O , o.O , o_O	surprise
:(, :-(, :C , :[	sad
:'(	crying

4.2.2 Base de datos *Stopwords*

Las stopwords son palabras que carecen de algún significado gramatical y usualmente son filtradas antes o después de procesar algún texto. No existe alguna lista definitiva de stopwords que ocupen todas las herramientas o aplicaciones, ya que cualquier grupo de palabras puede ser escogida dependiendo de para qué propósito se necesita el texto.

La base de datos de stopwords contiene principalmente preposiciones en inglés, las cuales son filtradas antes de calcular la distancia entre conceptos y palabras de las opiniones. En esta tabla hash, dentro de las llaves sólo se almacena la palabra *stopword*, mientras que en los valores de cada llave se asigna un uno. Esto es porque no es necesario remplazar las palabras sino saber si están almacenadas en la base, por lo que al momento de buscar una palabra para filtrarla del texto, se lee

[3]Lista de Abreviaciones de Palabras en Inglés, http://www.noslang.com
[4]Lista de Emoticones, http://en.wikipedia.org/wiki/List_of_emoticons

el contenido en busca de la llave que la contenga. Si la palabra está en la tabla hash devuelve un uno y se quita del conjunto de palabras a analizar.

4.2.3 Bases de datos de *WordNet*

Se implementaron dos bases de datos referentes al contenido de WordNet. Una de ellas con el nombre de *words*, contiene todas las palabras que existen en WordNet, es decir los sustantivos, verbos, adjetivos y adverbios. Esta base de datos, al igual que la de stopwords, sólo almacena las palabras en la parte de llaves dentro de la tabla hash, y asigna un uno en los valores de cada llave. Esto con el propósito de verificar si una palabra que se encuentre dentro del texto y previa a buscar su trayectoria en el grafo, existe en WordNet, ya que si no, se evita hacer dicha búsqueda y se filtra la palabra. A pesar de haber muchas palabras, estas se encuentran almacenadas en una tabla hash, por lo que consultar si existe una palabra en específico dentro de esta base es relativamente rápido.

La segunda es la base de datos más importante de todo el sistema, ya que representa el grafo principal de WordNet; recibe el nombre de *distances*. Para mencionar su contenido, es necesario explicar cómo se representó el grafo en la base de datos.

Tomamos como ejemplo la palabra *angry* y observamos el contenido de sus synsets así como los conceptos con los cuales comparte algún sentido. En la Tabla 8 se muestran los sentidos de la palabra *angry*. Este concepto sólo está representado para la categoría gramatical adjetivo y contiene solamente tres sentidos, de los cuales, el segundo lo comparte con cuatro conceptos más.

Se toman todos los conceptos que aparecen en todos los synsets de *angry* y los consideramos como nodos para crear un grafo, tomando como nodo raíz a *angry* y nodos hijos a los demás conceptos sin repetir en los niveles inferiores a los conceptos que ya hayan aparecido previamente. En la Figura 8 se puede observar el grafo resultante. Para este ejemplo, a partir de los nodos hijos de *angry*, *furious*, *wild*, *raging* y *tempestuous*, se aplica el mismo proceso de obtener todos los conceptos que se encuentran en cada uno de sus synsets y se agregan al grafo ya creado a partir de *angry*, incrementando el número de niveles y nodos. En la Figura 9 mostramos el grafo resultante obtenido después de este proceso.

Tabla 8. Synsets de "angry"

Sentido	Palabras en el mismo synset	Glosa
angry#n#1	angry	*feeling or showering anger*
angry#n#2	angry, furious, raging, tempestuous, wild	*(of elements) as if showing violent anger*
angry#n#3	angry	*severely inflamed and painful*

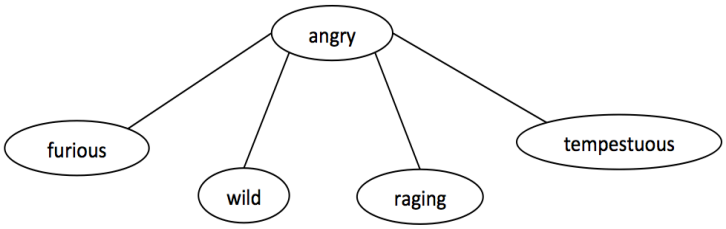

Figura 8. Grafo que representa los synsets de "angry"

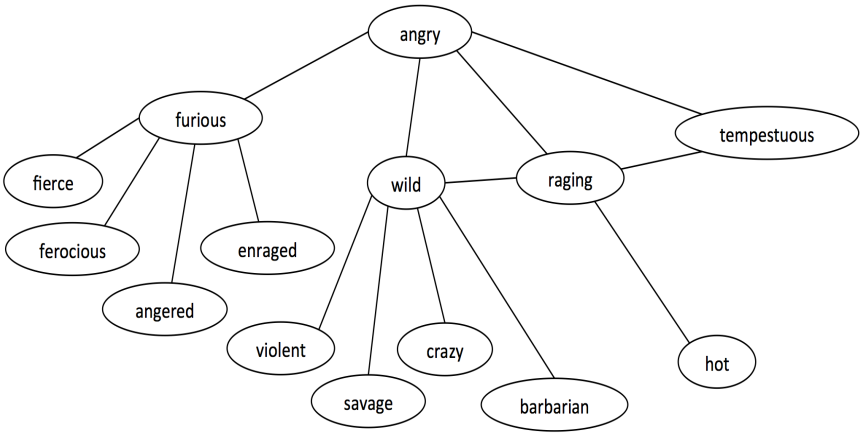

Figura 9. Representación de grafo después de haber expandido otro nivel

Recordemos que un synset es un conjunto de palabras que agrupa conceptos relacionados al mismo significado, por lo que dichos conceptos son sinónimos. Cada concepto representa un nodo por lo que dos nodos que estén unidos significa que son sinónimos para cierto sentido.

Para esta base de datos, se tomaron todos los conceptos existentes de cada categoría gramatical que contiene WordNet, y a partir de los synsets de cada uno de ellos, obtenemos los conceptos con los que están relacionados y se almacenan estos vínculos de sinonimia en la base de datos. Cada llave de la tabla hash utilizada para almacenar los vínculos en la base de datos, contiene un par de conceptos separados por una coma, mientras que en los valores de cada llave, se almacena un número asignado a partir del cálculo de logaritmos. En la sección 4.5 se explicará como se obtuvo dicho cálculo para cada par de conceptos. En la Tabla 9 se muestra el contenido de la base referente al grafo anteriormente explicado. Para la creación de esta base de datos, al igual que la de *words*, se utilizó un módulo de WordNet llamado *WordNet::QueryData*, mediante el cual se provee una interfaz para Perl directa a las bases de datos de WordNet y que previamente debe estar instalado para el uso de este módulo. *QueryData* permite al usuario acceso directo a todo el lexicón semántico de WordNet, soportando todas las categorías gramaticales con las que cuenta. De igual forma que *BerkeleyDB*, es de libre acceso y puede ser descargado del mismo sitio web.

Tabla 9. Representación del grafo en la base de datos

Vínculos de dos conceptos	
angry, furious	wild, savage
angry, wild	wild, crazy
angry, raging	wild, barbarian
angry, tempestuous	wild, raging
furious, fierce	wild, angry
furious, ferocious	raging, wild
furious, angered	raging, hot
furious, enraged	raging, tempestuous
furious, angry	raging, angry
wild, violent	tempestuous, raging

La Figura 10 muestra el uso de este módulo, en donde en la primera línea se incluye para su uso y segunda línea muestra la localización de la base de datos WordNet a QueryData invocando la función "new" con un parámetro que indica no cargar los índices al comienzo. Estas dos líneas son necesarias para el uso de este módulo. Existen dos funciones de consultas principales: 'querySense' y 'queryWord' las cuales, son las funciones que acceden a WordNets. QuerySense accede a las relaciones semánticas (sentido a sentido); queryWord accede a las relaciones léxicas (palabra a palabra). Las siguientes líneas de código en el ejemplo, acceden a WordNet, devolviendo distintos resultados y recibiendo diferentes parámetros. La tercera línea devuelve los conceptos con los que comparte el sentido número siete de sustantivo para *cat*. La cuarta línea devuelve las categorías gramaticales con las que cuenta cualquier palabra que se le dé como argumento y se encuentre en WordNet; para el ejemplo fue la palabra *run*. En la quinta línea la función obtiene como argumento una palabra junto con su etiqueta gramatical, devolviendo el número de sentidos que cuenta para esa categoría; si la palabra que recibe como argumento no cuenta con esa categoría gramatical, el resultado es una cadena vacía. La siguiente línea utiliza la función 'listAllWords', la cual puede recibir como argumento uno de los cuatro tipos de categorías gramaticales, devolviendo todas las palabras correspondientes que existan en WordNet. En el ejemplo utiliza la función *count*, por lo que devolverá la cantidad de sustantivos que existen. Finalmente, la última línea utiliza la función queryWord, devolviendo el antónimo de la palabra *dark* como sustantivo para el primer sentido. El segundo argumento se refiere a antónimo.

```
use WordNet::QueryData;

my $wn = WordNet::QueryData->new(noload => 1);

print "Synset: ", join(", ", $wn->querySense("cat#n#7", "syns")),"\n";
print "Parts of Speech:", join(", ", $wn->querySense("run")),"\n";
print "Senses: ", join(", ", $wn->querySense("run#v")),"\n";
print "Noun count: ", scalar($wn->listAllWords("noun")),"\n";
print "Antonyms: ", join(", ", $wn->queryWord("dark#n#1","ants")), "\n";
```

Figura 10. Uso del módulo WordNet::QueryData

Para las funciones querySense y queryWord, existen diferentes argumentos que pueden utilizar cada una para obtener una variedad de resultados. La combinación de estas funciones y sus argumentos, nos ayudaron a la creación de las bases de datos *words* y *distances*.

4.2.4 Base de datos de *Paths*

La base de datos Paths sirve para almacenar trayectorias de dos conceptos en el grafo principal. Ya que algunas de las opiniones escritas en las redes sociales contienen una cantidad moderada de palabras, además de que el propósito es poder medir un número considerado de opiniones a la vez, es importante que el tiempo que tarda el sistema en medir la relación que tienen las opiniones con el conjunto de conceptos sea muy rápido, por lo que esta base de datos sirve para que, conforme el sistema calcula diversas trayectorias, se almacenen y no sea necesario volverlas a calcular si las mismas palabras aparecen en otras opiniones. A pesar de que el algoritmo utilizado entrega las trayectorias de dos conceptos en el grafo con rapidez, la consulta de esta base de datos es siempre mucho más rápida, por lo que si se utiliza una gran cantidad de opiniones, el tiempo ahorrado es mayor.

4.3 Obtención del texto

La etapa de obtención del texto es la primera de las tres que conforman el sistema. En ella se realizan cinco tareas principales:

- Recibir cada una de las opiniones a través de un archivo de texto.
- Para cada opinión recibida, dividir su contenido en oraciones individuales.
- Búsqueda y remplazo de abreviaciones, contracciones y emoticonos.
- Identificar la etiqueta gramatical de cada palabra del texto.
- Búsqueda de palabras negadas.

Estas tareas fueron implementadas en dos pequeños programas realizados en Python y Perl. Ambos tienen como entrada y salida un archivo de texto lo que facilita la comunicación entre ellos.

4.3.1 División de oraciones

Las opiniones fueron recolectadas manualmente y guardadas en un archivo de texto, el cual es leído por un programa hecho en Python y es el encargado de dividir el texto en oraciones. Para esto se utilizó NLTK 3.0, una plataforma de libre acceso[5] para crear programas en Python y trabajar con datos de lenguaje humano. Dentro de esta plataforma se encuentra el módulo para realizar el proceso de tokenización el cual se realiza automáticamente al usar una serie de funciones ya establecidas.

[5] http://www.nltk.org

Este módulo se llama *Punkt Sentence Tokenizer* y es el encargado de dividir un texto en una lista de oraciones, mediante el uso de un algoritmo no supervisado [28] para construir modelos de abreviatura de palabras, colocaciones y palabras que inician oraciones. Detecta oraciones que terminan con diferentes puntuaciones: ';', ':', '.', '!' y '?'.

El programa recibe como parámetro el archivo de opiniones y lee una por una todas las que contenga dicho archivo, aplicando el proceso de tokenización de las oraciones. El resultado es la lista de oraciones obtenidas y almacenadas en otro archivo, en donde cada línea de texto es una oración diferente. Se realiza la tokenización del texto en oraciones, ya que posteriormente es necesario detectar negaciones.

4.3.2 Búsqueda y remplazo de texto

El siguiente programa está hecho en Perl. Lee cada una de las oraciones del archivo creado por el programa anterior y se almacenan temporalmente en el programa actual para su proceso, siendo la búsqueda de palabras a remplazar el primer proceso a realizar.

Primero es necesario tokenizar las palabras que conforman las oraciones para realizar una consulta en la base de datos *replacements*. Recordemos que se accede a la información de la base de datos mediante una tabla hash, por lo que si alguna abreviatura o contracción mal escrita es encontrada en alguna llave, se devuelve su valor, el cual representa la forma correcta de escritura, y remplaza a la que se mandó a verificar en la base. Este mismo proceso se realiza para los emoticonos, ya que al poderse separar del texto como tokens y buscarlo en la base, se devuelve su significado en concepto, remplazando al emoticono.

4.3.3 Etiquetado gramatical

El siguiente paso para la oración que se está analizando, es el etiquetado gramatical. Para esto, se utilizó el módulo de Perl *Lingua::EN::Tagger* de libre acceso, el cual es un etiquetador basado en corpus basado en probabilidades, entrenado para el inglés. Este módulo asigna etiquetas gramaticales a textos en inglés, basándose en un diccionario de búsqueda, un conjunto de valores de probabilidades y bigramas; es decir, examina la etiqueta previamente asignada a una palabra para determinar la etiqueta apropiada para la palabra actual. Aquellas palabras que son desconocidas se clasifican de acuerdo a la morfología de la palabra o pueden ser agrupadas para ser tratadas como sustantivos u otra etiqueta.

La razón para limpiar el texto mediante un diccionario de abreviaturas, contracciones y emoticones es para que el etiquetador realice un mejor desempeño. Al utilizar contracciones sin apóstrofos, el etiquetador no las reconoce, incluyendo la negación. De igual forma, es importante

remplazar los emoticones y abreviaciones, ya que pueden significar adjetivos o sustantivos que reflejen uno de los conceptos a analizar. Este etiquetador puede realizar el proceso de manera muy eficiente y rápida, ya que puede etiquetar de manera correcta las palabras como conjunciones o preposiciones, cualquier tipo de adjetivos o adverbios de tipo comparativos o superlativos, posesivos, verbos conjugados en cualquier tiempo como infinitivos, gerundios en pasado o en tercera persona. De igual forma, detecta cualquier tipo de puntuación y asigna su etiqueta correspondiente.

4.3.4 Negaciones

Una vez etiquetada la oración, es importante detectar si contiene alguna negación. Para esto, nos basamos en el algoritmo más sencillo propuesto por Das Sanjiv y Mike Chen en 2001 [53]. Ellos proponen en su trabajo que una vez encontrada una negación, se les agrega a las palabras que se encuentren entre la negación y la siguiente puntuación, la etiqueta "NOT_". Por ejemplo si se tiene la oración: "I didn´t like this movie, but I…" se modificaría a "I didn't NOT_like NOT_this NOT_movie, but I".

Este método es usualmente utilizado en el área de análisis de opinión, ya que el objetivo es clasificar un texto, reseña, opinión, etc., de acuerdo a su polaridad positiva o negativa, por lo que es importante detectar, en el caso del uso de cualquier negación, una opinión negativa.

La detección de la negación que se implementó, sigue el mismo principio. Una vez obtenidas las oraciones con sus palabras y sus respectivas etiquetas gramaticales, nos fijamos en aquellas etiquetas que indican una negación. Es importante denotar que ya que nosotros no buscamos una polaridad, no nos interesa cualquier negación sino sólo aquellas que modifiquen a los sustantivos en una oración. Por ejemplo: si se tiene la oración: *"I don't like big houses"* y nos interesa medir la relación que tiene dicha oración con el concepto *big* o *small*. A pesar de haber una negación no cambia el hecho de que la oración se refiere a que la casa es grande. Por otro lado, si la oración es: *"This house isn't big"* y queremos la relación con los mismos conceptos, dicha negación sí modifica el tamaño por lo que si es importante identificarla al medir la relación con dichos conceptos. Las oraciones que se identifican con negaciones son aquellas que van acompañadas del verbo *to be* además de la palabra *nothing*, como por ejemplo en oraciones como "nothing good" o "nothings bad" en donde el significado es el opuesto. Ya que el verbo *to be* en español, tiene dos significados, sólo se consideran aquellas oraciones cuando a la negación le sigue un adjetivo y se ignoran aquellas a las que le sigue un verbo o una preposición.

Cuando las negaciones han sido identificadas, se verifica palabra por palabra hasta encontrar todos los adjetivos utilizando las etiquetas previamente asignadas para la búsqueda de sus

respectivos antónimos. Si se encuentran palabras como "*but*", "*however*" o "*nevertheless*" , se detiene la búsqueda de adjetivos.

Para encontrar los antónimos de los adjetivos, se utiliza el módulo WordNet::QueryData. Por ejemplo, la línea de código utilizada para encontrar el antónimo de *good* seria: $wn->queryWord("good#a#1","ants"), devolviéndonos "bad#a#1", es decir la palabra *bad* para el sentido uno en la categoría gramatical adjetivo.

Una vez limpiada la oración, realizado el etiquetado gramatical de cada palabra y modificado adjetivos en caso de que se haya requerido con el uso de alguna negación, la oración analizada se escribe en un archivo de texto para el siguiente proceso que se llevará acabo. Este archivo contiene por cada línea de texto, la misma estructura de la oración que la de las oraciones del archivo anterior, pero cada palabra junto con su etiqueta gramatical.

4.4 Procesamiento del texto

La segunda etapa es el Procesamiento del Texto y fue implementada mediante un programa en Python y el uso de NLTK 3.0. Esta etapa realiza una única tarea, la cual es aplicar el proceso de lematización, además de diferenciar los adjetivos de las demás palabras para el siguiente proceso.

La lematización es un proceso lingüístico que consiste en hallar el lema correspondiente de una palabra en alguna de sus formas flexionadas (plural, conjugada, o en español femenino, etc.). El lema es la forma estándar que se acepta como representante de todas las formas flexionadas de una misma palabra, por ejemplo el lema de una palabra es aquella palabra que se puede encontrar en un diccionario tradicional de definiciones, es decir *notebook* es el lema de *notebooks* y *take* es el lema de takes, *took*, *taken o taking*.

El proceso de lematización puede realizarse automáticamente mediante programas de análisis morfológico. Existen grados de lematización: morfológica o bien sintáctica que tenga en cuenta el contexto en el que aparecen las palabras.

Este programa recibe el archivo creado por el programa anterior el cual contiene la lista de oraciones conformado por las palabras junto con su etiqueta gramatical correspondiente, por lo que el proceso de lematización de sustantivos, verbos y adjetivos, utiliza sus etiquetas para llevarlo a cabo. Cuando el proceso se realiza para una oración, se reescriben en un archivo en el orden correspondiente en el que aparecen en la oración, eliminando cualquier puntuación y agregando la etiqueta "/adj" a todos los adjetivos en la oración analizada.

El archivo escrito a partir de este programa contiene todas las oraciones conformadas por sus palabras lematizadas y sin ninguna puntuación.

4.5 Comparación de palabras con lista de conceptos

El objetivo de esta etapa es obtener la trayectoria más corta de dos conceptos dentro del grafo creado a través de la base de datos *distances*. Es aquí donde el usuario define qué pares de antónimos se utilizarán para medir la relación que tienen con el texto.

El programa tiene algunas tareas más sencillas que realizar antes de calcular las trayectorias entre palabras. Se lee el archivo creado a partir de la etapa anterior y se analizan palabra por palabra para identificar los adjetivos, ya que posteriormente se realizan experimentos solamente tomándolos en cuenta; se verifica si contienen letras mayúsculas (nombres propios o inicio de oración) y se cambian a minúsculas; se realiza la búsqueda de *stopwords* a través de la base de datos, para quitarlas del conjunto de palabras a analizar; y finalmente, para las palabras que quedan de igual forma que las stopwords, se buscan en la base de datos que contiene todas las palabras de WordNet para descartar todas aquellas que no se encuentren en dicha base. Una vez aplicadas estas tareas, obtenemos el conjunto de palabras con más relevancia y a las que se podrá calcular su trayectoria en el grafo ya que se aseguró que se encuentran en WordNet.

Dada la forma en la que se crea el grafo, el algoritmo de búsqueda implementado fue el de búsqueda bidireccional junto con el de búsqueda a lo ancho, ya que se conocen ambos conceptos para los cuales se quiere encontrar la trayectoria. Esto conlleva algunas ventajas, como el que la velocidad para buscar y encontrar una solución se incremente considerablemente, así como el hecho de que si un concepto no está unido al grafo principal, se evita que el algoritmo siga buscando hasta recorrer completamente todo el grafo. Una búsqueda a lo ancho se realiza desde el nodo raíz hacia el nodo meta y otra desde el nodo meta hasta el nodo raíz. Dado que a partir de un concepto (nodo raíz y meta) se expanden todos los conceptos en sus synsets, si alguno de ellos está en la frontera de exploración de la otra búsqueda, ambas se han encontrado y la solución será la composición del resultado de las dos búsquedas: el camino desde la situación inicial a la meta pasando por el estado o concepto en común. La pertenencia a la frontera de exploración se puede comprobar en sólo uno de los grafos o en los dos.

A pesar de que en dos conceptos pueda existir más de una trayectoria, la forma de obtener los conceptos hijos es expandiendo siempre a partir de los primeros synsets, por lo que la solución, además de ser la más corta, contendrá los conceptos más relevantes posibles.

4.5.1 Variaciones del método propuesto

Se realizaron cinco diferentes variaciones del método y el programa entrega los resultados de las cinco cuando se realiza el proceso. Las variaciones son:

- Conteo de aristas
- Uso de logaritmos
- Uso de logaritmos y etiquetas gramaticales
- Orientación semántica de palabras
- Orientación semántica individual

Cada una de estas variaciones está basada en el uso del grafo de WordNet y la búsqueda de la trayectoria más corta de dos conceptos.

4.5.1.1 Conteo de aristas

La primera variación del método consiste simplemente en utilizar la trayectoria de dos conceptos en el grafo, en donde para cada par de nodos conectados entre sí, se le asigna una distancia de uno. Se calcula la distancia de una trayectoria sumando todas las uniones entre palabras o aristas que la conforman. Para cada palabra del par de antónimos, se calculan las trayectorias con cada una de las palabras seleccionadas del texto, formando un conjunto de distancias. Tantas palabras del texto seleccionadas, es el número *n* de distancias que contendrá el conjunto.

Para un conjunto que representa las distancias de un concepto *c* con las palabras del texto de w_i hasta w_n, obtenemos el promedio de dichas distancias *dp*:

$$dp = \frac{\sum_{i}^{n} d(w_i c)}{n}$$

A continuación dividimos el promedio *dp* entre la distancia máxima dm que existe dentro del grafo de WordNet que es quince (previamente calculado) y restamos a uno el resultado del cociente.

$$rel(t,c) = 1 - \frac{dp}{dm}$$

Este resultado representa la relación que existe entre un concepto del par de antónimos con el texto. De igual forma es necesario aplicar dicho procedimiento a cada uno de los conceptos almacenados en cada par de antónimos con las palabras del texto.

4.5.1.2 Uso de logaritmos

La variación del método utilizando logaritmos se basa en el hecho de que una palabra está considerada en diversas categorías gramaticales y para cada una de ellas varios sentidos o synsets en donde los primeros tienen mayor importancia o son más usados que los demás, es decir, la relación de sinonimia es más fuerte en los primeros sentidos a pesar de que todos los conceptos agrupados en un synset son sinónimos. Es por esto que se considera no tomar una distancia de uno entre nodos representados por conceptos para todos los sentidos de una palabra, sino asignar una distancia menor a los primeros synsets y mayor a los demás en un rango de cero a uno utilizando la función de logaritmo.

Por ejemplo en la Tabla 10, se muestran los synsets de la palabra *like* que contiene tres categorías gramaticales. Cada categoría, tiene diversos sentidos en donde podemos observar que los primeros contienen mayor relevancia que los demás.

Tabla 10. Synsets de la palabra "like"

Sentido	Palabras en el mismo Synset	Glosa
like#n#1	like, the-like, the-like-of	*-A similar kind*
like#n#2	like, ilk	*-A kind of person*
like#a#1	like, similar	*-Resembling or similar; having the same or some of the same characteristics; often used in combination*
like#a#2	like, same	*-Equal in amount or value*
like#a#3	like, alike, similar	*-Having the same of similar characteristics*
like#a#4	like, comparable, corresponding	*-Conforming in every respect*
like#v#1	like, wish, care	*-Prefer or wish to do something*
like#v#2	like	*-Find enjoyable or agreeable*
like#v#3	like	*-Be found of*
like#v#4	like	*-Feel about or towards; consider, evaluate or regard*
like#v#5	like	*-Want to have*

Para asignar una distancia a cada concepto de cada sentido, utilizamos la fórmula de logaritmo, tomando en cuenta la cantidad de sentidos que tiene la palabra *like* en cada categoría gramatical. Recordemos algunas fórmulas básicas de los logaritmos:

$$\log_N N = 1 \qquad \log_N 1 = 0$$

El logaritmo base N del mismo número N siempre es igual a uno, y logaritmo base N de uno, siempre es cero siempre y cuando N sea diferente de uno. Ya que la cantidad de sentidos nunca es igual en las palabras de WordNet, es necesario definir una fórmula que tome en cuenta dicha cantidad en cada palabra y asigne una distancia de entre cero y uno, en donde a los primeros conceptos contenidos en los sentidos se les asigna una menor distancia que a los demás.

Entran en consideración dos casos a tomar en cuenta al asignar las distancias correspondientes:

- Si una palabra contiene más de un synset, se toma el número total de synsets N y se define una constante C como :

$$C = \frac{N-1}{N+1}$$

De igual forma, se define la formula de distancia d_n como:

$$d_n = \log_N(1+nC)$$

en donde n es el número que representa al sentido al que se le está asignando la distancia, asignando 1 al primer sentido y así sucesivamente. El resultado de la formula devuelve la distancia que se asignará a todos los conceptos agrupados en el synset *n*.

- Si una palabra contiene solamente un synset agrupando otros conceptos, se considera que *N* es igual a uno y la distancia para ese *n* synset es 0.

Para el ejemplo de la palabra *like*, podemos obtener la distancia que se le asignará a cada concepto dentro de cada synset. En la Tabla 11 se muestran las distancias asignadas. Es posible observar que la palabra *similar* aparece en dos sentidos, por lo que se le asignaría la mayor distancia.

Tabla 11. Distancias para cada synset de "like"

Sentido	Palabras en el mismo Synset	Distancia
like#n#1	like, the-like, the-like-of	$d_n = \log_2 (1+(1*1/3)) = 0.415$
like#n#2	like, ilk	$d_n = \log_2 (1+(1*1/3)) = 0.736$
like#a#1	like, similar	$d_n = \log_4 (1+(1*3/5)) = 0.339$
like#a#2	like, same	$d_n = \log_4 (1+(2*3/5)) = 0.568$
like#a#3	like, alike, similar	$d_n = \log_4 (1+(3*3/5)) = 0.742$
like#a#4	like, comparable, corresponding	$d_n = \log_4 (1+(4*3/5)) = 0.882$
like#v#1	like, wish, care	$d_n = 0$
like#v#2	like	$d_n = 0$
like#v#3	like	$d_n = 0$
like#v#4	like	$d_n = 0$
like#v#5	like	$d_n = 0$

La Figura 9 muestra el grafo resultante una vez que se asignen las distancias para cada concepto relacionado con la palabra *like*.

Estas distancias se calcularon para cualquier par de conceptos relacionados. Por lo que en la base de datos *distances*, se almacenan los pares de conceptos que están vinculados en las llaves y sus distancias calculadas con logaritmos en los valores de cada llave.

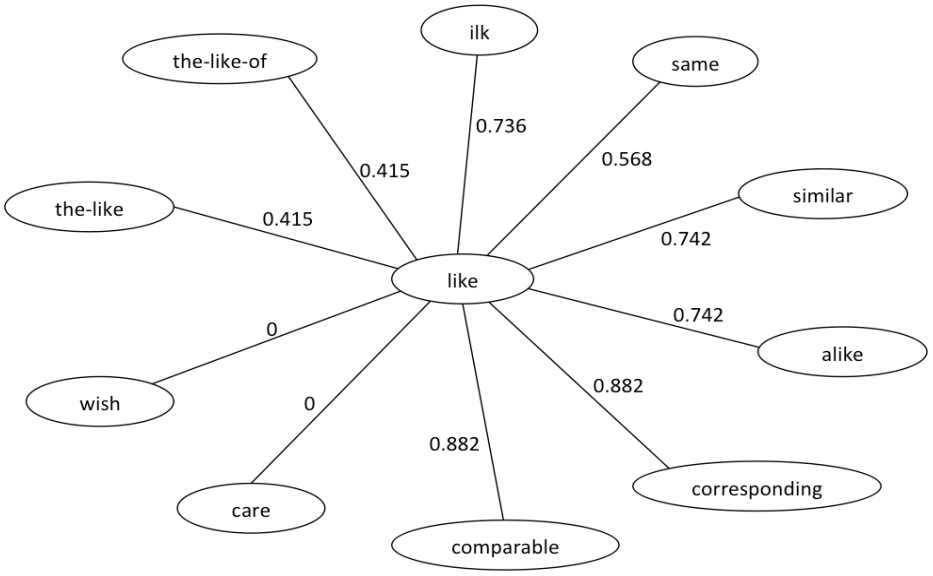

Figura 9. Grafo de la palabra "like"

Cuando se obtiene la trayectoria de dos conceptos conformada por varias distancias dentro del grafo, se obtiene el promedio para obtener la distancia total entre ambos conceptos. A continuación, para cada par de conceptos calculados, se obtiene el grado de relación para cada una de las trayectorias restando a uno su distancia total. Para obtener el grado de relación de cada concepto de los pares de antónimos y el texto, se obtiene el promedio de las relaciones de cada par de concepto previamente calculado.

4.5.1.3 Uso de logaritmos y categorías gramaticales

En esta variación del método se consideran, igual que el anterior, el uso de logaritmos para determinar la distancia entre nodos dentro del grafo; sin embargo, también se toma en cuenta la categoría gramatical de los sinónimos al definir una trayectoria conformada por varios nodos.

La forma de obtener una trayectoria es a través de synsets. Una palabra al estar en varios synsets y considerada en varias categorías gramaticales, se relaciona con otros conceptos, por lo que es posible hacer un cambio de categoría gramatical en el momento de buscar una trayectoria a través de dichos conceptos. Por ejemplo en la Figura 10, se muestra una trayectoria de dos conceptos pasando a través distintos synsets y conceptos relacionados.

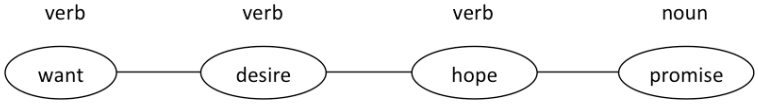

Figura 10. Trayectoria de *want* a *promise*

En esta trayectoria, podemos observar que es posible llegar de *want* a *promise* a través de distintos synsets pero cambiando de categoría gramatical. Dichas palabras son sinónimos para algún determinado synset, pero debemos recordar que los sinónimos deber pertenecer a la misma categoría gramatical.

Basándonos en las observaciones de que en la mayoría de las trayectorias que hacen uso de diferentes synsets para llegar de un concepto a otro lo hacen a través del uso de diferentes categorías gramaticales, en esta variación del método se propone que cada vez que en la trayectoria se realice un cambio de categoría gramatical de un nodo a otro, se sume un uno al total de la distancia obtenida y así compensar dicho cambio, es decir, penalizar la distancia.

La forma de obtener la relación de cada par de conceptos antónimos y el texto, es el mismo que la variación anterior.

4.5.1.4 Orientación semántica

En las variaciones utilizadas anteriormente, para obtener la relación que tiene un texto con una lista de pares de antónimos, se consideraba por separado cada uno de los conceptos de cada par para medir la distancia que existe entre ellos y cada una de las palabras que conforman el texto, sin que tuviera ninguna relación en ese proceso el antónimo del concepto utilizado. En esta variación, se pretende relacionar la distancia entre ambos antónimos en una sola formula y utilizarla para obtener la relación con el texto.

Como en el trabajo de Kamps, Marx, Mokken y Rijke [19] miden la orientación semántica de palabras tomando en cuenta tres factores, en esta variación proponemos medir de igual forma la distancia de las palabras que conforman el texto con los pares de antónimos propuestos por el usuario. La fórmula utilizada es:

$$OS(w) = \frac{dis(w,c_1) - dis(w,c_2)}{dis(c_1,c_2)}$$

Para medir la relación que tiene el texto con el par de antónimos (c_1 y c_2), tomamos cada una de las palabras que conforman dicho texto y utilizamos la orientación semántica tomando en cuenta los conceptos c_1 y c_2 al mismo tiempo en la formula. Dada una palabra *w*, se mide la distancia más corta con cada concepto c_1 y c_2 y que posteriormente se resta para ser normalizado por la distancia más corta que existe entre c_1 y c_2. Esta fórmula devuelve un resultado numérico entre -1 y 1. Si el resultado es menor que 0, se dice que la palabra *w* está del lado de c_1 y por lo tanto su grado de relación es más fuerte que con c_2. De igual forma, si el resultado es mayor a cero, *w* está del lado de c_2 y su relación es más fuerte.

Este proceso se realiza para todas las palabras que conforman el texto a analizar, y con todos los pares de antónimos. Una vez calculada la orientación semántica de cada palabra, se toman las palabras orientadas a cada uno de los conceptos c_1 y c_2 y se obtiene el promedio. El resultado es la relación que tiene el texto con cada uno de los conceptos antónimos.

4.5.1.5 Orientación semántica individual

La quinta y última variación del método, está basada en la orientación semántica propuesta en la variación anterior. Consiste en separar del cálculo de la formula, las distancias que tiene una palabra *w* de cada uno de los conceptos c_1 y c_2 y calcularlos de forma individual sin quitar la normalización de la distancia entre antónimos.

Para medir la orientación semántica de forma individual, se proponen las siguientes formulas, una para cada uno de los conceptos en los pares de antónimos propuestos.

$$OSi(w) = 1 - \frac{dis(w, c_1)}{dis(c_1, c_2)} \qquad OSi(w) = 1 - \frac{dis(w, c_2)}{dis(c_1, c_2)}$$

Al utilizar estas fórmulas, es posible que el resultado sea menor a cero cuando las distancias de la palabra w a algún concepto c_1 o c_2 es mayor que la distancia entre c_1 y c_2. Cuando esto sucede, simplemente se asigna cero a la orientación semántica de esa palabra *w*.

De igual forma que la variación anterior, una vez aplicadas estas fórmulas para todas las palabras de un texto, se calcula el promedio de distancias obtenido para cada concepto de antónimos y se obtiene la relación que tiene el texto con el par de antónimos calculados.

CAPÍTULO 5

EXPERIMENTOS Y RESULTADOS

En este capítulo se describen los experimentos realizados con las distintas variaciones del método propuesto. De igual forma, se describen los resultados y se comparan junto con las medidas de semejanza proporcionadas por WordNet contra un conjunto de datos elaborado por expertos.

5.1 Forma de evaluación

Para evaluar el desempeño del método propuesto y cada una de sus variantes, los resultados obtenidos mediante los experimentos se compararon con el criterio proporcionado por una encuesta a veinte personas. A ellas, se les dio la lista de las mismas opiniones a evaluar con el método y se les pidió que calificaran si la escritura de éstas, implicaba alguno de los pares de atributos proporcionados para cada opinión.

Cuando se obtienen los resultados del método propuesto para una determinada cantidad de opiniones y al compararlos con los resultados brindados por las personas, se puede saber qué tan precisa es cada una de las variaciones del método propuesto. La razón para utilizar estas encuestas para evaluar el desempeño del método, es porque a pesar de que existe una gran variedad de aplicaciones y trabajos realizados que utilizan semejanza entre conceptos, no existe uno que obtenga la semejanza entre un determinado texto y un par de conceptos establecidos.

De igual forma, se utilizaron las medidas de semejanza Hirst-St-Onge (HSO), Jiang-Conrath (JCN) y Resnik (RES) para obtener el grado de semejanza que tienen los conceptos antónimos con las opiniones y comparar estos resultados con los obtenidos en el método propuesto. Además se aplican las medidas de evaluación *Recall* y *Precision* y se comparan estos valores. Se utilizaron estas tres medidas porque tanto JCN y RES son dos de las más utilizadas, pero solamente se pueden comparar sustantivos, mientras que con HSO es posible comparar cualquier categoría gramatical.

5.2 Descripción de experimentos

A continuación se presentan las características tomadas en cuenta para el desarrollo de los experimentos para cada variación del método propuesto.

5.2.1 Conjunto de opiniones

Se recolectaron opiniones de las redes sociales de Facebook y Twitter. Además de ocupar las reseñas de dichos sitios, también se ocuparon opiniones de Google Shopping, ya que ofrece reseñas de productos con una mayor cantidad de palabras que las de las redes sociales. La recolección de opiniones a través de este sitio es solamente para contar con una mayor cantidad de palabras, haciendo que el método pueda ser probado con opiniones más complejas en comparación con las obtenidas de Facebook o Twitter. Además, facilita su recolección al mostrar varias de un solo producto.

Google Shopping es un servicio de Google que permite a los usuarios realizar búsquedas de una gran variedad de productos en sitios de venta en línea y comparar precios entre productos y vendedores. En algunos de los resultados devueltos en las búsquedas, se incluyen opiniones de los productos escritas por los usuarios que ya cuentan con ellos, asignando una calificación mediante

un sistema de estrellas que denota la calificación general del producto y no de aspectos individuales de éste. Estas calificaciones no se toman en cuenta para el presente trabajo.

Se recolectaron un total de cuarenta opiniones de diferentes productos o personas. En la Tabla 12 se muestra acerca de qué o de quién son estas opiniones así como la cantidad que se tomó en cuenta para realizar los experimentos. En el apéndice B, se muestra la encuesta aplicada así como las opiniones utilizadas para probar el método.

Tabla 12. Lista de opiniones

No Opiniones	Producto / persona
6	Consola de videojuegos
6	Productos electrónicos
6	Jugador futbol
5	Presidente
5	Videojuego
7	Jabón
5	Ciudad

Para cada uno de estos conjuntos de opiniones, se midieron diferentes aspectos que los usuarios reflejan en dichas opiniones. De igual forma, para todas y cada una de las opiniones, se consideró la medición entre los aspectos *good* y *bad*. Esto con el propósito de saber si el método puede detectar su polaridad. Los aspectos tomados en cuenta para las opiniones, se muestran en la Tabla 13.

Tabla 13. Aspectos de Opiniones

Producto/persona	Aspectos
consola de videojuegos	costo
productos electrónicos	apariencia, configuración
jugador de futbol	personalidad, habilidad
presidente	honestidad
videojuego	diversión
jabón	limpieza
ciudad	clima, apariencia

Para cada aspecto utilizado en cada opinión, se consideró su representación mediante un par de conceptos, los cuales son los dos conceptos más alejados entre sí dentro del grafo (antónimos) y así poder cuantificar la relación entre ambos y cada una de las palabras que conforman las opiniones.

5.2.2 Adjetivos

Una observación es que la mayoría de las medidas de semejanza o relación solamente se pueden aplicar a la relación de hiponimia de una jerarquía. Una notable excepción es la medida de Hirst y St-Onge [7], para la cual el método es aplicable a todas las categorías gramaticales de WordNet. Esta restricción provoca que las medidas de semejanza o relación sean solamente aplicables a las categorías gramaticales de sustantivos y verbos, descartando a las categorías de adjetivos y

adverbios. Sin embargo, estas categorías pueden ser cruciales para algunas aplicaciones, ya que son los modificadores o elaboran el significado de otras palabras.

Es por esto que para todas las opiniones anteriormente mencionadas, además de aplicar el método con sus variaciones tomando en cuenta todas las palabras que las conforman, también se aplican considerando solamente los adjetivos, de tal forma que para cada una de las variaciones del método propuesto, se muestran resultados tanto para todas las palabras, como para únicamente los adjetivos.

5.3 Resultados

Las tablas de los resultados obtenidos al aplicar cada una de las variaciones del método así como las medidas HSO, JCN y RES a todas las opiniones recolectadas se presentan en el apéndice A. Las tablas de los resultados de las variaciones del método propuesto representan los resultados para cada conjunto de opiniones, mostrando los resultados aplicando el método a) utilizando todas las palabras y b) solamente los adjetivos. Cada número sobre los resultados representa la opinión a la que se le aplicó la variación. Además, para cada resultado de los pares de antónimos obtenidos en cada una de las opiniones, se resalta la cantidad con mayor valor numérico, el cual representa que para esa opinión, dicho concepto tiene mayor relación que su contraparte antónima. Al final de esta sección se muestra la comparación de los resultados obtenidos para cada variación, con el criterio de las personas que realizaron el ejercicio, así como un análisis de los resultados.

5.3.1 Evaluación y comparación de resultados

Para evaluar los resultados obtenidos con el método propuesto, se compararon con los resultados de las encuestas aplicadas a 20 personas. A continuación, se muestran las tablas de los resultados obtenidos en las encuestas. Para cada conjunto de opiniones, la gente debía de calificar si cada una de ellas implicaba alguno de los conceptos que tenían como opción.

Tabla 14. Resultado de encuesta con conjuntos de opiniones "Consola de videojuegos" y "Productos electrónicos"

Consola de videojuegos		
Opinión	Polaridad	Costo
1	good	cheap
2	bad	costly
3	good	cheap
4	good	cheap
5	bad	cheap
6	good	cheap

Productos electrónicos			
Opinión	Polaridad	Apariencia	Configuración
1	good	pretty	easy
2	good	pretty	easy
3	bad	pretty	hard
4	good	pretty	easy
5	good	pretty	easy
6	bad	ugly	hard

Tabla 15. Resultado de encuesta con conjuntos de opiniones "Jugador de futbol " y "Presidente"

Jugador de futbol			
Opinión	Polaridad	Personalidad	Habilidad
1	good	proud	skillful
2	good	-	skillful
3	bad	proud	skillful
4	-	proud	skillful
5	good	modest	skillful
6	good	proud	skillful

Presidente		
Opinión	Polaridad	Honestidad
1	bad	corrupt
2	good	honest
3	bad	corrupt
4	good	honest
5	bad	corrupt

Tabla 16. Resultado de encuesta con conjuntos de opiniones "Videojuego" y "Ciudad"

Videojuego		
Opinión	Polaridad	Diversión
1	good	entertain
2	good	entertain
3	good	entertain
4	good	entertain
5	bad	boring

Ciudad			
Opinión	Polaridad	Clima	Apariencia
1	good	cold	pretty
2	bad	cold	pretty
3	good	cold	pretty
4	good	cold	pretty
5	good	hot	pretty

Tabla 17. Resultado de encuesta con conjunto de opiniones "jabón"

Jabón		
Opinión	Polaridad	Limpieza
1	good	clean
2	good	clean
3	good	clean
4	bad	dirty
5	good	clean
6	bad	dirty
7	good	clean

Con el número de conceptos pares propuestos para evaluar el grado de relación de cada opinión, podemos obtener el número de posibles aciertos que se pueden tener para cada uno de los conjuntos de opinión. En la Tabla 18 muestra la cantidad de cada conjunto.

Tabla 18. Aciertos de cada conjunto de opinión

Conjunto de opinión	Aciertos
Consola de videojuegos	12
Productos electrónicos	18
Jugador de futbol	18
Presidente	10
Videojuego	10
Jabón	14
Ciudad	15

Considerando las respuestas de las personas que realizaron la encuesta, al comparar los resultados con los obtenidos por cada variante y los obtenidos con las medidas de semejanza HSO, JCN y Res, y al aplicar las medidas de *Recall* y *Precision*, podemos saber qué tan preciso es el método propuesto. En la Tabla 19 podemos observar para cada variación y medida de semejanza, el número de aciertos y reactivos resueltos así como su porcentaje de Recall y Precision.

Podemos observar que la mejor variante del método propuesto es la de Orientación semántica individual considerando todas las palabras, ya que tuvo el mayor número de aciertos y reactivos resueltos. También los porcentajes de Recall y Precision son los más altos. La separación del valor numérico entre cada par de concepto proporcionado por esta variación, es la mayor en comparación con las demás.

Tabla 19. Evaluación de cada variante y medida de semejanza

Variación	Tipo	Aciertos	Resueltos	Recall	Precision
Aristas	Todas	72	96	74.22%	75.0%
	Adjetivos	63	89	64.94%	70.78%
Logaritmos	Todas	62	97	63.91%	63.91%
	Adjetivos	65	96	67.01%	67.70%
Log y POS	Todas	52	96	53.60%	54.16%
	Adjetivos	53	97	54.63%	54.63%
Orientación Semántica	Todas	72	95	74.22%	75.78%
	Adjetivos	62	92	63.91%	67.39%
OS Individual	Todas	78	95	**80.41%**	**82.10%**
	Adjetivos	58	86	59.79%	67.44%
Hirst-St-Onge	Todas	28	41	28.86%	68.29%
Jiang-Conrath	Sustantivos	51	95	52.57%	53.68%
Resnik	Sustantivos	13	21	13.40%	61.90%

CAPÍTULO 6

CONCLUSIONES Y TRABAJOS FUTUROS

En este capítulo se dan las conclusiones a las que se llegaron con el desarrollo del método propuesto. Se mencionan las aportaciones realizadas además de presentar propuestas para trabajos futuros.

6.1 Conclusiones

A pesar de que WordNet es un recurso léxico que se utiliza en una amplia variedad de aplicaciones y trabajos, éste tiene algunas limitaciones. Una de ellas es que muchas veces no es posible aplicar las definiciones de sentidos que se enlistan para cada una de las palabras que contiene, ya que son muy detalladas y la mayoría no se utilizan con mucha frecuencia en el mundo real. Esto provoca que muchas palabras contengan una amplia y extensa aplicabilidad en su uso, haciendo que cada uno de sus sentidos se relacionen a su vez con una gran cantidad de conceptos, y éstos, a formar más uniones entre los nodos del grafo principal, por lo que palabras con significados totalmente opuestos estén muy cercanas dentro del grafo. Por ejemplo, para las palabras *good* y *bad*, la trayectoria más corta tiene una distancia de cuatro aristas.

Para el método propuesto en el presente trabajo, la limitación principal en el uso de WordNet fue que no todas las palabras están unidas al grafo principal. Esto es un gran inconveniente, ya que muchas de ellas pueden aportar un gran valor a los experimentos realizados, pero no es posible aplicar el proceso para obtener la distancia entre dos palabras. Por ejemplo, la palabra *beautiful* no tiene ninguna relación de sinonimia con otra palabra al contar solamente con dos sentidos, a pesar de que para un criterio humano, puede ser sinónimo de *pretty* o por lo menos, tener una trayectoria relativamente cercana una con otra. Esto quiere decir que hay que ser muy cuidadosos al momento de elegir los conceptos organizados en pares de antónimos, para medir la distancia con cada una de las palabras de los textos a probar.

El conjunto de textos seleccionados para el sistema debe de reflejar información subjetiva, es decir, implicar una opinión de quien lo escribe, y no información objetiva, ya que no tendría caso obtener la relación entre conceptos o clasificar dichos textos. Las redes sociales son utilizadas diariamente por millones de personas en todo el mundo, por lo que sería relativamente sencillo obtener opiniones de los productos o servicios que ofrecen las empresas. Sin embargo, cuando se realiza una determinada publicación, la mayoría de la gente no expresa una opinión útil al producto, pero cuando lo hace, el número de palabras que utiliza es muy corto, incluyendo una o dos oraciones muy sencillas. Esto provoca que sea complicado obtener opiniones que aporten retroalimentación clara al producto.

Las opiniones que funcionan mejor para aplicar el proceso, son aquellas que realmente reflejen aspectos o características del producto, persona o servicio. En muchas opiniones, utilizan una gran cantidad de palabras que no reflejan o aportan alguna relación con los conceptos a medir. Es por ello que el método implementado trabaja de una mejor manera cuando no contiene demasiadas palabras, ya que todas se consideran para obtener su distancia y por lo tanto influyen en el resultado mejorándolo o perjudicándolo. Cuando se utilizan opiniones subjetivas, utilizando

palabras que sí reflejen atributos de los objetos y no una gran cantidad de palabras, se obtienen los mejores resultados.

Con respecto a las medidas JCN y RES, al no admitir éstas adjetivos en la comparación de semejanza, se utilizaron los sustantivos derivados de los adjetivos en los experimentos para comparar las opiniones de los usuarios. Desafortunadamente, al hacer esto puede perderse cierta fidelidad con respecto a la comparación original, pues ciertos sustantivos no pueden formarse directamente a partir del adjetivo, como en el caso de *costly* y *expensiveness*. El aplicar nuestro método utilizando solamente sustantivos y comparar con estas dos medidas, no fue posible de manera directa porque algunos de dichos sustantivos no estaban unidos al grafo principal de WordNet. Sin embargo, para mostrar que los resultados no son lo suficientemente buenos utilizando solamente sustantivos, incluimos aquí los cálculos de ambas medidas de semejanza, además como referencia para lo que tendría que hacerse si se deseara realizar un trabajo similar utilizando las medidas existentes definidas en WordNet. Ya que estas medidas no admiten adjetivos y nuestro método sí, podemos concluir que es muy importante tomarlos en cuenta por lo que obtenemos mejores resultados que con las medidas tradicionales existentes. Para el caso de la medida HSO, es posible realizar los experimentos utilizando las cuatro categorías gramaticales; sin embargo, podemos observar en la tabla 19, que su desempeño es menor comparando con nuestro método que incluye la medida de semejanza propuesta.

Al realizar los experimentos utilizando solamente adjetivos, se pudo comprobar que es mejor considerar todas las palabras. Esto es porque muchas opiniones contienen relativamente una cantidad pequeña de adjetivos por lo que no es muy conveniente probar el método con unas cuantas palabras. Se obtuvieron buenos resultados, incluso mejores que utilizando todas las palabras, cuando los adjetivos que se utilizan en las opiniones son los mismos que se encuentran en la lista de conceptos pares, por lo que los resultados numéricos que representan la relación entre pares de conceptos y texto, se separa, demostrando una mayor relación del texto con uno de ellos.

También podemos observar que en la mayoría de las variaciones, la relación numérica es muy cercana. Esto se debe a que muchos conceptos antónimos, están muy cercanos dentro del grafo principal por lo que sus trayectorias con cada una de las palabras del texto, contienen casi el mismo número de aristas.

La variación Orientación Semántica Individual fue el mejor método de las cinco variaciones, ya que fue la que obtuvo mejores resultados que las demás. La ventaja de esta variación se debe a que se normaliza por la distancia entre pares de conceptos antónimos, por lo que considera solamente a las palabras más cercanas, es decir, aquellas que contienen una distancia no mayor a la de sus conceptos a medir, y descartando aquellas más lejanas o con mayor distancia.

Es por esto que podemos observar que es mejor si las opiniones son más cortas y con mayor número de palabras que reflejan sus atributos, ya que las demás variaciones al considerar muchas y con trayectorias muy grandes, perjudicaban algunos de sus resultados.

Dado que la mayoría de las variaciones arrojaron valores numéricos no muy alejados entre sí, podemos concluir que el método no es muy conveniente para asignar una relación entre cada concepto de los pares de antónimos y el texto, ya que siempre existirá una trayectoria con una distancia entre conceptos y en algunos casos muy similar. Sin embargo, sí podemos decir que el uso de distancias en el grafo de WordNet, creado a partir de relaciones de sinonimia, es un método bueno para clasificar textos en una gran variedad de grupos que demuestren sus atributos, y no sólo clasificar de acuerdo a su polaridad.

6.2 Trabajos Futuros

- Incrementar la cantidad de relaciones que existen entre conceptos para crear el grafo principal. Dado que el grafo utilizado está hecho a partir de la relación de sinonimia, muchos conceptos no están unidos, por lo que limita el cálculo de distancias, al haber subgrafos o sólo conceptos que no están unidos al grafo principal.
- Considerar un análisis más exhaustivo, tomando en cuenta más características del texto como semántica, y no sólo trayectorias o distancias entre conceptos dentro del grafo de WordNet.

Referencias

[1] G. Varelas, E. Voutsakis, P. Raftopoulou. Semantic similarity methods in WordNet and their application to information retrieval on the web. In: 7th annual ACM international workshop on Web information and data management Bremen, Germany, 2005, pp. 10-16.

[2] R. Sinha, R. Mihalcea. Unsupervised graph-based word sense disambiguation using measures of word semantic similarity. In: ICSC, IEEE Computer Society, 2007, pp. 363–369.

[3] Kenneth E. Harper. Measurement of similarity between nouns. In: International Conference on Computational Linguistics, 1965.

[4] Agirre E, Alfonseca E, Hall K, Kravalova J, Pasca M, Soroa A. A study on similarity and relatedness using distributional and WordNet-based approaches. In: Proceedings of human language technologies: the 2009 annual conference of the North American chapter of the Association for Computational Linguistics NAACL '09. Association for, Computational Linguistics, 2009, pp. 19-27

[5] Budanitsky A, Budanitsky A. Lexical semantic relatedness and its application in natural language processing. Tech. Rep, 1999.

[6] Leacock and M. Chodorow C. Combining Local Context and WordNet Similarity for Word Sense Identification in WordNet. In: C.Fellbaum, editor, An Electronic Lexical Database, MIT Press, 1998, pp. 265-283.

[7] Graeme Hirst and David St-Onge. Lexical chains as representation of context for the detection and correction of malapropisms. In Fellbaum, chapter 13, 1998, pp. 305-332.

[8] Z. Wu and M. Palmer. Verb Semantics and Lexical Selection. In: Annual Meeting of the Associations for Computational Linguistics, Morgan Kaufmann Publishers, Las Cruces, New Mexico, 1994, pp. 133-138.

[9] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montréal Québec, 1995.

[10] Lin, Dekang. An information-theoretic definition of similarity. In: Proceedings of the 15th International Conference on Machine Learning, Madison, Wisconsin, 1998, pp. 296--304.

[11] Jiang, Jay J. and David W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of International Conference on Research in Computational Linguistics, TaiWan, 1997, pp. 19--33.

[12] Lesk, Michael. Automated Sense Disambiguation Using Machine-readable Dictionaries: How to Tell a Pine one from an Ice Cream Cone. Proceedings of 1986 SIGDOC Conference, Toronto, Canada, June 1986.

[13] Budanitsky, A. And Hirst, G. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In Workshop on WordNet and Other Lexical Resources, Second meeting of the North American Chapter of the Association of Computational Linguistics, Pittsburgh, June 2001.

[14] Patwardhan S, Pedersen T. Using WordNet-based context vectors to estimate the semantic relat- edness of concepts. In: EACL 2006 workshop making sense of sense-bringing computational linguistics and psycholinguistics together, Trento, Italy, 2006, pp. 1-8.

[15] Banerjee S, Pedersen T. Extended gloss overlaps as a measure of semantic relatedness. In: Proceedings of the eighteenth international joint conference on, artificial intelligence, 2003, pp. 805–810.

[16] Pedersen T, PakhomovSVS, PatwardhanS, ChuteCG. Measures of semantic similarity and relatedness in the biomedical domain. J Biomed Inform 40(3), 2007, pp. 288-299.

[17] Lingling Meng, Junzhong Gu. A new model for measuring Word sense similarity in WordNet, 2012.

[18] Rubenstein, Herbert and John B. Goodenough. Contextual correlates of synonymy. Comunications of the ACM, 8(10), 1965, pp.627-633.

[19] Q. Peng, L. Zhao, Y, Yu, W, Fang. A New Measure of Word Semantic Similarity based on WordNet Hierarchy and DAG Theory. International Conference on Web Information Systems and Mining. Shanghai China, 2009.

[20] J. Kamps, M. Marx, R. J. Mokken, and M. de Rijke, Using WordNet to Measure Semantic Orientations of Adjectives, Proceedings of the LREC'04, Lisbon, Portugal, 2004, pp.1115-1118.

[21] Taieb, M. A. H., Aouicha, M. B., & Hamadou, A. B. A new semantic relatedness measurement using WordNet features. Knowledge and Information Systems, 2005, pp. 1-31.

[22] T. Pedersen, S. Patwardhan, and J. Michelizzi. Wordnet::Similarity: Measuring the relatedness of concepts. In Proceedings of Fifth Annual Meeting of the North American Chapter of the Association for Computational Linguistics, Boston, MA, 2004, pp. 38-41,

[23] A. Sebti and A. A. Barfroush. A New Word Sense Similarity Meaure in WordNet. Proceedings of the International Multiconference on Computer Science and Information Technology, 2008. pp.369-373.

[24] M. Ghzizadeh, M. Naghibzadeh and S. Yasrebi. Using WordNet to Determine Semantic Similarity of Words. THe fifth International Symposium on Telecommunications 2010.

[25] M. Ghzizadeh and M. Naghibzadeh. Semantic Similarity Assessment of Words Using Weighted WordNet. December 2012.

[26] E. Agirre, E. Algonseca, K. Hall, J Kravalova, M Pasca and A Soroa. A study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL, Boulder Colorado, June 2099. pp. 19-27.

[27] Zhendong Dong and Qiang Dong. HowNet and the Computation of Meaning. World Scientific Publishing Co. Pte. Ltd, Singapore. 2010, pp. 53-56.

[28] Jinwu Hu, Liuling Dai, Bin Liu. Measure Semantic Similarity between English Words. The 9th Internaitonal Conference of Young Computer Scientists. 2008

[29] Xuexia Gao and Tao Kuang. Research of Words Similarity Model Based on HowNet. In International Conference on Electronic and Mechanical Engineering and Information Technology, 2011.

[30] J. Liu, J. Xu and Y. Zhang. An Approach of Hibrid Hierarchical Structure for Word Similarity Computing by HowNet. In International Joint Conference on Natural Language Processing, Nagoya, Japan, October 2013, pp. 927-931.

[31] Danushka Bollegala, Yutaka Marsuo and Mitsuru Ishizuka. Measuring Semantic Similarity between Words Using Web Search Engines. In International World Wide Web Conference Committee, Baniff, Alberta, Canada, 2007.

[32] Lu Zhiquang, Shao Werimin and Yu Zhenhua. Measuring Semantic Similarity between Words Using Wikipedia. In International Conference on Web Intformation Systems and Mining 2009.

[33] Milne, D., and Witten, I. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In Proceedings of the first AAAI Workshop on Wikipedia and Artificial Intelligence, 2008.

[34] T.K. Landauer, D. Laham, B. Rehder, and M.E. Schereiner. How Well Can Passage Meaning Be Derived Witout Using Word Order? A Comparation of Latent Semantic Analysis and Humans. Procedings 19th Ann. Meeting of the Cognitive Science Soc, 1997, pp.412-417.

[35] C.T. Meadow, B.R. Boyce and D.H. Kraft. Text Information Retrieval Systems. Second edition Academic Press, 2000.

[36] Li, Y., McLean, D., Bandar, Z., O'Shea, J., & Crockett, K. (2006). Sentence similarity based on semantic nets and corpus statistics. IEEE Transactions on Knowledge and Data Engineering, 18(8), 1138–1150.

[37] Kucera, H., Francis, W., and Caroll, J. (1967). Computational Analysis of Present Day American English. Brown University Press.

[38] N. Okazaki, Y. Matsuo, N. Marsumura and M. Ishizuka. Sentence Extraction by Spreading Activition with Refined Sentence Similarity. IEICE Trans. Information and Systems, vol. E86D, no 9. 2003. pp. 1683-1694.

[39] G. Tsatsaronis. Text Relatedness Based on a Word Thesaurus. Journal of Artificial Intelligence Reasearch 2010. pp. 1-39.

[40] B. Pang, L. Lee and S. Vaithyanathan. Thumbs up? Sentiment classification using Machine learning Techniques. EMNLP-2002, pp. 79-86

[41]A. Bermingham and A. Smeaton. Classifying Sentiment in Microblogs: Is Brevity and Advantage?. ACM International Conference on Information and Knowledge Management. Toronto, Ontario, Canada. 2010.

[42] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau. Sentiment Analysis of Twitter Data. In: Proc. ACL Workshop on Languages in Social Media 2011. pp.30-38.

[43] E. Kouloumpis, T. Wilson, J. Moore. Twitter Sentiment Analysis: The good the bad and the omg. In: Proceedings onf the ICWSM 2011.

[44] Liddy, E.D. Natural Language Processing. In: Encyclopedia of Library and Information Science, 2nd edn., Marcel Decker, Inc., New York, 2003.

[45] Locke, W.N.; Booth, D.A. Translation. Machine Translation of Languages. Cambridge, Massachusetts: MIT Press, 1965, pp. 15-23.

[46] Bath, England. Salton, G. (1971). The SMART Retrieval System: Experiments in Automatic Document Processing. Englewood Cliff, NJ: Prentice Hall.

[47] Terry Winograd Understanding Natural Language, Academic Press, Nueva York, 1972.

[48] Weizenbaum, Joseph. ELIZA: A Computer Program For the Study of Natural Language Communication Between Man And Machine, Communications of the ACM, v.9 n.1,January 1966, pp. 36-45.

[49] Kenneth Colby, PARRY Stanford University 1972.

[50] George A. Miller. WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 1995, pp. 39-41.

[51] Collins, A. M., and M. R. Quillian. Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behavior, 1969, pp. 240-247.

[52] Kiss, Tibor and Strunk. Unsupervised Multilingual Sentence Boundary Detection. Computational Linguistics 32, June, 2006, pp. 485-525.

[53] Das, Sanjiv and Mike Chen. Yahoo! for Amazon: Extracting market sentiment from stock message boards. In Proceedings of the Asia Pacific Finance Association Annual Conference (APFA), 2001.

Apéndice A

A.1 Conteo de aristas

Tabla 20. Resultados de conjunto de opiniones "Consola de videojuegos" con variación uno

	Todas las palabras					
	1	2	3	4	5	6
good	**0.659**	0.6118	**0.6567**	**0.6825**	**0.6606**	**0.6833**
bad	0.596	**0.6863**	0.5929	0.6349	0.6545	0.6222
cheap	**0.607**	0.5882	**0.6397**	**0.6413**	**0.6606**	**0.6611**
costly	0.548	**0.6431**	0.556	0.5778	0.5576	0.5833
	Adjetivos					
	1	2	3	4	5	6
good	**0.666**	**0.7111**	**0.6583**	**0.7143**	0.7	**0.7333**
bad	0.616	0.6778	0.6042	0.6571	**0.9**	0.6444
cheap	**0.633**	**0.6444**	**0.6417**	0.6095	**0.7667**	**0.7556**
costly	0.533	0.5889	0.5417	**0.619**	0.5667	0.6222

Tabla 21. Resultados de conjunto de opiniones "Productos electrónicos" con variación uno

	Todas las palabras					
	1	2	3	4	5	6
good	**0.6708**	**0.708**	0.65	**0.7818**	**0.6508**	**0.7451**
bad	0.6333	0.6391	**0.7167**	0.697	0.5587	0.6314
pretty	**0.5917**	**0.6046**	**0.6083**	0.6485	**0.5683**	**0.6588**
ugly	0.5708	0.5885	0.5917	**0.6545**	0.5333	0.6
easy	**0.675**	**0.6966**	0.6667	0.7091	**0.6286**	0.6784
hard	0.6417	0.6621	**0.6792**	**0.7394**	0.6063	0.6784
	Adjetivos					
	1	2	3	4	5	6
good	0.7	**0.75**	0.7	**0.8133**	**0.6667**	**0.8**
bad	**0.7222**	0.7333	**0.7111**	0.76	0.6	0.7333
pretty	0.6333	0.6667	**0.6444**	0.7067	**0.6267**	**0.7333**
ugly	**0.6556**	0.6667	0.6333	**0.7333**	0.6	0.68
easy	**0.7556**	**0.8667**	0.7111	0.76	**0.76**	**0.7867**
hard	0.7111	0.6833	**0.7333**	**0.8**	0.64	0.7067

Tabla 22. Resultados de conjunto de opiniones "Jugador futbol" con variaciones uno

	Todas las palabras					
	1	2	3	4	5	6
good	**0.7083**	**0.6848**	**0.7028**	**0.6444**	**0.7**	**0.6952**
bad	0.5667	0.6424	0.6139	0.5926	0.64	0.5952
proud	0.5417	**0.6364**	0.5667	0.4889	0.58	**0.581**
modest	**0.5667**	0.6242	**0.6028**	**0.6148**	**0.6067**	0.5762
skillful	**0.6583**	**0.6303**	**0.6528**	**0.5852**	**0.6533**	**0.6429**
unskilled	0.4333	0.4606	0.45	0.437	0.4333	0.4381
	Adjetivos					
	1	2	3	4	5	6
good	0.7333	0.7333	**0.6667**	**0.6533**	**0.7333**	**0.7733**
bad	**0.7333**	**0.8**	0.6133	0.5733	0.6933	0.6267
proud	0.6	**0.8**	**0.6**	0.4933	0.5867	**0.68**
modest	**0.9333**	0.6	0.5333	**0.5733**	**0.68**	0.5733
skillful	0.5333	**0.6667**	**0.6**	**0.5867**	**0.6933**	**0.7333**
unskilled	**1**	0.5	0.4133	0.4133	0.44	0.44

Tabla 23. Resultados de conjunto de opiniones "Presidente" con variación uno

	Todas las palabras				
	1	2	3	4	5
good	**0.6939**	**0.6923**	**0.6571**	**0.6744**	**0.6889**
bad	0.6152	0.5949	0.5857	0.6077	0.6194
honest	0.6455	0.6308	0.6048	0.6385	0.625
corrupt	**0.6576**	**0.6513**	**0.6476**	**0.6538**	**0.6639**
	Adjetivos				
	1	2	3	4	5
good	0.6667	0.6667	0.7333	**0.65**	**0.8333**
bad	**0.68**	**0.7333**	0.7333	0.5667	0.75
honest	0.6533	**0.6667**	0.6667	0.6167	**0.7667**
corrupt	0.6533	0.6	**0.7333**	**0.6667**	0.6167

Tabla 24. Resultados de conjunto de opiniones "Videojuego" con variación uno

	Todas las palabras				
	1	2	3	4	5
good	**0.7133**	**0.6615**	**0.7267**	**0.7367**	0.6063
bad	0.6267	0.6205	0.6733	0.65	**0.6952**
entertain	**0.6933**	**0.6821**	**0.7067**	**0.7167**	0.6254
boring	0.62	0.6359	0.6267	0.6633	**0.6921**
	Adjetivos				
	1	2	3	4	5
good	0.6667	**0.7111**	**0.8222**	**0.84**	0.6222
bad	0.6667	0.6444	0.7333	0.68	**0.6667**
entertain	**0.6**	0.6667	0.6444	0.6267	0.6
boring	0.5333	**0.6889**	0.6444	**0.7333**	**0.6667**

Tabla 25. Resultados de conjunto de opiniones "Jabón" con variación uno

	Todas las palabras						
	1	2	3	4	5	6	7
good	**0.697**	**0.7022**	**0.674**	**0.711**	**0.702**	0.611	**0.6611**
bad	0.648	0.6044	0.6	0.631	0.623	**0.677**	0.5944
clean	**0.757**	**0.7111**	**0.681**	**0.688**	**0.686**	0.638	**0.6889**
dirty	0.654	0.64	0.629	0.671	0.643	**0.738**	0.6306
	Adjetivos						
	1	2	3	4	5	6	7
good	0.766	**0.7833**	0.6	0.8	**0.733**	0.633	**0.6417**
bad	**0.833**	0.7	0.6	**0.9**	0.655	0.633	0.6
clean	**0.866**	**0.75**	0.6	0.7	**0.688**	0.6	**0.7**
dirty	0.711	0.6167	**0.666**	**0.733**	0.633	**0.8**	0.55

Tabla 26. Resultados de conjunto de opiniones "Ciudad" con variación uno

	Todas las palabras				
	1	2	3	4	5
good	**0.719**	**0.6556**	**0.719**	**0.7333**	**0.7263**
bad	0.6	0.5889	0.619	0.6364	0.6456
hot	**0.6524**	**0.6389**	**0.6714**	**0.6667**	**0.6912**
cold	0.5381	0.5833	0.581	0.6121	0.6
pretty	**0.6524**	**0.6056**	**0.6667**	**0.6667**	**0.6807**
ugly	0.5952	0.5556	0.6	0.6061	0.607
	Adjetivos				
	1	2	3	4	5
good	**0.7167**	**0.64**	**0.7733**	**0.7524**	**0.7429**
bad	0.6	0.56	0.6267	0.6571	0.7048
hot	**0.6167**	0.6	**0.6533**	**0.6571**	**0.7048**
cold	0.5333	0.6	0.56	0.6476	0.581
pretty	**0.6667**	**0.5867**	**0.76**	**0.6762**	**0.7333**
ugly	0.6083	0.5067	0.5867	0.6381	0.6381

A.2 Uso de logaritmos

Tabla 27. Resultados de conjunto de opiniones "Consola de videojuegos" con variación dos

	Todas las palabras					
	1	2	3	4	5	6
good	**0.436**	0.3785	0.3914	**0.3941**	0.3719	**0.4413**
bad	0.3969	**0.407**	**0.3973**	0.3734	**0.4688**	0.4067
cheap	0.3434	**0.3976**	0.3597	0.3629	0.3004	**0.4547**
costly	**0.3862**	0.3182	**0.3849**	**0.4029**	**0.3811**	0.4081
	Adjetivos					
	1	2	3	4	5	6
good	**0.444**	0.3029	**0.3703**	**0.325**	0.2017	0.4931
bad	0.4004	**0.3864**	0.3691	0.2908	**0.6026**	**0.5106**
cheap	**0.3761**	**0.4152**	0.3853	0.353	0.3059	**0.7427**
costly	0.324	0.3562	**0.4025**	**0.4057**	**0.3201**	0.4552

Tabla 28. Resultados de conjunto de opiniones "Productos electrónicos" con variación dos

	Todas las palabras					
	1	2	3	4	5	6
good	**0.4252**	**0.4328**	**0.3748**	0.3588	**0.4079**	0.3785
bad	0.4204	0.394	0.3522	**0.3706**	0.366	**0.4088**
pretty	**0.5491**	**0.5283**	**0.4801**	**0.4367**	**0.504**	**0.5765**
ugly	0.3722	0.3347	0.326	0.3258	0.3789	0.3687
easy	**0.4012**	0.3858	0.3146	**0.4097**	**0.3713**	**0.389**
hard	0.3773	**0.3881**	**0.3414**	0.263	0.3281	0.3473
	Adjetivos					
	1	2	3	4	5	6
good	**0.4304**	**0.5319**	**0.3324**	**0.3922**	**0.4073**	0.2762
bad	0.3761	0.2129	0.3081	0.3176	0.3488	**0.3526**
pretty	**0.486**	**0.5065**	**0.4752**	**0.4056**	**0.4778**	**0.5259**
ugly	0.312	0.2453	0.3294	0.2794	0.2618	0.3152
easy	**0.3852**	**0.6678**	**0.2223**	**0.4647**	**0.3976**	**0.2674**
hard	0.3079	0.3119	0.228	0.2209	0.3195	0.1629

Tabla 29. Resultados de conjuntos de opiniones "Jugador de futbol" con variación dos

	Todas las palabras					
	1	2	3	4	5	6
good	**0.522**	0.3739	**0.4183**	0.4174	0.3765	**0.429**
bad	0.4546	**0.3757**	0.3979	**0.4295**	**0.4282**	0.3786
proud	**0.6171**	**0.6526**	**0.6103**	**0.5898**	0.3899	**0.5843**
modest	0.3957	0.3671	0.3744	0.3599	**0.6476**	0.3738
skillful	**0.5483**	**0.5076**	0.4872	0.4886	**0.5886**	**0.4785**
unskilled	0.5126	0.4906	**0.4988**	**0.5683**	0.54	0.4746
	Adjetivos					
	1	2	3	4	5	6
good	**1**	0.2276	**0.5656**	**0.4218**	0.3641	**0.5025**
bad	0.3222	**0.373**	0.4401	0.3942	**0.4377**	0.3566
proud	**0.6003**	**0.7437**	**0.646**	**0.6538**	0.3678	**0.6653**
modest	0.35	0.3339	0.3908	0.3713	**0.6935**	0.3124
skillful	**0.585**	0.3086	**0.5928**	0.4474	**0.6051**	0.49
unskilled	0.5299	**0.4532**	0.5251	**0.5646**	0.5873	0.49

Tabla 30. Resultados de conjuntos de opiniones "Presidente" con variación dos

	Todas las palabras				
	1	2	3	4	5
good	**0.4203**	**0.3981**	0.4171	0.3841	0.4076
bad	0.4087	0.3859	**0.4593**	**0.4111**	**0.4169**
honest	0.4017	0.3852	0.39	0.3793	0.3626
corrupt	**0.4226**	**0.4386**	**0.4916**	**0.4545**	**0.4298**
	Adjetivos				
	1	2	3	4	5
good	0.3321	**0.3854**	0.1994	**0.4499**	0.2837
bad	**0.548**	0.3612	**0.345**	0.4128	**0.442**
honest	0.403	**0.494**	0.211	0.4479	0.2472
corrupt	**0.4686**	0.4824	**0.406**	**0.4614**	**0.4267**

Tabla 31. Resultados de conjuntos de opiniones "Videojuego" con variación dos

	Todas las palabras				
	1	2	3	4	5
good	**0.4146**	**0.3526**	**0.4274**	**0.5047**	0.4024
bad	0.3569	0.347	0.3366	0.3531	**0.4805**
entertain	0.3518	0.3472	0.3654	0.3722	**0.5466**
boring	**0.4876**	**0.5208**	**0.4843**	**0.5605**	0.3481
	Adjetivos				
	1	2	3	4	5
good	0.3242	0.3349	**0.4443**	**0.6309**	0.3664
bad	**0.3348**	**0.4511**	0.3384	0.3232	**0.4799**
entertain	0.3805	0.3099	0.3372	0.3012	**0.4994**
boring	**0.4245**	**0.5143**	**0.4026**	**0.6101**	0.3991

Tabla 32. Resultados de conjuntos de opiniones "Jabón" con variación dos

	Todas las palabras						
	1	2	3	4	5	6	7
good	0.39	**0.4064**	**0.439**	0.38	**0.463**	**0**.353	**0.4124**
bad	**0.433**	0.3749	0.388	**0.407**	0.35	**0.384**	0.3892
clean	**0.416**	0.3987	**0.429**	**0.353**	0.374	0.329	**0.3962**
dirty	0.356	**0.4085**	0.37	0.352	**0.377**	**0.401**	0.3777
	Adjetivos						
	1	2	3	4	5	6	7
good	0.265	0.2175	0.419	0.259	**0.385**	0.326	**0.38**
bad	**0.473**	**0.4127**	**0.477**	**0.632**	0.157	**0.357**	0.368
clean	**0.459**	**0.4307**	**0.503**	0.285	**0.353**	0.24	**0.3552**
dirty	0.275	0.4056	0.296	**0.362**	0.304	**0.512**	0.323

Tabla 33. Resultados de conjuntos de opiniones "Ciudad" con variación dos

	Todas las palabras				
	1	2	3	4	5
good	0.3831	0.3851	**0.3571**	**0.506**	0.342
bad	**0.4046**	**0.424**	0.3435	0.4049	**0.3827**
hot	0.3035	0.2709	0.2894	0.3507	0.3664
cold	**0.3868**	**0.4492**	**0.3938**	**0.4916**	**0.3788**
pretty	**0.571**	**0.4999**	**0.5627**	**0.5914**	**0.5462**
ugly	0.3672	0.3473	0.3696	0.3684	0.346
	Adjetivos				
	1	2	3	4	5
good	0.3696	**0.4125**	0.2968	**0.5409**	0.3132
bad	**0.3982**	0.3633	**0.342**	0.3977	**0.3726**
hot	0.3616	0.2637	0.3461	0.379	**0.4768**
cold	**0.3865**	**0.5174**	**0.4121**	**0.5126**	0.3842
pretty	**0.6041**	**0.4948**	**0.6903**	**0.5631**	**0.5259**
ugly	0.3812	0.3272	0.3906	0.3992	0.3552

A.3 Uso de logaritmos y categorías gramaticales

Tabla 34. Resultados de conjuntos de opiniones "Consola de videojuegos" con variación tres

	Todas las palabras					
	1	2	3	4	5	6
good	0.079	**0.1726**	0.1238	0.1412	0.0513	0.0848
bad	**0.217**	0.0933	**0.1959**	**0.2171**	**0.1959**	**0.1151**
cheap	0.0702	**0.1094**	0.1192	**0.1583**	0.0527	**0.1789**
costly	**0.11**	0.0388	**0.1238**	0.1384	**0.0874**	0.0971
	Adjetivos					
	1	2	3	4	5	6
good	0.185	**0.2695**	0.1726	0.1762	0.0985	0.0598
bad	**0.254**	0.1363	**0.2355**	**0.2908**	**0.6026**	**0.0972**
cheap	0.039	**0.189**	0.121	**0.1843**	0	**0.5538**
costly	**0.226**	0.0538	**0.1849**	0.1437	**0.3201**	0.1176

Tabla 35. Resultados de conjuntos de opiniones "Productos electrónicos" con variación tres

	Todas las palabras					
	1	2	3	4	5	6
good	0.1211	0.118	0.1573	0.1721	0.1315	0.1481
bad	**0.1689**	**0.138**	**0.1677**	**0.2071**	**0.1372**	**0.2358**
pretty	0.1567	**0.1474**	**0.1872**	0.1364	0.1491	**0.2338**
ugly	**0.1794**	0.1202	0.1716	**0.1892**	**0.1794**	0.2126
easy	**0.12**	**0.1459**	0.1125	**0.2264**	**0.1152**	**0.165**
hard	0.0976	0.0904	**0.1482**	0.0923	0.0924	0.1124
	Adjetivos					
	1	2	3	4	5	6
good	0.1133	**0.217**	**0.2768**	**0.3255**	**0.2168**	0.2762
bad	**0.2205**	0.1641	0.1831	0.3176	0.1988	**0.3526**
pretty	0.0403	**0.1971**	0.2168	0.1723	**0.1562**	0.2626
ugly	**0.1783**	0.1838	**0.2196**	**0.2794**	0.1507	**0.3152**
easy	**0.1265**	**0.6053**	0.1334	**0.398**	**0.2702**	**0.1874**
hard	0.0944	0.1022	**0.1947**	0.1709	0.108	0.1629

Tabla 36. Resultados de conjuntos de opiniones "Jugador de futbol" con variación tres

	Todas las palabras					
	1	2	3	4	5	6
good	**0.2423**	0.1049	0.1406	0.1139	0.0742	**0.1754**
bad	0.1994	**0.196**	**0.1417**	**0.1791**	**0.2051**	0.1676
proud	**0.248**	**0.2987**	**0.1885**	0.1275	0.2168	**0.2558**
modest	0.051	0.1636	0.088	**0.1398**	**0.25**	0.1217
skillful	**0.25**	**0.2506**	**0.1637**	**0.1569**	**0.218**	**0.1776**
unskilled	0.1082	0.1466	0.1136	0.1568	0.1164	0.0925
	Adjetivos					
	1	2	3	4	5	6
good	**1**	0.2276	**0.3999**	**0.1434**	0.1065	**0.3416**
bad	0.3222	**0.373**	0.2794	0.1339	**0.2286**	0.2167
proud	**0.1003**	**0.4104**	**0.2921**	0.0863	0.2678	**0.3219**
modest	0	0.3339	0.1278	**0.1699**	**0.3049**	0.1425
skillful	**0.585**	**0.3086**	**0.2619**	**0.1996**	**0.308**	**0.2589**
unskilled	0	0.1139	0.072	0.1214	0.0419	0.1031

Tabla 37. Resultados de conjuntos de opiniones "Presidente" con variación tres

	Todas las palabras				
	1	2	3	4	5
good	0.1059	0.0962	0.1508	0.0951	0.1533
bad	**0.1966**	**0.1123**	**0.164**	**0.1492**	**0.2084**
honest	0.1675	0.0846	0.1106	0.1322	0.1305
corrupt	**0.2038**	**0.1169**	**0.1842**	**0.2015**	**0.2258**
	Adjetivos				
	1	2	3	4	5
good	0.2153	**0.3854**	0.1994	**0.1784**	0.1587
bad	**0.4346**	0.3612	**0.345**	0.1628	**0.4003**
honest	**0.3363**	0.094	0.211	0.2188	0.1639
corrupt	0.2184	**0.3158**	**0.406**	**0.2238**	**0.1938**

Tabla 38. Resultados de conjuntos de opiniones "Videojuego" con variación tres

	Todas las palabras				
	1	2	3	4	5
good	0.0999	0.016	0.1749	**0.202**	0.1388
bad	**0.119**	**0.1493**	**0.229**	0.1613	**0.1766**
entertain	**0.1811**	0.0702	**0.1562**	0.2032	0.1499
boring	0.0895	**0.1001**	0.1339	**0.2719**	**0.1619**
	Adjetivos				
	1	2	3	4	5
good	0.3242	0.0682	**0.4443**	**0.5737**	**0.2235**
bad	**0.3348**	**0.1654**	0.3384	0.2232	0.2132
entertain	0.1305	0.0836	0.0664	0.0835	**0.1597**
boring	**0.1388**	**0.2476**	**0.3074**	**0.4801**	0.138

Tabla 39. Resultados de conjuntos de opiniones "Jabón" con variación tres

	Todas las palabras						
	1	2	3	4	5	6	7
good	0.078	0.1142	0.091	0.08	0.173	**0.144**	0.055
bad	**0.245**	**0.1375**	**0.154**	**0.134**	**0.279**	0.085	**0.138**
clean	**0.266**	0.1011	**0.222**	0.086	**0.186**	0.053	0.093
dirty	0.156	**0.1028**	0.138	**0.097**	0.162	**0.22**	0.093
	Adjetivos						
	1	2	3	4	5	6	7
good	0.034	0.2175	0.219	0.259	0.224	**0.326**	0.0999
bad	**0.473**	**0.3127**	**0.262**	**0.632**	**0.279**	0.158	**0.2264**
clean	**0.459**	**0.3057**	**0.403**	0.115	**0.311**	0.124	0.1484
dirty	0.221	0.2098	0.296	**0.362**	0.27	**0.512**	**0.201**

Tabla 40. Resultados de conjuntos de opiniones "Ciudad" con variación tres

	Todas las palabras				
	1	2	3	4	5
good	0.1561	0.0839	0.1191	**0.2375**	0.0979
bad	**0.2133**	**0.1107**	**0.1867**	0.2266	**0.1973**
hot	0.0848	0.0478	0.09	0.1459	0.1244
cold	**0.1245**	**0.1986**	**0.142**	**0.2715**	**0.1581**
pretty	**0.2526**	**0.1598**	**0.2984**	**0.2834**	**0.2403**
ugly	0.1911	0.1528	0.2427	0.2513	0.1546
	Adjetivos				
	1	2	3	4	5
good	0.2188	**0.1154**	0.1169	**0.3504**	0.1518
bad	**0.2762**	0.0407	**0.2467**	0.2584	**0.2631**
hot	0.1308	0.0806	0.1312	0.2112	**0.2303**
cold	**0.1335**	**0.3096**	**0.135**	**0.2908**	0.1789
pretty	**0.3457**	**0.2138**	**0.536**	**0.3393**	**0.3432**
ugly	0.2443	0.1412	0.2989	0.338	0.2075

A.4 Orientación semántica

Tabla 41. Resultados de conjuntos de opiniones "Consola de videojuegos" con variación cuatro

	Todas las palabras					
	1	2	3	4	5	6
good	**0.308**	0.0536	**0.3403**	**0.3269**	**0.225**	**0.375**
bad	0.058	**0.3929**	0.0278	0.0385	0.2	0.1
cheap	**0.211**	0.0513	**0.2763**	**0.2396**	**0.2833**	**0.3**
costly	0.033	**0.2308**	0.0175	0.0312	0	0.0667
	Adjetivos					
	1	2	3	4	5	6
good	**0.25**	0.125	**0.2344**	**0.25**	0	**0.5**
bad	0.062	**0.25**	0.0312	0.0357	**0.75**	0.1667
cheap	**0.25**	0.1111	**0.25**	0.0476	**0.5**	**0.4444**
costly	0	**0.25**	0	**0.0714**	0	0.1111

Tabla 42. Resultados de conjuntos de opiniones "Productos electrónicos" con variación cuatro

	Todas las palabras					
	1	2	3	4	5	6
good	**0.325**	**0.34**	0.0536	**0.4**	**0.4028**	**0.5769**
bad	0.1	0.04	**0.3393**	0.05	0	0.0192
pretty	**0.1667**	**0.1373**	**0.1515**	0.125	**0.1771**	**0.2273**
ugly	0.0909	0.0686	0.0909	**0.1667**	0.0625	0
easy	**0.2708**	**0.35**	0.2045	0.1667	**0.3036**	0.25
hard	0.1042	0.1	**0.2727**	**0.3056**	0.1786	0.25
	Adjetivos					
	1	2	3	4	5	6
good	0.0417	**0.125**	0.0833	**0.3**	**0.25**	**0.3**
bad	**0.125**	0.0625	**0.125**	0.1	0	0.05
pretty	0.0556	0.0833	**0.0833**	0.0333	**0.1667**	**0.1333**
ugly	**0.1111**	0.0833	0.0556	**0.1**	0.1	0
easy	**0.3333**	**0.6875**	0.2083	0.2	**0.6**	**0.4**
hard	0.1667	0	**0.2917**	**0.35**	0.15	0.1

Tabla 43. Resultados de conjuntos de opiniones "Jugador de futbol" con variación cuatro

	Todas las palabras					
	1	2	3	4	5	6
good	**0.6786**	**0.3125**	**0.4167**	**0.2812**	**0.3611**	**0.4423**
bad	0.0714	0.0938	0.0357	0.0625	0.1111	0.0385
proud	0.0952	**0.1429**	0.0779	0	0.1071	**0.1607**
modest	**0.1667**	0.1143	**0.1623**	**0.3036**	**0.1786**	0.1429
skillful	**0.4219**	**0.3182**	**0.3802**	**0.3125**	**0.4125**	**0.3839**
unskilled	0	0	0	0	0	0
	Adjetivos					
	1	2	3	4	5	6
good	**1**	0	**0.35**	**0.35**	**0.35**	**0.55**
bad	0	**0.25**	0.15	0.05	0.2	0
proud	**0.2857**	**0.4286**	**0.2**	0	0.0571	**0.2571**
modest	0	0	0.0571	**0.3036**	**0.2571**	0.0286
skillful	**0.75**	**0.3125**	**0.35**	**0.325**	**0.475**	**0.55**
unskilled	0	0	0	0	0	0

Tabla 44. Resultados de conjuntos de opiniones "Presidente" con variación cuatro

	Todas las palabras				
	1	2	3	4	5
good	**0.369**	**0.4545**	**0.3077**	**0.3523**	**0.3875**
bad	0.0595	0.0227	0.0192	0.0568	0.075
honest	0.0714	0.0926	0.0167	0.1019	0.0877
corrupt	**0.119**	**0.1667**	**0.1667**	**0.1574**	**0.2105**
	Adjetivos				
	1	2	3	4	5
good	0.15	0	0	**0.3125**	**0.5**
bad	**0.2**	**0.25**	0	0	0.1875
honest	0.0667	**0.1667**	0	0.0417	**0.375**
corrupt	0.0667	0	**0.1667**	**0.1667**	0

Tabla 45. Resultados de conjuntos de opiniones "Videojuego" con variación cuatro

	Todas las palabras				
	1	2	3	4	5
good	**0.4167**	**0.2727**	**0.275**	**0.4167**	0.025
bad	0.0556	0.0909	0.075	0.0556	**0.375**
entertain	**0.2667**	**0.2444**	**0.2889**	**0.2778**	0.0444
boring	0.0222	0.0444	0.0222	0.1	**0.2778**
	Adjetivos				
	1	2	3	4	5
good	0.125	**0.3333**	**0.4167**	**0.6**	**0.0833**
bad	0.125	0.0833	0.0833	0	**0.25**
entertain	**0.2**	0	0.0667	0	0
boring	0	**0.0667**	0.0667	**0.32**	0.2

Tabla 46. Resultados de conjuntos de opiniones "Jabón" con variación cuatro

	Todas las palabras						
	1	2	3	4	5	6	7
good	**0.312**	**0.3833**	**0.5**	**0.392**	**0.403**	0.027	**0.3553**
bad	0.062	0.0167	0.083	0.071	0.019	**0.361**	0.0395
clean	**0.34**	**0.3091**	**0.3**	**0.155**	**0.236**	**0.022**	**0.2941**
dirty	0	0.0182	0.0667	0.066	0.036	**0.422**	0.0471
	Adjetivos						
	1	2	3	4	5	6	7
good	0	**0.375**	0.25	0.125	**0.333**	0.062	**0.1875**
bad	**0.25**	0.0625	0.25	**0.5**	0.041	0.062	0.0312
clean	**0.4667**	**0.4**	0	0	**0.2**	0	**0.45**
dirty	0	0	**0.2**	**0.1**	0.033	**0.6**	0

Tabla 47. Resultados de conjuntos de opiniones "Ciudad" con variación cuatro

	Todas las palabras				
	1	2	3	4	5
good	**0.4464**	**0.3**	**0.4375**	**0.3864**	**0.375**
bad	0	0	0	0.0227	0.0156
hot	**0.3077**	**0.2576**	**0.2917**	**0.25**	**0.2812**
cold	0	0.1061	0.0278	0.1	0.0104
pretty	**0.2667**	**0.1667**	**0.2667**	**0.2**	**0.2556**
ugly	0.0667	0.0303	0.0333	0.0333	0.0222
	Adjetivos				
	1	2	3	4	5
good	**0.4375**	**0.3**	**0.55**	**0.3929**	**0.1786**
bad	0	0	0	0.0357	0.0357
hot	**0.2083**	0.2	**0.2333**	**0.1667**	**0.3095**
cold	0	0.2	0	0.1429	0
pretty	**0.2292**	**0.2**	**0.4333**	**0.1429**	**0.2619**
ugly	0.0833	0	0	0.0476	0.0238

A.5 Orientación semántica individual

Tabla 48. Resultados de conjuntos de opiniones "Consola de videojuegos" con variación cinco

	Todas las palabras					
	1	2	3	4	5	6
good	**0.25**	0.125	**0.3333**	**0.3333**	0	**0.5**
bad	0.083	**0.375**	0.1667	0	**0.625**	0.0833
cheap	**0.24**	0.1212	**0.2099**	**0.2222**	**0.3056**	**0.3889**
costly	0.074	**0.2424**	0.0556	0.1111	0.0278	0.2222
	Adjetivos					
	1	2	3	4	5	6
good	0	0.0417	0.0156	**0.0714**	0	0.1667
bad	**0.062**	**0.125**	**0.0312**	0	**0.625**	**0.0833**
cheap	**0.083**	0.1111	**0.1146**	0.0476	**0.4167**	**0.3889**
costly	0	**0.1389**	0.0208	**0.0952**	0	0.1111

Tabla 49. Resultados de conjuntos de opiniones "Productos electrónicos" con variación cinco

	Todas las palabras					
	1	2	3	4	5	6
good	0	**0.3**	0.1875	**0.5**	**0.25**	**0.45**
bad	**0.75**	0	**0.4375**	0.2083	0	0.1
pretty	**0.2**	**0.25**	**0.3**	0.2778	**0.3333**	**0.2833**
ugly	0.1667	0.2222	0.25	**0.3333**	0.2667	0.1833
easy	**0.75**	**0.75**	0.375	**0.625**	**0.875**	**0.75**
hard	0.5	0.0833	**0.625**	0.4375	0	0.375
	Adjetivos					
	1	2	3	4	5	6
good	0	**0.125**	0.0417	**0.35**	**0.05**	**0.25**
bad	**0.125**	0	**0.0833**	0.2	0	0.1
pretty	**0.1667**	0.1667	**0.2222**	0.2667	0.1667	**0.3333**
ugly	0.0833	0.1667	0.1389	**0.3333**	**0.2**	0.2
easy	0.125	**0.5**	0.125	0.25	**0.35**	**0.3**
hard	**0.1667**	0	**0.2083**	0.25	0	0.1

Tabla 50. Resultados de conjuntos de opiniones "Jugador de futbol" con variación cinco

	Todas las palabras					
	1	2	3	4	5	6
good	**0.5833**	**0.25**	**0.8333**	0	**0.75**	**0.5**
bad	0	0.1667	0	0	0.5	0
proud	**0.25**	**0.3214**	**0.2857**	0	0.2143	**0.2857**
modest	0.2286	0.254	0.2105	**0.2449**	**0.3714**	0.2143
skillful	**0.4286**	**0.3068**	**0.3807**	**0.25**	**0.35**	**0.3854**
unskilled	0.0357	0.0455	0.0398	0.0156	0.0125	0.0625
	Adjetivos					
	1	2	3	4	5	6
good	**1**	0.125	**0.2**	0	**0.15**	**0.3**
bad	0	**0.25**	0	0	0.1	0
proud	**0.4286**	**0.5714**	**0.2**	0	0.1143	**0.3429**
modest	0.1429	0.1429	0.0571	**0.1143**	**0.3143**	0.1143
skillful	**0.875**	**0.375**	**0.25**	**0.225**	**0.425**	**0.5**
unskilled	0.125	0.0625	0.025	0	0	0.05

Tabla 51. Resultados de conjuntos de opiniones "Presidente" con variación cinco

	Todas las palabras				
	1	2	3	4	5
good	0.2	**0.5**	0.25	**0.375**	**0.3125**
bad	**0.25**	0	0.25	0.0625	0.1875
honest	**0.2778**	0.2222	0.1905	**0.3333**	**0.3194**
corrupt	0.2444	**0.25**	**0.2167**	0.2745	0.3222
	Adjetivos				
	1	2	3	4	5
good	0	0	0	0	**0.375**
bad	**0.2**	0	0	0	0.3125
honest	0.1667	**0.1667**	0.1667	0.125	**0.4167**
corrupt	0.1667	0	**0.3333**	**0.1667**	0.0833

Tabla 52. Resultados de conjuntos de opiniones "Videojuego" con variación cinco

	Todas las Palabras				
	1	2	3	4	5
good	**0.3333**	0.25	**0.625**	**0.5833**	0
bad	0	**0.5**	0.25	0	**0.25**
entertain	**0.2667**	**0.35**	**0.32**	**0.38**	0.1
boring	0	0.1	0.04	0.12	**0.425**
	Adjetivos				
	1	2	3	4	5
good	0	0.0833	**0.4167**	**0.55**	0
bad	0	**0.1667**	0.1667	0	0
entertain	0	0	0	0.04	0
boring	0	**0.0667**	**0.0667**	**0.2**	**0.0667**

Tabla 53. Resultados de Conjuntos de Opiniones "Jabón" con variación cinco

Todas las palabras							
	1	2	3	4	5	6	7
good	**0.25**	**0.4167**	**0.5**	0.5	**0.583**	0	**0.375**
bad	0.2	0.1667	0	**0.625**	0	**0.375**	0.125
clean	**0.459**	**0.3333**	**0.25**	**0.28**	**0.275**	0.085	**0.3667**
dirty	0.085	0.0667	0	0.24	0.1	**0.4**	0.1667
Adjetivos							
	1	2	3	4	5	6	7
good	0.125	**0.1875**	0	0.25	**0.166**	0	**0.0938**
bad	**0.375**	0.125	0	**0.625**	0	**0.062**	0.0625
clean	**0.6**	**0.25**	0	0.1	**0.166**	0.05	**0.175**
dirty	0.133	0	0	**0.2**	0.066	**0.45**	0.025

Tabla 54. Resultados de conjuntos de opiniones "Ciudad" con variación cinco

Todas las palabras					
	1	2	3	4	5
good	**0.375**	0	**0.4**	**0.5833**	**0.25**
bad	0	0	0	0	0.125
hot	**0.2857**	0.2333	**0.2778**	**0.2292**	**0.2708**
cold	0.0238	**0.2667**	0.0926	0.1667	0.0417
pretty	**0.381**	**0.2083**	**0.4286**	**0.3333**	**0.3214**
ugly	0.119	0.0417	0.0952	0.1429	0.0595
Adjetivos					
	1	2	3	4	5
good	**0.125**	0	**0.25**	**0.25**	0.0714
bad	0	0	0	0	0.0714
hot	**0.0625**	**0.0333**	**0.1333**	0.1429	**0.2619**
cold	0.0208	0.2	0.0333	**0.1905**	0.0238
pretty	**0.25**	**0.0333**	**0.4333**	**0.2381**	**0.3571**
ugly	0.1042	0	0.0333	0.1429	0.0952

Tabla 55. Resultados de conjuntos de opiniones "Consola" con medias de semejanza de WordNet

HSO	1	2	3	4	5	6
good	**0.0729**	0.019	0.0102	0.0085	0.1197	**0.111**
bad	0.0338	**0.029**	0.0102	**0.0170**	0.1197	0.049
cheap	**0.0416**	0	0	0	0	**0.093**
costly	0.013	0	0	0	0	0
JCN	1	2	3	4	5	6
goodness	**0.060**	**0.0893**	**0.076**	**0.038**	**0.08**	**0.048**
badness	0.054	0.074	0.067	0.033	0.066	0.042
cheapness	0	0	0	0	0	0
expensiveness	**0.044**	**0.057**	**0.061**	**0.027**	**0.086**	**0.035**
RES	1	2	3	4	5	6
goodness	0.389	0.9435	1.089	0.194	1.2205	0.519
badness	0.389	0.9435	1.089	0.194	1.2205	0.519
cheapness	0.389	0.9435	1.268	0.194	2.0705	0.519
expensiveness	0.389	0.9435	1.268	0.194	2.0705	0.519

Tabla 56. Resultados de conjuntos de opiniones "Productos electrónicos" con medias de semejanza de WordNet

HSO	1	2	3	4	5	6
good	**0.0131**	**0.0302**	**0.0803**	0.1538	**0.024**	0.0284
bad	0.009	0.0221	0.0505	0.1538	0.0216	**0.0426**
pretty	0	0	0	0	0	0.0227
ugly	0	0	0	0	0	0.0227
easy	**0.0164**	0.0645	0	0.0769	0.0384	0
hard	0.0131	0.0645	0	0.0769	0.0384	0
JCN	1	2	3	4	5	6
goodness	**0.074**	**0.0865**	**0.0907**	**0.0842**	**0.0854**	**0.0643**
badness	0.0623	0.0746	0.0772	0.0732	0.0741	0.0576
beauty	0.0473	0.0586	0.0601	0.0585	0.0591	0.0481
ugliness	**0.0646**	**0.0768**	**0.0797**	**0.0755**	**0.0765**	**0.0591**
easiness	**0.0503**	**0.0563**	**0.0571**	**0.0585**	**0.0602**	**0.0481**
hardness	0	0	0	0	0	0
RES	1	2	3	4	5	6
goodness	1.5986	0.9752	1.04	0.989	0.899	0.1298
badness	1.5986	0.9752	1.04	0.989	0.899	0.1298
beauty	1.5986	0.9752	1.04	0.989	0.899	0.1298
ugliness	1.5986	0.9752	1.04	0.989	0.899	0.1298
easiness	**1.8556**	0.7857	0.7964	0.989	**1.0091**	0.1298
hardness	1.5986	0.7857	**1.0095**	**1.362**	0.899	0.1298

Tabla 57. Resultados de conjuntos de opiniones "Jugador de futbol" con medias de semejanza de WordNet

HSO	1	2	3	4	5	6
good	**0.1651**	**0.1079**	**0.1428**	**0.0742**	0.0178	0.0909
bad	0.1473	0.0511	0.0535	0.0488	0.0178	0.0909
proud	**0.0848**	**0.0909**	0	**0.0683**	0	0.017
modest	0.0401	0.034	0	0.0253	0	0.017
skillful	0	**0.0909**	0	0	0	0
unskilled	0	0.034	0	0	0	0
JCN	1	2	3	4	5	6
goodness	**0.072**	**0.0955**	**0.072**	**0.0755**	**0.0757**	**0.0826**
badness	0.0643	0.0815	0.064	0.067	0.0653	0.072
pride	0	0	0	0	0	0
modest	**0.0653**	**0.083**	**0.065**	**0.068**	**0.0617**	**0.0768**
skill	**0.0663**	**0.085**	**0.074**	**0.0722**	**1609587.49**	**2575339.97**
incapacity	0	0	0	0	0	0
RES	1	2	3	4	5	6
goodness	0	0	0.3895	0.1947	1.2995	0.947
badness	0	0	0.3895	0.1947	1.2995	0.947
pride	0	0	0.3895	0.1947	1.0863	**1.1012**
modest	0	0	0.3895	0.1947	**1.2995**	0.947
skill	0	0	0.8885	**0.445**	**1.8213**	**2.9164**
incapacity	0	0	0.3895	0.19475	1.2995	0.947

Tabla 58. Resultados de conjuntos de opiniones "Presidente" con medias de semejanza de WordNet

HSO	1	2	3	4	5
good	0.0292	**0.1049**	0	0	0.037
bad	**0.0312**	0.0491	0	0	0.037
honest	0	0	0	0	0
corrupt	**0.0078**	**0.0133**	0	0	**0.0115**
JCN	1	2	3	4	5
goodness	**0.0648**	**0.092**	**0.0787**	**0.0755**	**0.0855**
badness	0.0565	0.0789	0.0697	0.0667	0.0743
honesty	**0.0522**	**0.0721**	**0.0647**	**0.0623**	**0.0686**
corruption	0.0448	0.0614	0.056	0.0543	0.0593
RES	1	2	3	4	5
goodness	0.8791	1.0103	0.779	0.6892	1.2887
badness	0.8791	1. 0103	0.779	0.6892	1.2887
honesty	0.8791	1. 0103	0.779	0.6892	1.2887
corruption	0.8791	1. 0103	0.779	0.6892	1.2887

Tabla 59. Resultados de conjuntos de opiniones "Videojuego" con medias de semejanza de WordNet

HSO	1	2	3	4	5
good	**0.0568**	0.0923	0.0909	**0.1217**	0.0170
bad	0.0255	0.0923	0.0909	0.0526	0.0170
entertain	0	0	0	0	0
boring	0	**0.0081**	0	**0.0197**	0
JCN	1	2	3	4	5
goodness	**0.0643**	**0.0811**	**0.082**	**0.0671**	**0.0847**
badness	0.0568	0.0711	0.072	0.059	0.0737
entertainment	**0.0711**	**0.1108**	**0.1142**	**0.0943**	**0.0895**
boringness	0	0	0	0	0
RES	1	2	3	4	5
goodness	0.5193	0.779	1.1837	0.6491	1.61
badness	0.5193	0.779	1.1837	0.6491	1.61
entertainment	**0.9533**	**2.41**	**2.732**	**2.118**	**1.7987**
boringness	0.5193	0.779	1.1837	0.6491	1.61

Tabla 60. Resultados de conjuntos de opiniones "Jabón" con medias de semejanza de WordNet

HSO	1	2	3	4	5	6	7
good	0.0044	0.0429	0.0065	0.0588	0	**0.0098**	**0.0260**
bad	0.0044	0.0429	**0.0263**	0.0588	0	0	0.0104
clean	0	0	**0.0098**	0	0	0	0.0833
dirty	0	0	0	0	0	0	0.0833
JCN	1	2	3	4	5	6	7
goodness	**0.0742**	**0.0661**	**0.07**	**0.0775**	**0.068**	**0.0788**	**0.068**
badness	0.066	0.0596	0.0625	0.0686	0.0613	0.0693	0.0625
cleanness	0	0	0	0	0	0	0
dirtiness	**0.0667**	**0.0605**	**0.0632**	**0.0695**	**0.062**	**0.073**	**0.073**
RES	1	2	3	4	5	6	7
goodness	1.1455	0.3895	0.1947	0.779	0.5193	1.0488	**0.5193**
badness	1.1455	0.3895	0.1947	0.779	0.5193	1.0488	0.1947
cleanness	1.1455	0.3895	0.19475	0.779	0.5193	1.1773	0.779
dirtiness	1.1455	0.3895	0.19475	0.779	0.5193	1.1773	**1.1773**

Tabla 61. Resultados de conjuntos de opiniones "Ciudad" con medias de semejanza de WordNet

HSO	1	2	3	4	5
good	0	0.0781	0	0	0.039
bad	0	**0.0859**	**0.0104**	**0.0117**	**0.0781**
hot	0.0333	0.0625	0	0	0.0625
cold	0.0333	0.0625	0	0	0.0625
pretty	**0.0333**	0.0156	**0.0694**	**0.0781**	0
ugly	0.0125	0.0156	0.0347	0	0
JCN	1	2	3	4	5
goodness	**0.0758**	**0.0895**	**0.0822**	**0.081**	**0.0788**
badness	0.067	0.0775	0.072	0.0712	0.0695
beauty	**0.0687**	**0.0795**	**0.074**	**0.073**	**0.0712**
ugliness	0.0545	0.0612	0.0576	0.0574	0.0561
hotness	0.0533	0.0582	0.0594	0.0556	**0.065**
coldness	**0.0668**	**0.072**	**0.074**	**0.0698**	0.05742
RES	1	2	3	4	5
goodness	0.4395	0.7942	0.4674	0.6354	0.3338
badness	0.4395	0.7942	0.4674	0.6354	0.3338
beauty	0.4395	0.7942	0.4674	0.6354	0.3338
ugliness	0.4395	0.7942	0.4674	0.6354	0.3338
hotness	0.2596	0.3895	0.6676	0.3116	**0.4768**
coldness	**0.7623**	**0.7942**	**1.0484**	**0.9338**	0.3338

Apéndice B

A continuación se presenta la encuesta utilizada y aplicada a las 20 personas para realizar la evaluación del método. En ella se puede observar la lista de cada una de las opiniones utilizadas en inglés.

Encuesta de Opinión

A continuación se enlistan algunas opiniones en inglés obtenidas de redes sociales de siete diferentes productos o personas. En la parte de debajo de cada opinión, se muestran algunos conceptos que pueden tener dichas opiniones. Agrega una marca junto a cada concepto si crees que las opiniones contienen dichos aspectos, sino, deja el espacio en blanco.

Consola de videojuegos: Play Station 4

Para cada opinión se consideraron si son buenas o malas opiniones además del precio.

Easy to set up, the graphics are outstanding. Easy to use the interface. The controller is much better than the PS3 design. The PSN network is much more affordable than the 'other box' and more cheaper. I do not regret this purchase. Just waiting for more awesome games!

good _____ bad _____ cheap _____ costly _____

The controller feels nice but bigger. Its a major upgrade from the previous gen, The battery doesn't last as long though. I get about 5hrs of play before it says it's low. Maybe cause that blue light always on. Charges quick on my power charge station. The console looks fine but it become hot after some hours of gaming. The system is not good and sometimes get freeze so I have to turn it off and let it rest. Why always new gens have so high price and they don't work well?.

good _____ bad _____ cheap _____ costly _____

Wow! Sony did it again. The PS4 is a wonderful awesome system in every way. The system is quiet, fast, has a very sleek design and produces some amazing looking graphics. I'm very happy with my purchase. $399.99 is a great price for this system. My only real gripe is the battery life. The new dual shock is very comfy and improves design wise on the dual shock 3. I love the controller has it's very comfortable. Battery life is approximately 6 hours compared to 30 on the 3 which is 1/5 of the time... That stinks. Other then that everything is too notch and I would t expect any less from sony.

good _____ bad _____ cheap _____ costly _____

This is an amazing console. Very fast UI, games load and download fast, graphics are very nice. Games look and run very smooth and crisp. It's really unbelievable how far we've come in the gaming world. The price for this wonderful piece of technology is low and acceptable.

good _____ bad _____ cheap _____ costly _____

I do appreciate the price of sonys console, cause is lower than its competence, but the game system sucks and not good exclusives. Not fun at all...

good _____ bad _____ cheap _____ costly _____

The best console ever! i just love it. For what they did, its very cheap!, every penny is worthy. The only pity is that lack of games now, but we are gonna get them very soon.

good _____ bad _____ cheap _____ costly _____

Productos marca Apple

Para cada opinión se consideraron si son buenas o malas opiniones, la apariencia de cada producto y facilidad de configuración.

This is my first smartphone. Even high price but is worthy. I found it easy to use and master. I really like the speak function for texting and adding to notes for tasks. Taking pictures and sending to my contacts is also easy. I look forward to using the face time to 'visit' with my out-of-state contacts. I highly recommend this phone.

good _____ bad _____ pretty _____ ugly _____ easy _____ hard _____

This phone was a great price for being an open box item and one of the few that were left. It is much light compared to my iPhone 4, more durable, better overall pictures and sound, plus amazingly fast.

good _____ bad _____ pretty _____ ugly _____ easy _____ hard _____

This is the first time I buy an iphone. I feel so desperate cause many things I don't know how to use them or set them up, so hard for me to use this phone. But I have to admit, it looks nice. It seems this cellphone is just pretty about, just looks good. If you don't know much about technology, you cant use.

good _____ bad _____ pretty _____ ugly _____ easy _____ hard _____

Battery life is awesome...easy to use...I really have nothing negative to say about this MacBookPro. I am an early adopter mac from the mac-n-crash generation and avid PC user. I have owned this unit for almost a year now and have only restarted it 2-3 times. Just primarily because I thought I should. It was not because it was acting up in any way. It's not heavy at all.

good _____ bad _____ pretty _____ ugly _____ easy _____ hard _____

I love my Macbook Pro. It is so nice looking and it is really nice to type on. I love how the keyboard lights up also. The retina display feels like you are looking through a window. It is a great computer and really fast, and it's so light.

good _____ bad _____ pretty _____ ugly _____ easy _____ hard _____

This is my first mac laptop and it sucks. First of all I just purchased it today and it is already freezing up and acting slow. Then I tried downloading Microsoft Office for it, but it's tough cause there are too many problems with the download! I'm so disappoint

good _____ bad _____ pretty _____ ugly _____ easy _____ hard _____

Jugador de futbol Cristiano Ronaldo

Para cada opinión se consideraron si son buenas o malas opiniones, la personalidad y habilidad del jugador.

No doubt his skills and talent as a footballer are pretty good. Only problem is no one will ever love him quite as much as he does.

good _____ bad _____ proud _____ modest _____ skillful _____ unskillful _____

One of the greatest players ever to grace the beautiful game. Pace, agility, control he has all the skills needed in football. Wonderful to watch.

good _____ bad _____ proud _____ modest _____ skillful _____ unskillful _____

Arrogant, childish, so prideful and over dramatic. He's a very good footballer, but spoils himself by trying his best to get an opposing player sent off. A cheat, especially near the penalty area, and wouldn't think twice about rolling over and over again (as if he had been shot!!) even when the tackle was nowhere near him!! I have no respect whatsoever for him!

good _____ bad _____ proud _____ modest _____ skillful _____ unskillful _____

In my opinion not just the best footballer ever but the best footballer possible. However, he also epitomizes the whole stereotype of the temperamental, selfish cheating Latin, and therefore I cannot like him.

good _____ bad _____ proud _____ modest _____ skillful _____ unskillful _____

He is now the best player, so skillful and talent, but above all, so handsome!. Besides, he had helped and visited kids in hospitals!! Always with so cute smile!! Sometimes he is proud but all people famous and rich are just like that!

good _____ bad _____ proud _____ modest _____ skillful _____ unskillful _____

There is no better player than him. He is so good player. No body else can compare with him. His moves with the ball are so nice that it make so enjoyable to watch. Too pity that he is so proud of himself

good _____ bad _____ proud _____ modest _____ skillful _____ unskillful _____

Presidente Barack Obama

Para cada opinión se consideraron si son buenas o malas opiniones además de la honestidad del presidente.

In my opinion all he has done is lie, multiply national debt and cut army funds to hire new government employees, also he over-defends the minorities, I'm not racist but i think there is a certain limit to everything even when it comes to defending people. He has done bad things for our country, and make us weak as a nation.

good _____ bad _____ honest _____ corrupt _____

He is doing a wonderful job for this country. I reckon he should be president forever because he has provided me hope and honor and respect for others. He has also shown us that there are no losers and that everyone is a winner. One love, brother. Respect from Ambrose and Troy.

good _____ bad _____ honest _____ corrupt _____

He increased the taxes and in a down economy is not bright. Suggesting it spurns economy growth shows you don't understand the economy to any degree. You messing up Obama.

good _____ bad _____ honest _____ corrupt _____

Obama has been cleaning up George W. Bush's mistakes for the past years and trying to repair the damage down to our economy as well as our relations with the rest of the world. Especially in comparison to Mitt Romney, Obama is the clear and best choice, bsolutelly the best choise. We can be blaming Obama for people being unemployed or gas prices, these are functions of our economy. Obama 2012

good _____ bad _____ honest _____ corrupt _____

I do not see him doing anything but making life harder and more costly. I just see him making the government bigger and richer, taxing Americans more, taking our freedoms away, and shoving his far left agenda and his crappy healthcare, with a million hoops to jump through, down our throats. What about the transparency...and all the jobs. The only thing he does is lie to us.

good _____ bad _____ honest _____ corrupt _____

Videojuego The Last of Us

Para cada opinión se consideraron si son buenas o malas opiniones además de la diversión del videojuego.

This game is wonderful. I loved it all from start to finish. The storyline was awesome and it left me wondering what would happen next.

good _____ bad _____ entertain _____ boring _____

This is the best game of 2013 by far. Possible the best game ever released on the PS3. Naughty-Dog has outdone itself this time. Awesome graphics, great story and interesting characters. This game is definitely worth buying a ps3 for.

good _____ bad _____ entertain _____ boring _____

Awesome and good game and gameplay. It requires a lot of thought, planning, and patience. It's been a while since I have played games and this is great.

good _____ bad _____ entertain _____ boring _____

Few games can make you cry; The Last of Us is one of those. The game is good, but I don't think it is as good as reviews and people say. The story is intriguing, but I feel it is slow paced compared to other Naughty-Dog games. Still this is an exceptionally well made game by Naughty-Dog

good _____ bad _____ entertain _____ boring _____

Just another game of zombies. I don't know why people think is the best game ever. The way to create weapons its good and new, but that's all. The story is fine but boring. All you got to do is survive and not waste your resources. There are better games with x box.

good _____ bad _____ entertain _____ boring _____

Jabón Palmolive

Para cada opinión se consideraron si son buenas o malas opiniones además de la limpieza del jabón.

Palmolive soap is big in size and it smells like heaven (olive oil). After the shower, it makes your skin feels like baby soft and so clean.

good _____ bad _____ clean _____ dirty _____

I fell in love with Palmolive soap 45 years ago. It is real soap, it has a great scent, and is not dry. I prefer it over the high-priced handmade soaps out there. This is the real thing.

good _____ bad _____ clean _____ dirty _____

I've had a really odd obsession with Palmolive Milk & Honey soap since I was a kid and I could literally eat it because it smells so yummy...

good _____ bad _____ clean _____ dirty _____

This has to be the worst smelling soap I have ever used. The smell is so strong it gave me a headache. I tried it two days in a row. On day three it went out with the trash.

good _____ bad _____ clean _____ dirty _____

Nice soap. It lathers well, smells nice and classic. I use it for both body and hair. Palmolive leaves my skin and hair feeling soft smooth and never dry.

good _____ bad _____ clean _____ dirty _____

This is probably the worst bar of soap I have ever purchased. Considered throwing it out just from the way it smelled even before I opened the box. I decided to at least give it a try. It smells like something I would wash my car with and it does not give you a rich lather. Thank goodness the scent does not linger when you get out of the shower. I will not be buying this again.

good _____ bad _____ clean _____ dirty _____

I am so glad that I purchased this soap! The smell is great and the lather is what I would call "plush" its almost a thick foamy lather. The only other soap that I have used that has a lather like this is a department store facial soap. My skin feels soft hydrated. Love it and will definitely purchase again.

good _____ bad _____ clean _____ dirty _____

Ciudad de Nueva York

Para cada opinión se consideraron si son buenas o malas opiniones además del clima y apariencia de la ciudad.

There are building everywhere and so awesome cause they are so bright! its so pretty and modern city. The weather was freezing but it was ok. The streets are so clean and the people so nice. I just love it.

good _____ bad _____ cold _____ hot _____ pretty _____ ugly _____

It sucks! The people discriminate you cause you are from other contry. The weather too damn cold, and too many cars everywhere, traffic all the time and smoke. The streets are clean and some areas looks fine, but is not enough. The worst city i been.

good _____ bad _____ cold _____ hot _____ pretty _____ ugly _____

The city is nice so pretty. lights everywhere and many people. It is clean and most of the time freezing. I was shaking all the time in the street. I shouldn't have come in winter. I could go to skate at least.

good _____ bad _____ cold _____ hot _____ pretty _____ ugly _____

Prices of things are expensive, people nice, weather ok but little cold. Streets are clean and sometimes crowded. I think new york is good city.

good _____ bad _____ cold _____ hot _____ pretty _____ ugly _____

I have seen others great cities. New York has that times square and is just soso. Its pretty but not too much. There are many streets vendors of fast food. I think the food not so clean cause there are a lot of traffic and yellow cabs, usually mess on the streets. The weather just fine, Its not cold but i think depends on the season. I think is as good as any other city: many people, traffic and buildings.

good _____ bad _____ cold _____ hot _____ pretty _____ ugly _____