Benemérita Universidad Autónoma de Puebla
Facultad de Ciencias de la Computación

Análisis de sentimientos basado en aspectos: un modelo para identificar la polaridad de críticas de usuarios.

Tesis profesional para obtener el Título de:
Licenciado en Ciencias de la Computación

Presenta:
Miguel Ángel Rosales Quiroga

Asesores:
Dra. Darnes Vilariño Ayala
Dr. David Eduardo Pinto Avendaño

Puebla, Puebla. Marzo 2016

Agradecimientos

Dedico esta tesis a mis padres, **Regina Quiroga** y **Ferrer Rosales**, que siempre me han apoyado en mis estudios y mi formación personal. Sin duda alguna, en gran parte, lo que soy es gracias a ustedes.

A mi hermano **Jesús Manuel Rosales** y su esposa **Adi Janai Jiménez**, por el apoyo brindado durante mi preparación. Además de sus consejos durante esta última etapa de mi formación académica.

A mis amigos. **Silvia Serrano**, muchas gracias por apoyarme durante estos últimos años. **Luis Silva, Francisco Alejo y Alejandro Serrano**, por su apoyo y compañía durante toda la carrera.

Y finalmente, a mis asesores, la **Dra. Darnes Vilariño** y el **Dr. David Pinto** por su tiempo, apoyo y consejos para la realización de esta tesis.

Índice

Resumen

Actualmente, con el crecimiento de los usuarios de internet también ha aumentado la cantidad de datos generados en la red. Por lo que se hace importante desarrollar modelos que permitan obtener información a partir de dichos datos.

Esta tesis pretende detectar la polaridad de enunciados, párrafos o fragmentos de texto, con respecto a las entidades mencionadas (ej. Laptops, batería, pantalla) y sus atributos (ej. precio, diseño, calidad). El objetivo es identificar los aspectos de las entidades y el sentimiento expresado por cada aspecto para así generar una lista con la polaridad total de todos estos.

Se plantea la creación de un modelo para la identificación de características léxicas sintácticas para la detección de aspectos y la clasificación del sentimiento expresado en las reseñas en tres categorías: positivo, negativo y neutro. Se analizaron características como el etiquetado gramatical de las palabras, la similitud semántica entre palabras, la co-ocurrencia de palabras en un conjunto de documentos y la identificación de patrones para la identificación de aspectos.

Esta problemática forma parte de un conjunto de tareas propuestas por el SemEval 2016. Son un conjunto de tareas en el área del análisis semántico y procesamiento del lenguaje natural para las cuales no hay una solución total. Por lo cual, cada equipo participante presenta su propuesta para obtener los mejores resultados.

Capítulo 1. Descripción general

El Análisis de Sentimientos es una tarea de clasificación de textos dentro del área del Procesamiento del Lenguaje Natural, su objetivo es dado una opinión de usuario poder detectar la polaridad de ésta, ya sea positiva, negativa o neutra. El conocer la opinión que una persona tiene hacia un producto o servicio, es de gran ayuda para la toma de decisiones, ya que permite a otros posibles consumidores detectar la calidad del producto o servicio evaluado para utilizarlo. Es necesario el desarrollo de modelos automáticos debido a la gran cantidad de opiniones que puede tener un producto, manualmente es una tarea muy tediosa.

En esta tesis se plantea un modelo no supervisado para resolver el problema del Análisis de Sentimientos basado en aspectos. Para la detección de entidades y atributos, la detección de aspectos y la polaridad de las opiniones se realizan el análisis léxico de las reseñas así como también la semántica de las mismas. Se trabajó con un conjunto de datos de entrenamiento conformado por reseñas sobre Restaurantes y Laptops, estas reseñas proporcionadas por el comité de organización del SemEval 2016[1]. Estas se encuentran etiquetadas correctamente para su utilización como entrenamiento para los modelos a desarrollar.

1.1 Planteamiento del problema

El SemEval (*Semantic Evaluation*) es una serie de evaluaciones de sistemas computacionales de análisis semántico. Se presentan tareas sobre el análisis semántico de textos. El propósito de estas evaluaciones es explorar la naturaleza del significado en el lenguaje. Mientras que el significado es intuitivo para los seres humanos, la transferencia de esas intuiciones de análisis computacional ha demostrado ser difícil de alcanzar. En el marco del SemEval 2016, se presentó la tarea 5: Análisis de Sentimientos Basado en Aspectos. Esta se basa en la minería

[1] http://alt.qcri.org/semeval2016/

y agregado de opiniones sobre entidades específicas y sus aspectos. Se pretende obtener, dado una opinión de usuario, el aspecto mencionado en el texto. Es decir, el objeto del cual se está haciendo referencia. Para cada aspecto se debe detectar la categoría del aspecto, estas dadas dentro de un conjunto ya predefinido. Además, la tarea aborda la detección de la polaridad sobre cada aspecto, esto implica clasificar el sentimiento expresado por el usuario, ya sea positivo, negativo, neutro o conflictivo.

Esta tarea se conforma de 4 fases o *slots*. En el slot 1, las reseñas son procesadas para identificar las entidades y atributos mencionados en la crítica. Los conjuntos de entidades y atributos están predefinidos por el comité organizador de la tarea para cada dominio. El Slot 2 consiste en la detección de los aspectos mencionados en cada reseña. Estos aspectos deben ser parte de la sentencia analizada. Para el Slot 1-2, una vez encontrados los aspectos de cada *review*, cada uno de estos es clasificado con su entidad y atributo correspondiente. Y finalmente, el Slot 3, se basa en la detección de la polaridad expresada para cada elemento identificado.

1.2 Objetivos

Objetivo General

Desarrollar un modelo que permita descubrir el/los aspectos mencionados en una opinión y el sentimiento que esta expresa.

Objetivos Específicos

1. Estudiar los artículos publicados en el marco del SemEval 2015 para esta tarea.
2. Estudiar el lenguaje de programación Python y las herramientas Clips-Pattern.
3. Pre-procesar los datos de entrada.
4. Desarrollar el modelo para descubrir el aspecto.

5. Desarrollar un modelo para descubrir las entidades y atributos asociadas a cada aspecto.
6. Desarrollar un modelo que permita detectar la polaridad del aspecto descubierto (positivo, negativo, neutro, conflictivo).
7. Validar el modelo desarrollado con el corpus de restaurantes y de laptops en el marco de la Conferencia SemEval 2016.

1.3 Justificación

Hoy en día, existe un aumento exponencial de los datos generados en internet. Por esto es de gran utilidad desarrollar modelos y técnicas que permitan a usuarios obtener información de mayor calidad. El desarrollo de un modelo para la detección de la polaridad de una reseña sobre un producto es importante para la toma de decisiones sobre el mismo. Permite tener una idea sobre el producto basado en experiencias anteriores de otras personas. Debido a que existen infinidad de datos y reseñas en internet es necesario automatizar la detección de opiniones relevantes. Opiniones que ofrezcan información de utilidad para el posible consumidor, saber si la opinión es positiva o negativa. Además de conocer el sentimiento de las personas sobre diferentes aspectos o características de dicho producto.

1.4 Organización de la tesis

El presente trabajo de tesis está estructurado de la siguiente manera:

- **Capítulo 1 Descripción general:** En éste capítulo se detalla el planteamiento del problema, los objetivos a considerar y los resultados esperados.
- **Capítulo 2 Estado del arte:** Se exponen los trabajos relacionados realizados anteriormente. Se describen los estudios y artículos presentados sobre la tarea del Análisis de Sentimientos, así como también se exponen algunas propuestas presentadas en la conferencia del SemEval 2014.
- **Capítulo 3 Marco teórico:** Se abordan los conceptos teóricos para el desarrollo de este trabajo, como son: el Procesamiento del Lenguaje Natural, Análisis de Sentimientos, Etiquetadores Gramaticales, entre otros.
- **Capítulo 4. Modelo propuesto:** Aquí se expone el modelo propuesto para resolver el problema. Se describen las fases y los pasos seguidos para las diferentes sub-tareas así como su implementación.
- **Capítulo 5. Resultados obtenidos:** Se detallan los resultados obtenidos aplicando el modelo propuesto para los conjuntos de datos de Restaurantes y Laptops.
- **Capítulo 6. Conclusiones:** En esta sección se discuten los resultados obtenidos y si cumplen estos con los objetivos planteados inicialmente.

Capítulo 2. Estado del arte

En este capítulo se exponen propuestas interesantes sobre el análisis de sentimientos. Se exponen las herramientas utilizadas y las características analizadas por los autores de dichas propuestas. De igual manera, se muestran algunos enfoques presentados en el foro de competición del SemEval 2014.

2.1 Sobre el Análisis de sentimientos

En 2002, Pang y Lee, publicaron su trabajo sobre la clasificación de documentos en base al sentimiento expresado en estos [1]. Analizando reseñas sobre películas, encontraron que las técnicas de Aprendizaje Automático mejoran el rendimiento de las líneas base generadas por humanos. Emplearon tres algoritmos de Aprendizaje Automático: Naive Bayes (NV), Máxima Entropía (ME) y Máquinas de Soporte Vectorial (SVM). Obteniendo alrededor de un 80% de precisión siendo SVM la que mostró mejores resultados.

Uno de los primeros trabajos que introdujeron el término de Análisis de Sentimientos fue el presentado en 2003 por Nasukawa y Yi [2]. En esta publicación definen esta tarea como encontrar expresiones de sentimientos para un sujeto dado y determinar la polaridad de los mismos. En las investigaciones anteriores a esta, se realizaba el análisis de la polaridad general de un documento, sin embargo en este enfoque se trata de identificar la opinión de cada sujeto mencionado en el texto. El algoritmo expuesto consiste primero en generar el etiquetado gramatical de las palabras del texto para posteriormente identificar límites de frases y dependencias locales. Una vez realizado este paso, se analizan las frases y se identifican aquellas con un término que exprese sentimiento. Se emplea un diccionario creado manualmente para detectar la polaridad de este término, en caso de incluir expresiones negativas como “not” o “never” el valor de polaridad es invertido. Para evaluar su propuesta crearon un conjunto de datos

con 175 casos extraídos de páginas web, cada uno identificado manualmente con el sentimiento a cada sujeto, se modificó el diccionario agregando términos apropiados y se obtuvo un 94.3% de precisión identificando sentimientos, de este porcentaje solo el 28.6% fue correctamente clasificado en comparación con los sentimientos encontrados manualmente.

Para probar el prototipo propuesto en el uso práctico se realizó una prueba sobre 2,000 reseñas relacionadas a cámaras, también de páginas de internet. Del total de las reseñas, solo 1000 contenían opiniones, ya sean favorables o desfavorables. Con este conjunto de datos obtuvieron un 94.5% de precisión al identificar correctamente el sentimiento expresado en 241 críticas de un total de 255 en las que se encontró un sentimiento expresado. Esto implica un total de 24% de eficiencia del total de reseñas. Algunas de las limitantes identificadas por los autores para esta propuesta fue el tamaño del diccionario empleado además de la necesidad de mejorar el algoritmo para tratar expresiones más complejas.

Minqing Hu y Bing Liu (2004) [3], expusieron una propuesta para minar y resumir reseñas de consumidores. Los objetivos de este trabajo fueron encontrar las características a las cuales se hacían referencia en las críticas, identificar los enunciados que expresaban opiniones y polaridad sobre las mismas y resumir los resultados. Se propuso la creación de una pequeña lista de adjetivos “semilla” etiquetados manualmente dependiendo si expresan sentimiento positivo o negativo. Posteriormente esta lista es aumentada usando *WordNet*. Para la detección de los aspectos se emplearon características de etiquetado de las partes del enunciado (*Part of speech tagging*). Se identifican las características frecuentes, aunque solo se analizan las que se presentan de manera explícita en los enunciados. A continuación, se realiza una extracción de palabras que expresan opinión, se tiene preferencia por los adjetivos cercanos a los aspectos para tener enunciados de opinión. La identificación de la orientación de estas palabras se realiza mediante un análisis de sinónimos y antónimos. Se analizaron 100 reseñas de usuarios de la página de *Amazon*, estas críticas fueron obtenidas y clasificadas de manera manual. La propuesta tiene algunas limitantes, entre las

más importantes es el análisis solo de aspectos explícitos, se ignora la posibilidad del análisis de pronombres para identificar aspectos no mencionados. En la sub-tarea de la detección de la polaridad, únicamente se toman en cuenta los adjetivos como posibles palabras que expresan sentimientos y se propone a futuro poder trabajar con verbos, ya que también expresan opiniones. Los resultados obtenidos en la extracción de enunciados que expresan opiniones son de un 69% de precisión mientras que se obtuvo un 84% de precisión en la identificación de la polaridad.

Kim y Hovy (2006), presentaron un método para identificar una opinión con su autor y el tema en enunciados obtenidos de noticias en línea y medios sociales [4]. Este método usa un etiquetado de roles semánticos como un paso intermedio para etiquetar un autor de opinión y el tema usando datos de *FrameNet*. La tarea se divide en tres fases: identificar palabras que producen una opinión, etiquetar el rol semántico relacionados para cada palabra en el enunciado y encontrar el autor y tema de la palabra de opinión entre los roles semánticos etiquetados. Se realizaron pruebas con un conjunto de datos etiquetados manualmente por dos humanos. En la tarea de identificar palabras que producen opiniones se obtuvo 64% y 55% de precisión para los resultados propuestos por los humanos 1 y 2 respectivamente. 64.7% y 55.8% en la sub-tarea de la detección del tema y por último 47.9% y 36.6% al obtener el autor de la opinión, cada valor con respecto a los resultados propuestos por los humanos.

Ganu, Elhadad y Marian (2009) [5], publicaron una propuesta que se enfoca en identificar la estructura y el sentimiento mencionado en el contenido de reseñas de restaurantes, esto para mejorar la experiencia de usuario al acceder a las reseñas. Se trabajó con un corpus de reseñas sobre restaurantes que incluye información adicional como una puntuación en un rango de 5 estrellas y la fecha de la crítica. La hipótesis de este trabajo es que el texto de una crítica es un mejor indicador de sentimiento que la puntuación en un rango de estrellas. Se proponen dos alternativas de puntuación que incorporan información basada en el texto: una puntuación que se basa solo en la información de los sentimientos expresados y

otra que incorpora temas y sentimientos en una puntuación basada en regresión. Esto último motivado por la hipótesis de que diferentes temas tienen una importancia diferente en un escenario de recomendaciones. Los resultados obtenidos muestran que usar la información textual para predecir la puntuación de una reseña da mejores resultados que utilizar la puntuación de estrellas proporcionada por los usuarios.

En 2011, Mikalai Tsytsarau y Themis Palpanas realizaron un estudio sobre el Análisis de la Subjetividad en la Web [6], este abarca tres aspectos importantes como son la Minería de opiniones, la Agregación de opiniones y el Análisis de Contradicción. Su propuesta se divide en tres partes, identificar los temas mencionados en los datos de entrada y asociar cada tema con sus sentencias de opinión. Posteriormente, se clasifican las opiniones en un binario, ya sea positivo o negativo. Finalmente, la agregación de opinión para determinar la opinión total de la comunidad sobre un producto específico más que la opinión de un solo usuario sobre un producto.

Se hace uso de un enfoque de Aprendizaje Automático (Machine Learning) para solucionar el problema de clasificación, este se describe en dos pasos principales: 1) aprendizaje del modelo con un corpus de datos de entrenamiento y 2) clasificar los datos basados en el modelo entrenado.

En este estudio se hace mención de tres enfoques para la detección de la polaridad: el enfoque de diccionario, el estadístico y el semántico.

En el enfoque de Diccionario se utilizan diccionarios con puntajes de polaridad definidos, se mencionan algunos como *General Inquirer, Dictionary of Affect of Language, WordNet-Affect* y *SentiWordNet*, siendo este último el más popular. El problema más usual de este enfoque es que no siempre es fiable, ya que los diccionarios no pueden ser ajustados a conjuntos de datos particulares. Es decir, usualmente no se pueden adaptar los valores de polaridad a contextos particulares.

El enfoque estadístico busca superar los inconvenientes anteriores. Se propone la deducción de polaridades mediante la co-ocurrencia de adjetivos en un corpus. Esto permite la creación de un corpus específico y se identifica la polaridad analizando frecuencias.

El enfoque semántico provee valores de sentimiento directamente, se utiliza el cálculo de la similitud entre palabras. El principio de este enfoque es que palabras semánticamente similares deben recibir valores de polaridad similares.

Aborda la agregación de opinión con el fin de poder identificar la opinión prevalente de un grupo de personas sobre un tema, además de seguir su evolución a través del tiempo.

Se realiza, también, un estudio sobre la calidad de las opiniones y el *spam*. Este es un proceso previo a la agregación de opiniones y aunque existen pocos estudios consiste en la identificación de *spammers*.

Finalmente el análisis de contradicción. Una contradicción puede ser definida con una forma de vinculación textual en la cual dos frases expresan información completamente diferente sobre el mismo tema. La agregación de opiniones puede producir que haya pérdida de algunos datos relevantes ignorando la diversidad de los datos. Para esto es importante el análisis de la contradicción, ya que permite encontrar cambios en las opiniones a través del tiempo y el espacio.

2.2 Propuestas presentadas en el SemEval 2014

A continuación se presentan algunas propuestas presentadas para la tarea del Análisis de Sentimientos basado en aspectos presentadas en el SemEval 2014.

Una de las propuestas más interesantes y que además ha obtenido los mejores resultados en esta tarea es la presentada por Mohammad, Zhu, Cherry y Kiritchenko [7], presentan técnicas como la creación de diccionarios para la detección de sentimientos, estos creados automáticamente utilizando fórmulas

que analizan la información mutua. Las palabras de negación son analizadas en un contexto diferente y se crearon diccionarios para estos casos. Adicional a estos se implementaron diccionarios sobre el dominio de laptops y restaurantes.

Para el problema de clasificación, se implementó una Máquina de Soporte Vectorial. Los enunciados primero son tokenizados para obtener el etiquetado gramatical. Algunas de las características consideradas fueron los unigramas (palabras individuales) y bigramas (secuencias de dos palabras). Se tomó en cuenta el número de palabras positivas y negativas, la suma de todos los valores de sentimiento de cada palabra y el valor máximo de sentimiento. El mejor resultado obtenido fue en la tarea de la identificación de categorías. En esta tarea obtuvieron un 91% de precisión, 86% clasificado correctamente.

Otro trabajo que es importante destacar es el desarrollado por Pavel Blinov y Eugeny Kotelnikov [8], que proponen un método para el Análisis de Sentimientos basado en Aspectos para un conjunto de opiniones sobre laptops y restaurantes. El método propuesto para la extracción de aspectos consiste en dos pasos: la selección de candidatos y la extracción de términos. En el primer paso se analizan el número de palabras y la estructura morfológica. En base a esto se decide si procesar los términos de manera individual o en pareja de palabras. Cuando se trata de forma individual se manejan solo sustantivos singulares y plurales como posibles candidatos. Mientras que las conjunciones de Sustantivo-Sustantivo y Sustantivo-Sustantivo Plural también son candidatos debido a que también es muy común este tipo de aspectos. El segundo paso consiste en extraer los términos, una vez obtenida la lista de posibles candidatos, esto se realiza mediante una medida de similitud entre los candidatos y las categorías ya predefinidas. Si esta medida es mayor a un límite por cada categoría este término candidato es marcado como un aspecto.

A continuación se realiza la Detección de la Categoría, esto mediante la representación de cada palabra de manera vectorial y calculando la distancia entre todas las categorías y estos vectores eligiendo así la distancia mínima.

La Detección de la Polaridad se realiza de manera similar, con una representación vectorial. Inicialmente encontrando los términos candidatos, en este caso palabras que puedan representar sentimientos y posteriormente calculando la similitud con un conjunto de términos que expresan sentimientos creados manualmente.

Los resultados obtenidos por Pavel Blinov y Eugeny Kotelnikov para la extracción de aspectos es de un 52% de precisión para el corpus de laptops y un 71% para el de restaurantes, en tanto para las detección de categorías es de un 75%, esta tarea está solo disponible para los datos de restaurantes.

En la tarea de detección de polaridad obtuvieron un 52% de precisión para el conjunto de datos de laptops, mientras que para el de restaurantes un 63%.

En la investigación desarrollada por Schouten, Frasincar y De Jong, presentan un enfoque basado en co-ocurrencias para la detección de categorías y uno basado en diccionarios de sentimientos para la clasificación [9]. En la sub-tarea de la extracción de aspectos utilizaron un conjunto de entrenamiento para contar que tan frecuente una palabra aparecía con un aspecto, una medida simple de probabilidad fue computada para calcular la probabilidad de que esa palabra representara un aspecto.

Para encontrar las categorías, se construyó un algoritmo basado en co-ocurrencias. La idea central de este algoritmo es una matriz de co-ocurrencias entre palabras en los enunciados y las categorías ya definidas.

La clasificación de sentimientos se realizó mediante la creación de un diccionario de sentimientos. Para la creación de este diccionario se analizaron palabras que comúnmente aparecen cerca de palabras positivas o negativas.

El resultado obtenido para la detección de aspectos fue de un 83%, mientras que solo el 14% de estos fueron clasificados correctamente. De manera similar, para el corpus de restaurantes se obtuvo un 90% de precisión, mientras que 38% fue detectado correctamente. Para la sub-tarea de la clasificación de sentimientos se obtuvo un porcentaje de efectividad de 57% y 66% para las reseñas de laptops y restaurantes respectivamente.

Gupta y Ekbal [10], implementaron su propuesta para la tarea del SemEval de Análisis de Sentimientos basados en Aspectos. La propuesta comienza con un pre-procesamiento de los datos, primero cada reseña es tokenizada usando etiquetado gramatical (PoS tagging). La extracción de términos de aspecto se basó en un algoritmo supervisado. Se propuso la implementación de una Máquina de Soporte Vectorial. Para la extracción de términos se utilizaron características como: analizar palabras cercanas a la que es analizada, en este caso se analizaron las dos palabras previas y siguientes, Etiquetado gramatical, palabras cerradas, prefijos y sufijos, se analizó también el tamaño de las palabras. De un corpus de entrenamiento se extrajeron los aspectos frecuentes, si alguno de estos se encuentra en una reseña es tomado como candidato para ser aspecto.

Para la identificación de la polaridad se utilizaron algunas características mencionadas como el contexto local, etiquetado gramatical, prefijos, sufijos, etc. Además se utilizó un diccionario para obtener la puntuación de polaridad de las palabras. Los resultados obtenidos para los corpus de restaurantes y laptops fueron de un 67.37% y 67.07% de precisión respectivamente.

Sapna Negi y Paul Buitelaar [11], exponen un análisis de varias características sintácticas y léxicas para el análisis de sentimientos basado en aspectos. Se analizan cuatro conjuntos de características: primero las características no contextuales como son unigramas, bigramas, adjetivos y verbos. En este primer caso el sentimiento de cada aspecto es igual al sentimiento total de la oración. Se analizan también características no contextuales pero en este caso las polaridades de cada palabra son obtenidas de *SentiWordNet*. El tercer grupo de características son las contextuales, en estas se incluyen los adjetivos y verbos en las oraciones, los términos de aspecto, y las palabras que mantienen alguna relación con los aspectos. Finalmente, se analizan características Contextuales usando el diccionario de *SentiWordNet* para obtener la polaridad de las palabras. Para obtener la polaridad se realiza una desambiguación del sentido de cada palabra, es decir, se obtienen los diferentes sentidos que puede tener una palabra. *SentiWordNet* provee por *default* un valor positivo y uno negativo para cada

sentido de una palabra. Posteriormente se proponen dos formas de asignar la polaridad a cada aspecto, uno como el promedio de los valores de polaridad de cada palabra. Y la segunda tomando el promedio de los valores positivos y por separado el promedio de los valores negativos. El algoritmo muestra los mejores resultados utilizando unigramas y Máquinas de Soporte Vectorial. Los resultados obtenidos son de aproximadamente un 60% de precisión. Además de mencionar que se observó que las palabras cerradas y que eliminarlas por completo disminuye la precisión del clasificador.

La propuesta presentada por Patra, Mandal y su equipo [12] realiza una combinación de algunas características como son: Etiquetado Gramatical (*POS*), los términos de aspecto son básicamente representados por frases nominales. Se observó que los términos de aspecto están rodeados por un sustantivo o adjetivo. Otra característica importante es que usualmente los sustantivos que ocurren antes de verbos "*be*" en la mayoría de los casos son términos de aspecto. Se analizan también relaciones de dependencia en enunciados etiquetados. Se contaron los términos de aspecto en los datos de entrenamiento, los términos que ocurren más de cinco veces en el corpus son considerados para los datos de prueba.

Se usó *SentiWordNet* para identificar la polaridad de las palabras que expresan sentimientos. También fue usado *WordNet* para la identificación de la categoría del aspecto, se analizó un árbol de hiperónimos, palabras que refieren a una clase y pueden reemplazar una enumeración, para cada palabra.

Se encontró que muchas reseñas contienen más de una sentencia. Las reseñas son divididas en enunciados. En esta propuesta las palabras cerradas son excluidas.

Los mejores resultados obtenidos en la sub-tarea de la identificación de Términos de Aspectos son de 74.5% y 84% de precisión para laptops y restaurantes respectivamente. Para la detección de la Categoría de Aspecto, solo disponible para el corpus de Restaurantes, el máximo porcentaje de efectividad fue del 88%.

Finalmente en la clasificación de polaridad se obtuvo un 53% de precisión para laptops y 65% para restaurantes.

Otra propuesta para abordar esta tarea fue la implementada por Malhotra, Vij, Nandan y Dahlmeier [13]. Inicialmente para la tarea de Extracción de Términos de Aspecto se analizaron características como son: N-gramas de palabras, todos los unigramas, bigramas y trigramas de las reseñas. Se estudia la presencia o ausencia de letras mayúsculas. Etiquetado gramatical de cada palabra y sus cercanas. Dependencias y relaciones entre los aspectos, presencia/ausencia de adjetivos y adverbios. Y por último los signos de puntuación: *?, !.*

En la sub-tarea de la estimación de la polaridad las características usadas fueron: N-gramas de palabras, todas los unigramas, bigramas y trigramas de palabras en minúsculas. La polaridad de los adjetivos vecinos a los términos de aspecto, estos valores obtenidos del diccionario de *SentiWordNet*. El Etiquetado Gramatical de las tres palabras vecinas al término de aspecto también es estudiado. El mejor resultado obtenido al evaluar su propuesta fue de 79%, esto en la tarea de la detección de categorías en el corpus de Restaurantes, ubicándose en el séptimo lugar de veintiún participantes de esa tarea. Su mejor posición fue el quinto lugar de veinticinco participantes en la sub-tarea de la identificación de la polaridad de las categorías, este resultado sobre el corpus de Restaurantes.

En la investigación de Pekar, Azfal y Bohnet [14], para extraer los Términos que son Aspecto encuentran dependencias. Se identifican palabras que expresan sentimientos usando un diccionario. Los términos candidatos son extraídos, para estos son considerados los sustantivos, frases sustantivas, adjetivos y verbos. Se toman en cuenta algunas observaciones como: las palabras que expresan sentimientos no pueden ser parte de los términos, las frases sustantivas con todos los elementos en mayúsculas y acrónimos son excluidos, esto bajo la idea que hacen referencia a marcas y no a aspectos del producto. Tampoco se toman en cuenta los sustantivos que hacen referencia al producto en general (“laptop”, “restaurant”). Para cada término se analizan algunas características para su clasificación como son: unigramas, bigramas, adjetivos + término, sentimiento +

termino, Be + término. Para la detección de la polaridad se utilizaron diccionarios como *SentiWordNet* y *General Inquirer*. Los resultados obtenidos para la sub-tarea de la extracción de términos de aspecto fueron superiores a 90% en el corpus de restaurantes y superior a 88% en los datos sobre laptops. En tanto la detección de sentimientos se obtuvo un 76% y 63.6% de precisión en los datos de restaurantes y laptops respectivamente.

Por último en el trabajo desarrollado por Brychcín, Konkol y Steinberger [15] se utilizó un enfoque basado en Aprendizaje Automático usando el clasificador de Máxima Entropía. Las características analizadas por esta propuesta son: la ocurrencia de palabras en un texto, bigramas, se utiliza un diccionario de términos basado en los datos de entrenamiento, sufijos de dos a cuatro caracteres, diccionarios de sentimientos creados semi-automáticamente, *SentiWordNet* y *Cluster* de palabras. En la evaluación de su propuesta para SemEval 2014 se ubicaron, para la tarea de extracción de términos, en cuarto lugar con un porcentaje de precisión de 77.69% para el corpus de restaurantes y 66.67% para el conjunto de datos de laptops. En la sub-tarea de clasificación de categorías se ubicaron en el puesto 8 con 72.78% de precisión.

Capítulo 3. Marco teórico

En este capítulo se describen de manera general las herramientas y conceptos aplicados en la realización de esta propuesta.

3.1 Procesamiento del lenguaje natural

El procesamiento del lenguaje natural (PLN o NLP del inglés *Natural Language Processing*) estudia la interacción entre las computadoras y el lenguaje humano. Su objetivo es la investigación y formulación de mecanismos computacionales que permitan la comunicación entre personas y máquinas por medio de una lengua hablada o escrita por humano [16]. Estas lenguas son llamadas lenguajes naturales.

Por Procesamiento del Lenguaje Natural se entiende la habilidad de la máquina para procesar la información comunicada, no simplemente las letras o los sonidos del lenguaje [16]. El PLN es un fascinante campo de las ciencias de la computación que trata el lenguaje natural en sus diferentes formas. Algunas aplicaciones importantes relacionadas con PLN son: traducción, recuperación y extracción de información, reconocimiento y síntesis del habla, identificación y verificación del parlante, sistemas de diálogo, y sistemas de preguntas y respuestas [17].

Como lo menciona la Asociación Mexicana para el Procesamiento del Lenguaje Natural, lo que para nosotros es conocimiento, para las computadoras son archivos, secuencias de caracteres y nada más. Una computadora puede copiar tal archivo, respaldarlo, transmitirlo, borrarlo. Pero no puede buscar las respuestas a las preguntas en este texto, hacer inferencias lógicas, generalizar y resumirlo. Esto porque no lo puede entender. [18]

3.2 Análisis de sentimientos

El Análisis de Sentimientos, también conocido como Minería de Opiniones, es el campo de estudio que analiza la opinión de las personas, sentimientos, evaluaciones, actitudes y emociones en lenguajes escritos. El crecimiento en importancia del análisis de sentimientos coincide con el crecimiento de medios sociales como reseñas, foros de discusión, blogs, micro-blogs, Twitter y redes sociales. [19]

Actualmente el Análisis de Sentimientos es utilizado por empresas y personas de dominio social debido a la importancia de conocer la opinión sobre distintos productos, servicios o incluso personas y la influencia que tienen estas sobre otras personas. Puede ser utilizada para conocer las deficiencias de un producto, de igual manera para medir la aceptación de una persona como un candidato a algún puesto público o una persona pública en general.

3.3 Lenguajes y herramientas

En esta sección se describen de manera general las herramientas y lenguajes de programación utilizados en la implementación del modelo propuesto.

3.3.1 Lenguaje Python

Python[2] es un lenguaje de programación fácil de aprender. Tiene estructuras de datos de alto nivel eficientes y un simple pero efectivo enfoque a la programación orientada a objetos. La elegante sintaxis de Python y su escritura dinámica, junto con su naturaleza interpretada, lo hacen un lenguaje ideal para el scripting y el desarrollo rápido de aplicaciones en varias áreas y muchas plataformas.

[2] www.python.org

El intérprete de Python y la extensa biblioteca estándar están disponible gratuitamente en forma binaria y de código fuente para las principales plataformas desde el sitio web de Python y puede ser distribuido libremente. El intérprete de Python es fácilmente extendido con nuevas funciones y tipos de datos implementados en C o C++. Python también es adecuado como un lenguaje de extensión para aplicaciones personalizadas [22].

3.4 Etiquetado gramatical

El Etiquetado Gramatical (en inglés *POS tagging, Part of speech tagging*) consiste en asignar a una palabra una etiqueta indicando la parte gramatical que representa. Las etiquetas o partes gramaticales incluyen sustantivos, verbos, adjetivos, entre otros (Figura 3.1). El poder etiquetar una palabra es un proceso complejo que necesita del contexto en el cual está inmersa esta. Existen palabras que para un contexto pueden fungir como sustantivos y para otro como adjetivos o verbos.

Etiqueta	Descripción	Ejemplo
CC	Conjunción	And, or, but
DT	Determinante	The, a, these
JJ	Adjetivo	Nice, easy
JJR	Adjetivo Comparativo	Nicer, easier
JJS	Adjetivo Superlativo	Nicest, easiest
NN	Sustantivo, singular	Tiger, chair, dog
NNS	Sustantivo, plural	Tiger, chairs, dogs
NNP	Sustantivo, propio singular	Germany, God, Alice
NNPS	Sustantivo, propio plural	Christmases
RB	Adverbio	Extremely, loudly, hard
RBR	Adverbio, Comparativo	Better
RBS	Adverbio, Superlativo	Best
VB	Verbo	think

Figura 3.1 *Conjunto de Etiquetas Part of Speech (POS)*

Existen varios etiquetadores gramaticales, a continuación se mencionan los etiquetadores utilizados en el modelo.

3.4.1 CLiPS Pattern

CLiPS[3] (*Computational Linguistics & Psycholinguistics*) es un centro de investigación asociado con el Departamento de Lingüística de la facultad de Artes de la Universidad de Antwerp en Bélgica. El objetivo de CLiPS es producir investigaciones y recursos reconocidos internacionalmente en psicolingüística, lingüística y lingüística computacional. Además de investigar las combinaciones interdisciplinarias entre estas disciplinas.

Una de las herramientas proporcionadas por CLiPS es *Pattern*, este es un módulo para minería web implementado para el lenguaje de programación Python. Contiene herramientas para la minería de datos, procesamiento del lenguaje natural, aprendizaje automático así como para el análisis y visualización de redes.

Pattern está organizado en módulos separados que pueden ser utilizados conjuntamente. *Pattern.en* es un módulo que contiene un etiquetador para el inglés basado en expresiones regulares. Este identifica los elementos que constituyen una oración, como son sustantivos, verbos, etc. Emplea un etiquetador de partes gramaticales (*POS tagger*) de estados finitos. Tiene una precisión superior a 95%. Este módulo incluye una clase *Sentence* que provee de funciones para singularizar o pluralizar palabras, conjugación, modalidad y análisis de sentimientos. Además brinda una instancia de *WordNet* y *PyWordNet*.

Pattern está escrito en Python, esto significa que se sacrifica el rendimiento por la velocidad de desarrollo y la legibilidad. El código fuente se encuentra publicado para una licencia BSD, lo que permite ser incorporada en productos propietarios o usado en combinación con otros paquetes de código abierto [20].

[3] www.clips.ua.ac.be/

3.4.2 Stanford Log-linear Part-Of-Speech Tagger

Originalmente escrito por Kristina Toutanova del Departamento de Ciencias de la Computación de la Universidad de California. Desde entonces, varios colaboradores del Grupo de Procesamiento del Lenguaje Natural de Stanford han mejorado su velocidad, rendimiento, usabilidad y soporte para otros lenguajes. Este etiquetador presenta un rendimiento superior principalmente por el enriquecimiento de la información utilizada para el etiquetado. En particular se mejoraron los resultados incorporando características como el aumento extensivo del tratamiento de mayúsculas para palabras desconocidas. Características para desambiguar verbos, y características para la desambiguación de partículas de preposiciones y adverbios. El mejor resultado para el etiquetador es de 96.86% en promedio [21].

Inicialmente está desarrollado en Java, pero se puede utilizar en Python mediante la clase *StanfordPOSTagger* de NLTK.

3.4.3 NLTK Tagger

NLTK (*Natural Language Toolkit*) es una destacada plataforma para desarrollar programas en Python para trabajar con datos del lenguaje humano. Provee una interfaz sencilla de utilizar con más de 50 corpus y recursos léxicos como *WordNet*. Además de una serie de bibliotecas para el procesamiento de textos utilizadas para clasificar, tokenizar, derivar, etiquetar, etiquetado gramatical, y razonamiento semántico además de contar con un foro de discusión activo. NLKT ofrece funciones para el etiquetado gramatical de una palabra mediante la función *pos_tag()* dentro de la biblioteca NLTK [23].

3.5 Posting List

Un Posting List consiste en una lista de palabras con un conjunto de índices de documentos asociados. Dichos documentos son aquellos en los cuales este término ocurre. Estas estructuras son generalmente usadas para realizar búsquedas de términos que devuelven una lista de documentos relevantes con relación a estos.

Dada una colección de documentos, un Posting list consiste en una lista con el vocabulario incluido en todos estos documentos, exceptuando las palabras cerradas, con una lista de índices asociadas a cada palabra. Esta lista contiene el índice de todos los documentos en donde aparece esta palabra (Figura 3.2). El posting list es de gran utilidad para agilizar la búsqueda de términos en una colección de documentos.

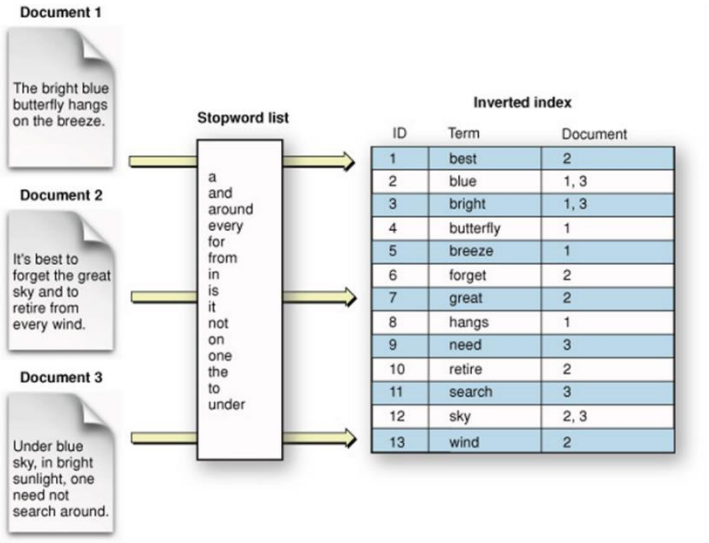

Figura 3.2 *Posting List*

3.6 Gensim – Word2Vec

Gensim es una biblioteca de Python diseñada para extraer automáticamente temas semánticos de documentos, de la manera más eficiente y fácil como sea posible. El fin de Gensim es el procesamiento de textos no estructurados (“texto plano”). Los algoritmos implementados en gensim, como son Análisis Semántico Latente (LSA, *Latent Semantic Analysis*), LDA *(Latent Dirichlet Allocation*) o Proyecciones aleatorias, descubren la estructura semántica de documentos, examinando patrones de co-ocurrencia dentro de un corpus de documentos de entrenamiento. Estos algoritmos son no supervisados, lo que significa que no es necesaria la intervención humana; sólo se necesita un corpus de documentos en texto plano [24].

Gensim proporciona la herramienta Word2Vec, esta provee una implementación eficiente de una continua bolsa-de-palabras y arquitecturas para el cómputo de la representación vectorial de las palabras. Estas representaciones pueden ser posteriormente utilizadas en muchas aplicaciones del procesamiento del lenguaje natural y futuras investigaciones. Word2Vec toma un corpus textual como entrada y produce vectores de palabra como salida. Primero construye un vocabulario del conjunto de datos de entrenamiento y entonces realiza la representación vectorial de las palabras. El archivo de vectores de palabra resultante puede ser usado como características en muchas aplicaciones del procesamiento del lenguaje natural y aprendizaje automático [25].

Word2Vec incluye funciones que permiten identificar el conjunto de palabras más similares a una palabra elegida por el usuario. Además de proporcionar un valor de similitud entre dos palabras específicas. Estas distancias calculadas de acuerdo al corpus de entrenamiento de entrada proporcionado [26].

3.7 Beautiful Soup

Beautiful Soup[4] es una biblioteca de Python diseñada para procesar datos de archivos en formato XML y HTML. Provee una manera de navegar, buscar y modificar la estructura de estos archivos. Esta biblioteca crea un árbol de todos los elementos del documento y puede ser utilizado para extraer información. Esta biblioteca es útil para extraer datos de la web [27].

Capítulo 4. Modelo propuesto

El modelo propuesto para resolver el problema del Análisis de Sentimientos basado en aspectos, se compone de las siguientes fases (Ver Figura 4.1).

1. Fase de Pre-procesamiento.
2. Fase de Identificación de Entidades y Atributos.
3. Fase de Identificación de Aspectos.
4. Fase de Clasificación de Aspectos en Entidades y Atributos.
5. Fase de Identificación de Polaridad.

A continuación se detallarán cada una de las fases mencionadas.

[4] www.crummy.com/software/BeautifulSoup

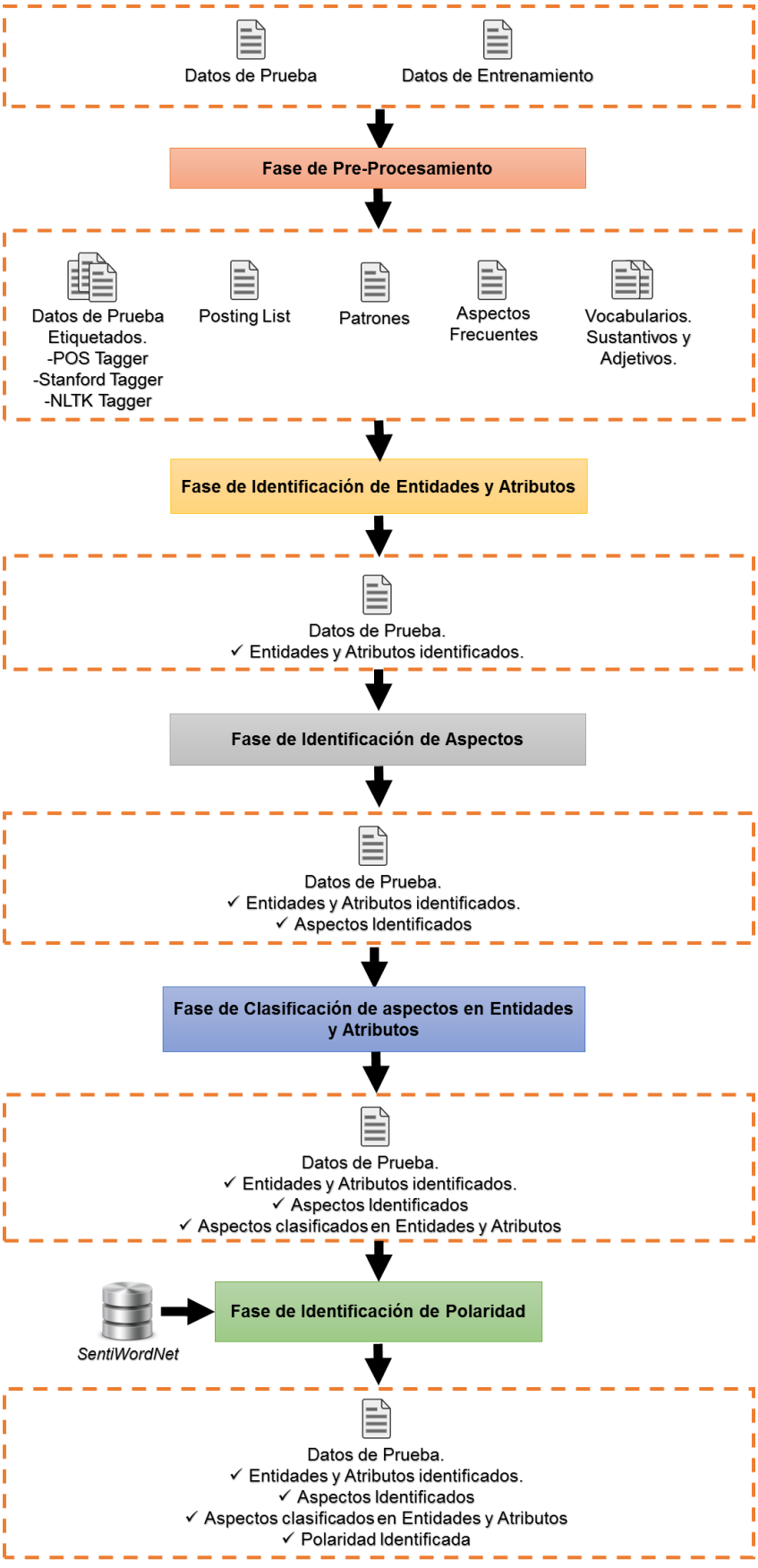

Figura 4.1 *Diagrama del modelo propuesto*

4.1 Fase de pre-procesamiento

Durante esta fase se realiza el análisis de los datos de entrenamiento proporcionados por el SemEval para la tarea, además de reunir los datos y generar los archivos de entrada necesarios para el modelo propuesto.

Los datos de entrenamiento proporcionados por el SemEval se encuentran en formato XML (Figura 4.2). Estos datos contienen el conjunto de reseñas con las categorías correctas, el objeto al cual se hace referencia en la reseña y la polaridad expresada hacia éste. Para el dominio de Laptops sólo se proporciona la categoría correcta identificada en la reseña.

```xml
<Review rid="1014458">
    <sentences>
        <sentence id="1014458:0">
            <text>I have eaten at Saul, many times, the food is always consistently, outrageously good.</text>
            <Opinions>
                <Opinion target="food" category="FOOD#QUALITY" polarity="positive" from="38" to="42"/>
            </Opinions>
        </sentence>
        <sentence id="1014458:1">
            <text>Saul is the best restaurant on Smith Street and in Brooklyn.</text>
            <Opinions>
                <Opinion target="Saul" category="RESTAURANT#GENERAL" polarity="positive" from="0" to="4"/>
            </Opinions>
        </sentence>
        <sentence id="1014458:2">
            <text>The duck confit is always amazing and the foie gras terrine with figs was out of this world.</text>
            <Opinions>
                <Opinion target="foie gras terrine with figs" category="FOOD#QUALITY" polarity="positive" from="42" to="69"/>
                <Opinion target="duck confit" category="FOOD#QUALITY" polarity="positive" from="4" to="15"/>
            </Opinions>
        </sentence>
        <sentence id="1014458:3">
            <text>The wine list is interesting and has many good values.</text>
            <Opinions>
                <Opinion target="wine list" category="DRINKS#STYLE_OPTIONS" polarity="positive" from="4" to="13"/>
                <Opinion target="wine list" category="DRINKS#PRICES" polarity="positive" from="4" to="13"/>
            </Opinions>
        </sentence>
```

Figura 4.2 *Datos de entrenamiento en Formato XML*

Del conjunto de datos de entrenamiento se obtienen varios elementos de entrada. Primero, se genera un diccionario con los aspectos que se encontraron en esas reseñas (Figura 4.3).

Restaurantes	Laptops
Food	Screen
Service	Battery
Place	Keyboard
Restaurant	Mouse
Staff	Programs
Pizza	Software
Athmosphere	Windows
Decor	Hard Drive
Ambience	System
Menu	Graphics
Fish	Processor

Figura 4.3 *Ejemplos de aspectos comunes en datos de entrenamiento*

Después, cada reseña del conjunto de datos de entrenamiento es tratada mediante la herramienta Clips Pattern para obtener su etiqueta POS (*Part of Speech*). Se extraen los patrones de cada aspecto y para cada aspecto se obtienen sus elementos gramaticales adyacentes. Esto para generar un diccionario con los patrones que cumple un aspecto además de conocer que elementos gramaticales se encuentran generalmente junto a este (Figura 4.4).

POS Aspecto	Patrón	Ejemplo
NN	JJ + ***NN*** + **.** Adjetivo – **Sustantivo** – Fin de Sentencia	...really good ***sushi***. Good ***food***! ...a delicious ***meal***. ...always great ***service.***
NN	DT + **NN** + VB Determinante + **Sustantivo** + Verbo	The ***sushi*** was... The **price** was... The ***food*** is... This ***restaurant*** is...
NN-NNS	DT + **NN/NNS** + VB Determinante + **(Sustantivo Singular -- Sustantivo Plural)** + Verbo	...the ***cheese fries*** are... ... this ***gourmmet pizzas*** are...
J/NN	JJ + **JJ/NN** + . Adjetivo + (**Adjetivo-Sustantivo**) + Fin de Sentencia	...faboluous ***Italian Food***! Nice ***Mexican Restaurant***.
NN	JJS + ***NN*** + IN Adjetivo Superlativo + Sustantivo + Conjunción	...best ***sushi*** in... ... worst **restaurant** in...

Figura 4.4 *Ejemplo de patrones comunes en los datos de entrenamiento*

Una vez que se tienen el conjunto de aspectos de los datos de entrenamiento se genera un ***Crawler*** con estos términos para la generación de dos corpus con datos específicos relacionados a Laptops y Restaurantes. Este consiste en para cada aspecto encontrar documentos relacionados con el mismo en artículos de internet. Es decir, para cada aspecto son extraídos un conjunto de párrafos relacionados. Para su implementación se utilizaron las bibliotecas de BeautifulSoup, que permite acceder a páginas web y obtener su contenido.

Para el conjunto de datos de Restaurantes se obtuvieron 299 aspectos diferentes, con los cuales se obtuvieron un total de 44620 páginas de internet con datos relacionados con estos aspectos. Posteriormente se extrajo el texto de estas páginas, generando así, un documento con más de 1 millón de párrafos. Para los datos de Laptops se identificaron 259 aspectos diferentes, se generó una lista con un total de 35722 páginas con texto relacionado, para así obtener finalmente un archivo con aproximadamente 1 millón de párrafos relacionados con los aspectos identificados en reseñas de laptops. Finalmente se realiza una limpia de estos corpus eliminando las palabras cerradas del idioma inglés.

A continuación, teniendo estos corpus específicos se generó un Posting List para cada corpus. Esto consiste para cada palabra generar una lista indexada con el identificador del párrafo que la contiene. Por ejemplo, para el aspecto “Food” se genera una lista con el índice de los párrafos en los que aparece. Con esto, se generaron dos Posting List con 554356 y 463924 palabras para Restaurantes y Laptops respectivamente.

El conjunto de datos de entrenamiento se analizó para generar un diccionario con el conjunto de aspectos y sustantivos, estos asociados con el porcentaje de aparición asociado a cada entidad predefinida. Por ejemplo, para el aspecto “FOOD”, se analiza cuantas veces aparece asociada a las entidades predefinidas, es decir, el número de veces que está relacionada con la entidad RESTAURANT, FOOD, SERVICE, etc. Se devuelve el aspecto FOOD con la lista de entidades asociadas con el porcentaje en rango 0 a 1, donde 1 implica que aparece siempre asociado a esa entidad (Figura 4.5).

<Opinion target="place" category="RESTAURANT#GENERAL"

<Opinion target="place" category="AMBIENCE#GENERAL"

Cada aspecto en el conjunto de datos de entrenamiento es relacionado con la entidad identificada.

place

SERVICE|0.083850931677,
LOCATION|0.0124223602484,
FOOD|0.136645962733,
RESTAURANT|0.515527950311,
AMBIENCE|0.240683229814,
DRINKS|0.0108695652174

El valor de relación de cada aspecto con las entidades es la relación entre el número de veces que aparece relacionado con esa entidad entre el total de veces que aparece el aspecto.

Figura 4.5 *Porcentaje de relación del aspecto "place" con las entidades asociadas.*

De manera similar, se genera un diccionario con los adjetivos, adverbios y verbos incluidos en los datos de entrenamiento. En este caso con su relación con los atributos predefinidos. Finalmente, como entrada para la siguiente fase (Figura 4.6), también se genera el etiquetado gramatical del conjunto de datos de prueba (Data test) con los tres etiquetadores propuestos (Stanford Parser, Clips Pattern y Nltk).

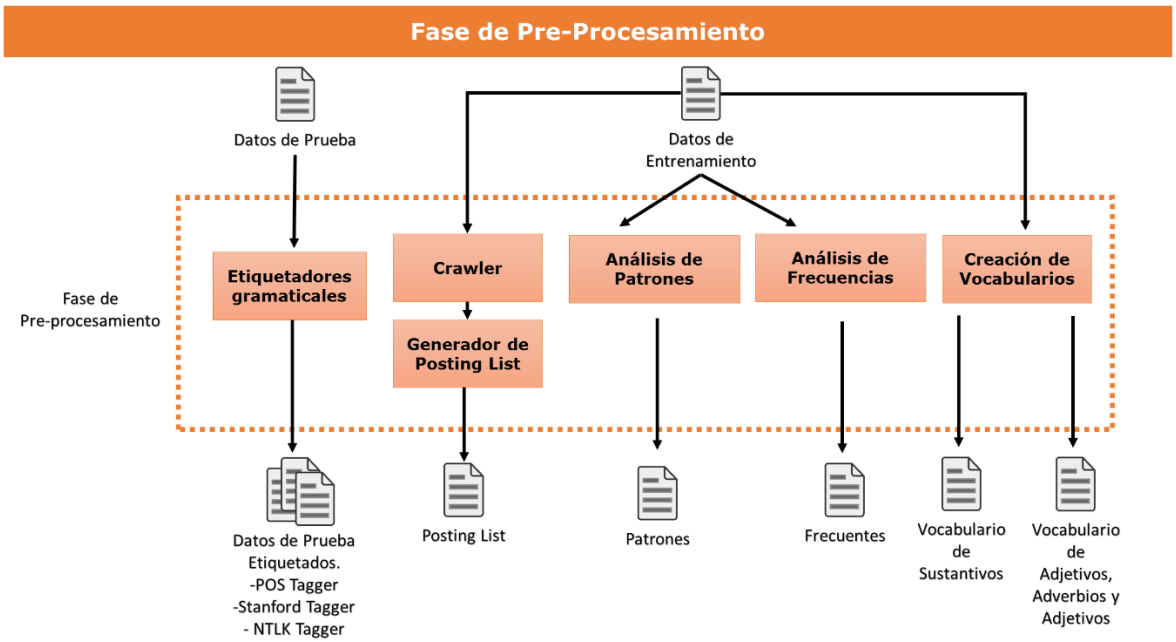

Figura 4.6 *Fase de Pre-procesamiento*

4.2 Fase de identificación de Entidades y Atributos

Esta fase consiste en identificar las entidades y atributos mencionados en las diferentes reseñas. Para resolver esta fase, cada reseña es dividida en frases. Esto mediante la identificación de elementos que fungen como conjunciones, disyunciones y elementos gramaticales como los puntos y comas. Esto utilizando las reseñas etiquetadas mediante la herramienta de Clips Pattern. Posteriormente, para cada frase encontrada, se extraen el conjunto de sustantivos y estos son buscados en el vocabulario generado en la fase anterior. Cada sustantivo tiene asociado un conjunto de entidades con un valor entre 0 y 1. Estos, para cada entidad, son acumulados en un diccionario, y finalmente para cada frase la entidad con un mayor valor es devuelta como la entidad identificada. De igual manera, para identificación del atributo mencionado se realiza el mismo proceso, pero analizando los adjetivos, adverbios y verbos. Finalmente, la salida de esta fase son las tuplas ENTIDAD#ATRIBUTO identificados para las reseñas (Figura 4.7).

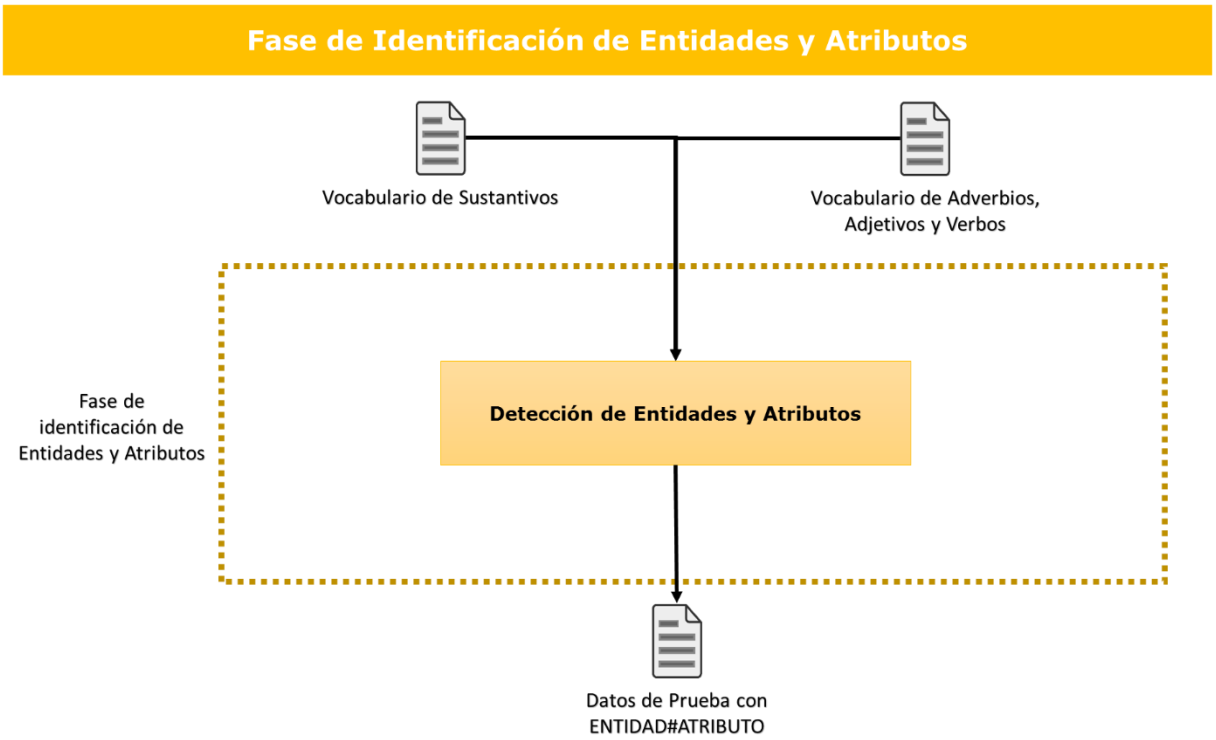

Figura 4.7 *Fase de Identificación de Entidades y Atributos*

4.3 Fase de identificación de aspectos

En esta fase se realiza la detección de los aspectos mencionados en cada reseña (Figura 4.8). Primero, para cada crítica, se extraen los sustantivos o secuencia de sustantivos y son agregados a un diccionario de candidatos a aspectos indicado con un peso que cumple con esta característica. Posteriormente, bajo la hipótesis de que dos palabras que se encuentran en el mismo párrafo están relacionadas, se realiza una búsqueda de cada uno de los elementos que se encuentran en el diccionario de candidatos en el Posting List generado en la fase anterior. Una vez realizada la búsqueda, cada candidato tiene un conjunto de párrafos en los que se encuentra.

A continuación se realiza una intersección de cada uno de los candidatos con el resto, y aquellos que se encuentran relacionados con al menos la mitad de candidatos más uno se les aumenta un valor en su peso, para así indicar que cumplen con esta segunda característica.

Otra característica más es el análisis de similitud semántica mediante la biblioteca gensim y el módulo Word2Vec. Esta herramienta necesita un corpus de tamaño considerable para su funcionamiento. El corpus proporcionado a la herramienta fue el corpus generado mediante la implementación del Crawler en la fase anterior. De manera similar a lo realizado con el Posting List, cada candidato es comparado con el resto de elementos y aquellos candidatos que estén relacionados con los demás se aumentan su valor de peso. Además cada candidato es comparado también con el conjunto de entidades predefinidas por el SemEval. Cuando la medida de similitud encontrada por el modelo generado por Word2Vec entre el candidato y alguna entidad, el valor de peso del candidato es aumentado, esto ya que si es muy similar a una entidad, lo más probable es que sea un aspecto.

La siguiente característica analizada son los patrones encontrados en la primera fase. Estos patrones están conformados por las secuencias de etiquetas de POS

(*Part of Speech*) de los aspectos en los datos de entrenamiento. Para obtener estos patrones cada reseña de los datos de entrenamiento es procesada mediante la herramienta de CLiPS para obtener la etiqueta *POS* (*Part of Speech*) de cada palabra. Posteriormente se forman los patrones, estas son las secuencias de la forma *POS_Izquierda* + *POS_Aspecto* + *POS_Derecha*. Donde *POS_Aspecto* es la etiqueta o secuencia de etiquetas gramaticales del aspecto, POS_Izquierda y POS_Derecha son las etiquetas gramaticales de la palabra izquierda y derecha al aspecto. Para cada reseña, se realiza la búsqueda de la posible n-grama que cumplan con el patrón de aspecto detectado. Una vez identificados estos patrones, también son estudiados sus elementos adyacentes izquierdo y derecho. Si cumplen con alguno de los patrones identificados en la fase de pre-procesamiento son agregados al diccionario de candidatos, si este elemento ya se encuentra, su valor de peso es aumentado.

Finalmente, la última característica tomada en cuenta es la búsqueda de los candidatos encontrados hasta este momento en la lista de aspectos del conjunto de datos de entrenamiento. Si es encontrado el candidato en esta lista se aumenta su valor de peso.

Ya realizado el análisis de las características mencionadas (Identificación de sustantivos, análisis de contexto, identificación de patrones, identificación de aspectos frecuentes), el criterio de selección de los candidatos es que cumplan con tener un peso mayor o igual a 3, es decir, cumplen con 3 o más características de las mencionadas anteriormente.

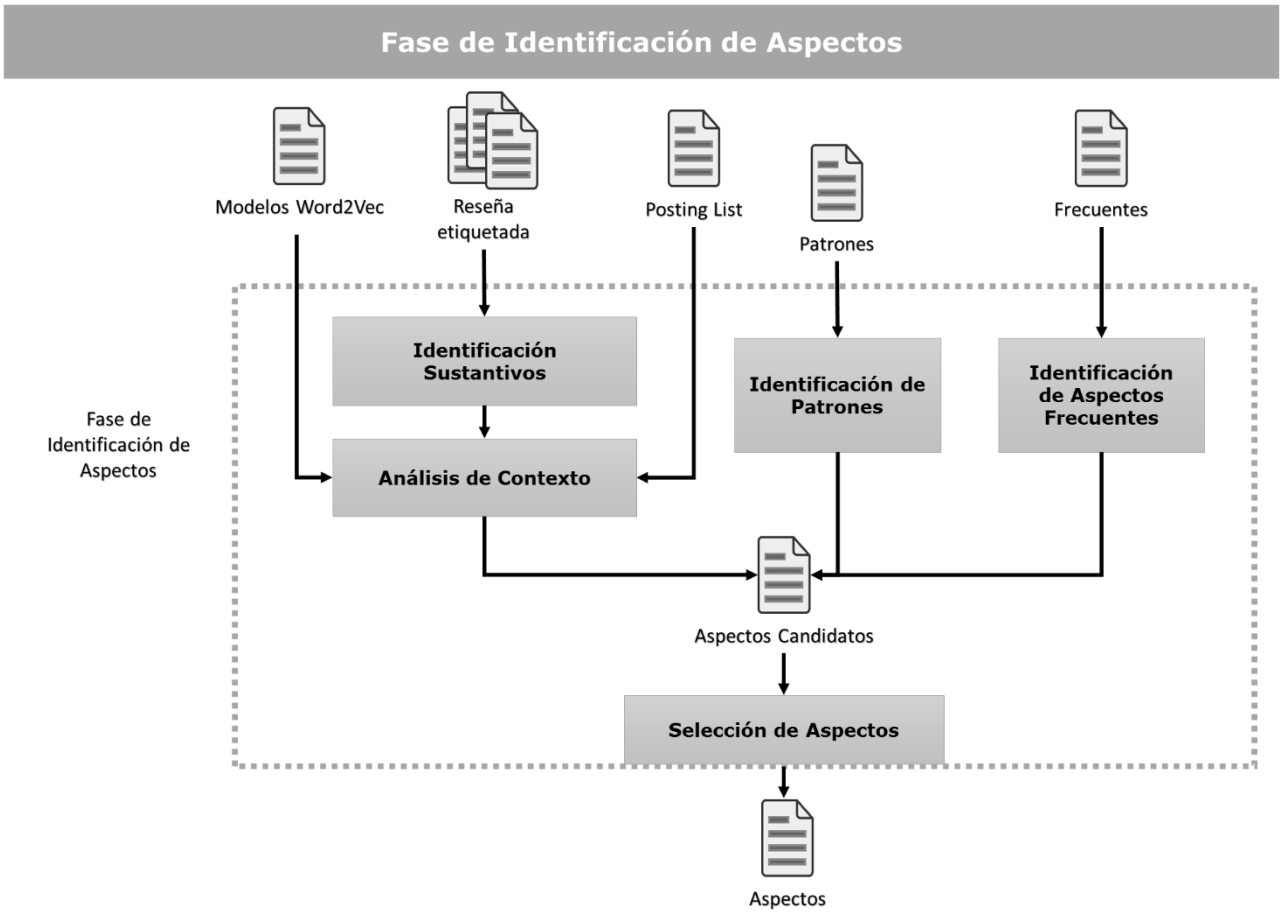

Figura 4.8 *Fase de Identificación de Aspectos*

4.4 Fase de clasificación de aspectos en Entidades y Atributos

En esta fase, para cada candidato identificado, se clasifica dentro de las Entidades predefinidas para Restaurantes y Laptops (tabla 4.1). Similar a la fase anterior donde se realiza identificación de las Entidades y Atributos mencionados, cada reseña es dividida en frases. La frase donde aparece el aspecto identificado es procesada. Primero en caso de que el aspecto detectado esté conformado por una sola palabra, ésta es buscada en el vocabulario de sustantivos, en caso de que esté conformada por más de una palabra, cada palabra es buscada en el vocabulario de sustantivos para obtener el porcentaje en el que cada sustantivo está relacionado con las entidades predefinidas. Estos valores son sumados y posteriormente la entidad con un mayor valor es elegida como la entidad a la cual

hace referencia el aspecto. También se cuenta con una lista de atributos predefinidos por el SemEval para Restaurantes y Laptops (Tabla 4.1). Para la frase en la cual se detectó el aspecto, cada elemento que sea adverbio, adjetivo o verbo es buscado en el vocabulario generado, esto para obtener el porcentaje de relación con el conjunto de atributos. Estos porcentajes son acumulados y el valor más alto es elegido como el atributo mencionado (Figura 4.9).

Restaurantes		Laptops	
Entidades	**Atributos**	**Entidades**	**Atributos**
- FOOD - DRINKS - SERVICE - AMBIENCE - LOCATION - RESTURANT	- GENERAL - QUALITY - STYLE & OPTIONS	- LAPTOP - DISPLAY - CPU - MOTHERBOARD - HARD DISC - MEMORY - BATTERY - POWER SUPPLY - KEYBOARD - MOUSE - FANS & COOLING - OPTICAL DRIVERS - PORTS - MULTIMEDIA DEVICES - HARDWARE - OS - SOFTWARE - WARRANTY - SHIPPING - SUPPORT - COMPANY	- GENERAL - QUALITY - OPERATION PERFORMANCE - USABILITY - DESIGN & FEATURES - PORTABILITY - CONNECTIVITY - MISCELLANEOUS

Tabla 4.1 *Entidades y Atributos definidos por el SemEval.*

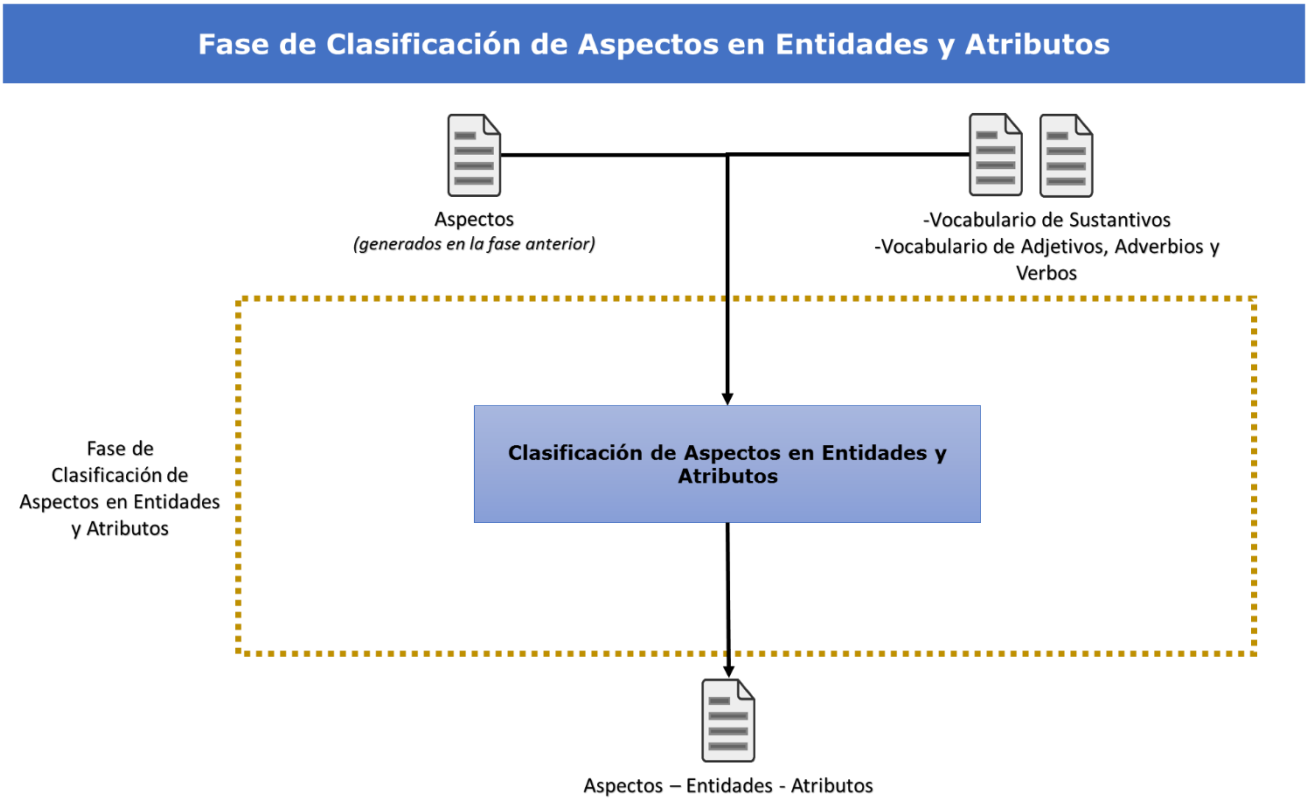

Figura 4.9 *Fase de Clasificación de Aspectos en Entidades y Atributos*

4.5 Fase de identificación de la polaridad.

Teniendo ya los candidatos propuestos, las entidades y atributos identificadas. La última parte consiste en identificar la polaridad del sentimiento expresado sobre cada aspecto. Para esto, se utilizó un enfoque basado en diccionario, para tener el valor de sentimiento para cada palabra. En esta propuesta se utilizó un diccionario creado en base a los datos de entrenamiento. Se creó mediante el procesamiento de adjetivos, adverbios y verbos en cada enunciado. Los datos de entrenamiento se encuentran etiquetados con su polaridad correcta, por tanto, para cada elemento encontrado en la sentencia se le da un valor de sentimiento en un rango de -1 a 1, donde -1 es muy negativo y 1 es muy positivo. Para el cálculo de este valor se realiza la división del número de ocasiones que aparece el elemento para cada polaridad, positiva, negativa y neutra. De estos resultados, el

que tenga mayor valor es elegido como la polaridad predominante de esta palabra y el resultado de la división es asignado como valor de sentimiento. Para aquellas palabras que no se encuentren en el conjunto de palabras encontradas en los datos de entrenamiento se utilizó el diccionario SentiWordNet que proporciona un valor numérico de sentimiento para cada palabra.

Posteriormente para detectar la polaridad de cada aspecto encontrado, se realiza el promedio de las polaridades de las palabras de la frase donde el aspecto se encuentra. Se analizan palabras que invierten el valor de polaridad como son “NOT”. Cuando este tipo de palabras aparecen, el valor de polaridad de las siguientes palabras de la sentencia es invertido. Palabras como “TOO”, “VERY” entre otras también causan un efecto en las palabras, estas aumentan el valor de polaridad de los siguientes elementos de la sentencia. Al finalizar, si el valor promedio encontrado es positivo y mayor a un rango establecido, la sentencia es clasificada como positiva. De lo contrario si es negativo y menor al rango es clasificada como negativa. Si el valor promedio es igual a cero o si está dentro del rango establecido es marcada como neutra. El rango mencionado se establece de manera manual, en donde el valor de polaridad identificado es mínimo, lo que implica que el sentimiento expresado sobre un aspecto no es relevante para ser clasificada como positiva o negativa. En adición al promedio de las polaridades, se estudia la polaridad individual de las palabras adyacentes al aspecto identificado. Se analiza el valor de polaridad más alto además del total de palabras positivas, negativas y neutras. La salida generada por esta fase es el conjunto de aspectos, entidades y atributos con su polaridad identificada (Figura 4.10).

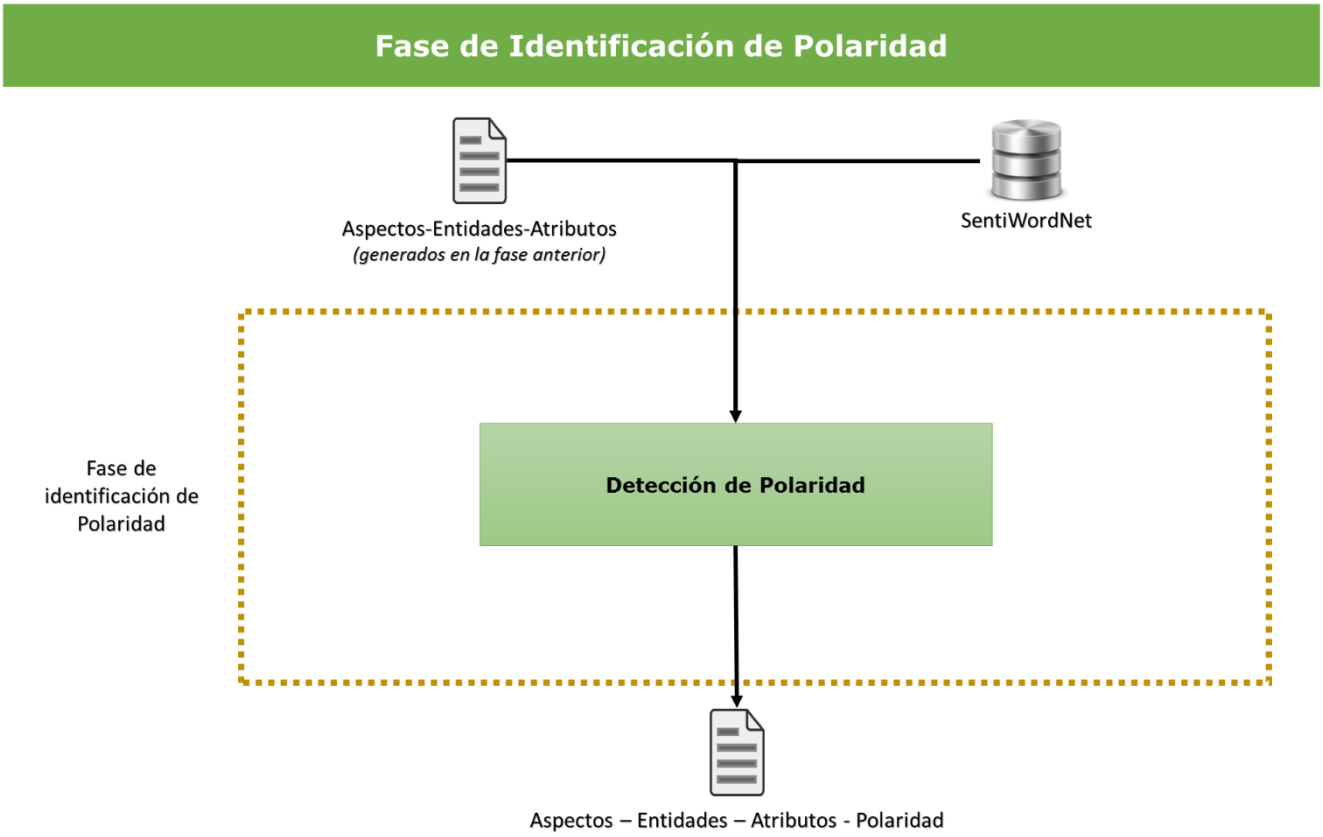

Figura 4.10 *Fase de Identificación de Polaridad*

Capítulo 5. Resultados obtenidos

A continuación, se analizan los resultados conseguidos por el modelo con las diferentes características implementadas sobre los dos dominios de datos.

5.1 Conjunto de datos

El modelo se aplicó a dos conjuntos de datos. El primer conjunto de datos es del dominio de Restaurantes, incluye un total de 3044 reseñas con 4724 aspectos mencionados en estas. Por su parte, el conjunto de datos sobre Laptops tiene un total de 3048 críticas de usuarios con un total de 3930 aspectos mencionados en estas (Tabla 5.1). Estos datos son proporcionados por el SemEval en formato XML. Ambos conjuntos de datos se encuentran en idioma Inglés.

	Conjunto de Datos.	
Dominio	**Restaurantes**	**Laptops**
Reseñas	**3044**	**3048**
Aspectos	**4724**	**3930**

Tabla 5.1 *Conjunto de Datos.*

5.2 Identificación de aspectos

Una vez implementadas las características mencionadas en el *capítulo 4*. A continuación se muestran los resultados obtenidos con las diferentes características propuestas en la tarea de la identificación de los aspectos mencionados en las reseñas. Primero, para el dominio de Restaurantes, los resultados obtenidos con la implementación sólo de la característica de selección de aspectos mediante la identificación de sustantivos fue de un 69.09% de precisión con respecto al total de aspectos proporcionados como correctos por el

SemEval y de un 40.48% de precisión con respecto al total de aspectos identificados por el modelo propuesto (Tabla 5.2).

RESTAURANTES	
Análisis de Sustantivos	
Aspectos Reales	4724
Aspectos Identificados por el Modelo	8063
Aspectos Correctamente Identificados	3264
Precisión sobre el total de Aspectos Reales (Aspectos Correctos / Aspectos Reales)	69.09%
Precisión sobre el total de Aspectos Identificados (Aspectos Correctos / Aspectos Identificados)	40.48%

Tabla 5.2 *Resultados obtenidos del Análisis de Sustantivos.*

Para el dominio de Laptops los porcentajes fueron de 47.53% sobre Aspectos Reales y 22.92% sobre los Aspectos Identificados (Tabla 5.3).

LAPTOPS	
Análisis de Sustantivos	
Aspectos Reales	3930
Aspectos Identificados por el Modelo	8150
Aspectos Correctamente Identificados	1868
Precisión sobre el total de Aspectos Reales (Aspectos Correctos / Aspectos Reales)	47.53%
Precisión sobre el total de Aspectos Identificados (Aspectos Correctos / Aspectos Identificados)	22.92%

Tabla 5.3 *Resultados obtenidos del Análisis de Sustantivos.*

La siguiente característica analizada es la implementación del Posting List para elegir los candidatos a aspectos que pertenecen al mismo contexto. Una vez implementada y evaluada esta característica junto con la anterior la precisión sobre el total de aspectos reales fue de un 62.51% y sobre el total de aspectos identificados fue de 41.53% (Tabla 5.4). Para el dominio de Laptops los porcentajes de precisión fueron de 43.07% y 23.46% (Tabla 5.5) sobre el total de aspectos reales e identificados respectivamente.

RESTAURANTES	
Análisis de Sustantivos Posting List	
Aspectos Reales	4724
Aspectos Identificados por el Modelo	7109
Aspectos Correctamente Identificados	2953
Precisión sobre el total de Aspectos Reales (Aspectos Correctos / Aspectos Reales)	62.51%
Precisión sobre el total de Aspectos Identificados (Aspectos Correctos / Aspectos Identificados)	41.53%

Tabla 5.4 *Resultados obtenidos del Análisis de Sustantivos junto con el Posting List.*

LAPTOPS	
Análisis de Sustantivos Posting List	
Aspectos Reales	3930
Aspectos Identificados por el Modelo	7215
Aspectos Correctamente Identificados	1693
Precisión sobre el total de Aspectos Reales (Aspectos Correctos / Aspectos Reales)	43.07%
Precisión sobre el total de Aspectos Identificados (Aspectos Correctos / Aspectos Identificados)	23.46%

Tabla 5.5 *Resultados obtenidos del Análisis de Sustantivos junto con el Posting List.*

Posteriormente, se incluyó el análisis de la similitud entre los candidatos a aspectos. Cada candidato identificado es comparado con el resto, aquellos que se encuentren relacionados son aceptados, su valor de peso es aumentado. De igual manera, cada aspecto es comparado con las entidades propuestas por el SemEval y si tiene una similitud alta con alguno, su valor de peso es aumentado. Los resultados una vez implementada esta característica fueron de 66.78% y 45.8% de precisión para Restaurantes (Tabla 5.6) y Laptops (Tabla 5.7).

RESTAURANTES	
Análisis de Sustantivos Posting List Similitud Word2Vec	
Aspectos Reales	4724
Aspectos Identificados por el Modelo	7580
Aspectos Correctamente Identificados	3155
Precisión sobre el total de Aspectos Reales (Aspectos Correctos / Aspectos Reales)	66.78%
Precisión sobre el total de Aspectos Identificados (Aspectos Correctos / Aspectos Identificados)	41.62%

Tabla 5.6 *Resultados obtenidos del Análisis de Sustantivos, Posting List y Similitud Word2Vec.*

LAPTOPS	
Análisis de Sustantivos Posting List Similitud Word2Vec	
Aspectos Reales	3930
Aspectos Identificados por el Modelo	7631
Aspectos Correctamente Identificados	1800
Precisión sobre el total de Aspectos Reales (Aspectos Correctos / Aspectos Reales)	45.80%
Precisión sobre el total de Aspectos Identificados (Aspectos Correctos / Aspectos Identificados)	23.58%

Tabla 5.7 *Resultados obtenidos del Análisis de Sustantivos, Posting List y Similitud Word2Vec.*

Como se puede observar en las tablas, pese a que la precisión sobre el total de aspectos disminuye, la precisión sobre el total de aspectos identificados aumenta. Esto debido a que con la aplicación de las características se disminuye el número de aspectos identificados. La siguiente característica es la detección de patrones dentro de las reseñas. Los porcentajes obtenidos para restaurantes fueron de 61.60% y 43.07% (Tabla 5.8) sobre aspectos reales e identificados y para laptops fueron de 47.18% y 22.95% (Tabla 5.9).

RESTAURANTES	
Análisis de Sustantivos Posting List Similitud Word2Vec Identificación de Patrones	
Aspectos Reales	4724
Aspectos Identificados por el Modelo	6755
Aspectos Correctamente Identificados	2910
Precisión sobre el total de Aspectos Reales (Aspectos Correctos / Aspectos Reales)	61.60%
Precisión sobre el total de Aspectos Identificados (Aspectos Correctos / Aspectos Identificados)	43.07%

Tabla 5.8 *Resultados obtenidos del Análisis de Sustantivos, Posting List y Patrones.*

LAPTOPS	
Análisis de Sustantivos Posting List Similitud Word2Vec Identificación de Patrones	
Aspectos Reales	3930
Aspectos Identificados por el Modelo	5430
Aspectos Correctamente Identificados	1413
Precisión sobre el total de Aspectos Reales (Aspectos Correctos / Aspectos Reales)	47.18%
Precisión sobre el total de Aspectos Identificados (Aspectos Correctos / Aspectos Identificados)	22.95%

Tabla 5.9 *Resultados obtenidos del Análisis de Sustantivos, Posting List y Patrones.*

La precisión con respecto al total de aspectos identificados por el modelo nuevamente aumenta. Esto debido a que con la inclusión de características se realiza un filtrado de los candidatos y así se disminuyen los aspectos identificados incorrectamente.

La última característica fue la de los aspectos obtenidos en los datos proporcionados por el SemEval para laptops y restaurantes. **Esta es una característica importante, ya que bajo la hipótesis de que elementos que han sido identificados como aspectos antes con seguridad son buenos candidatos a ser nuevamente aspectos.** Los resultados al implementar esta característica en conjunto con todas las demás fue de 75% y72.18% (Tabla 5.10) sobre aspectos reales e identificados para restaurantes. Para laptops fueron de 68.47% y 67.42% (Tabla 5.11) con respecto a aspectos reales e identificados.

RESTAURANTES	
Análisis de Sustantivos Posting List Similitud Word2Vec Identificación de Patrones Aspectos Entrenamiento	
Aspectos Reales	4724
Aspectos Identificados por el Modelo	4908
Aspectos Correctamente Identificados	3543
Precisión sobre el total de Aspectos Reales (Aspectos Correctos / Aspectos Reales)	75%
Precisión sobre el total de Aspectos Identificados (Aspectos Correctos / Aspectos Identificados)	72.18%

Tabla 5.10 *Resultados obtenidos del Análisis de Sustantivos, Posting List,Patrones y Aspectos de Entrenamiento.*

LAPTOPS	
Análisis de Sustantivos Posting List Similitud Word2Vec Identificación de Patrones Aspectos Entrenamiento	
Aspectos Reales	3930
Aspectos Identificados por el Modelo	3991
Aspectos Correctamente Identificados	2691
Precisión sobre el total de Aspectos Reales (Aspectos Correctos / Aspectos Reales)	68.47%
Precisión sobre el total de Aspectos Identificados (Aspectos Correctos / Aspectos Identificados)	67.42%

Tabla 5.11 *Resultados obtenidos del Análisis de Sustantivos, Posting List y Patrones.*

Finalmente se obtienen los resultados más altos al implementar las cinco características juntas. En comparación con las propuestas presentadas en el SemEval 2014, para ambos dominios se obtuvieron resultados cercanos al promedio de los resultados en la tarea de identificación de aspectos (Ver Tablas 5.12 y 5.13).

Comparación de Resultados – SemEval 2014 (Identificación de aspectos).	
LAPTOPS	
Nuestra Propuesta	**68.4%**
Mejor Propuesta (Equipo IHS R&D)	84.7%
Promedio	68.9%

Tabla 5.12 *Comparación de resultados obtenidos en la Identificación de aspectos. Dominio: Laptops*

Comparación de Resultados – SemEval 2014 (Identificación de aspectos).	
RESTAURANTES	
Nuestra Propuesta	**75%**
Mejor Propuesta (Equipo COMMIT)	90.9%
Promedio	76.7%

Tabla 5.13 *Comparación de resultados obtenidos en la Identificación de aspectos. Dominio: Restaurantes*

5.3 Identificación de polaridad de Aspecto.

En la fase de la identificación de la polaridad expresada para los aspectos detectados por el modelo para los dominios trabajados fueron los siguientes. Para el dominio de restaurantes se obtuvo 53.73% de precisión sobre los aspectos identificados correctamente en la fase de detección de aspectos. Para el dominio de laptops, su precisión fue de 58.19% (Tabla 5.14). En comparación con los resultados obtenidos por los participantes del SemEval 2014, nuestros resultados se encuentran cercanos al promedio de las propuestas presentadas (Ver Tablas 5.15 y 5.16).

Identificación de Polaridad		
	Conjunto de Datos.	
Dominio	**Restaurantes**	**Laptops**
Precisión	**53.73%**	**59.19%.**

Tabla 5.14 *Resultados obtenidos en la Identificación de Polaridad.*

Comparación de Resultados – SemEval 2014 (Identificación de polaridad).	
LAPTOPS	
Nuestra Propuesta	**59.19%.**
Mejor Propuesta (NRC-Canada)	70.48%
Promedio	59.01%

Tabla 5.15 *Comparación de resultados obtenidos en la Identificación de polaridad. Dominio: Laptops*

Comparación de Resultados – SemEval 2014 (Identificación de polaridad).	
RESTAURANTES	
Nuestra Propuesta	**53.73%**
Mejor Propuesta (DCU)	80.9%
Promedio	68.15%

Tabla 5.16 *Comparación de resultados obtenidos en la Identificación de polaridad. Dominio: Restaurantes*

5.4 Participación en el SemEval 2016.

Una vez desarrollado el modelo, se generaron las salidas necesarias para participar en el Foro de Competición del SemEval 2016. Para la tarea de la identificación de aspectos, se obtuvo un 50.25% de precisión al extraer los aspectos de las reseñas. Este resultado supera a las líneas base implementadas por los organizadores de la competición (Tabla 5.17). En la tarea de la identificación de la polaridad nuestra propuesta obtuvo un porcentaje de efectividad de 60.88% para el dominio de restaurantes (Tabla 5.18) y 62.79% para el dominio de laptops (Tabla 5.19). Ambos resultados inferiores a las líneas base.

Comparación de Resultados – SemEval 2016 (Identificación de aspectos).	
RESTAURANTES	
Nuestra Propuesta	**50.25%**
Mejor Propuesta (Equipo NLANGP)	72.34%
Líneas Base	44.07%

Tabla 5.17 *Comparación de resultados obtenidos en la Identificación de aspectos. Dominio: Restaurantes*

Comparación de Resultados – SemEval 2016 (Identificación de polaridad).	
LAPTOPS	
Nuestra Propuesta	**62.797%**
Mejor Propuesta (IIT-T)	82.72%
Lineas Base	70.03%

Tabla 5.18 *Comparación de resultados obtenidos en la Identificación de polaridad. Dominio: Laptops*

Comparación de Resultados – SemEval 2016 (Identificación de polaridad).	
RESTAURANTES	
Nuestra Propuesta	**60.885%**
Mejor Propuesta (XRCE)	88.12%
Líneas Base	76.48%

Tabla 5.19 *Comparación de resultados obtenidos en la Identificación de polaridad. Dominio: Restaurantes*

Capítulo 6. Conclusiones

Evaluado el modelo propuesto para resolver la tarea del análisis de sentimientos basado en aspectos propuesto por el SemEval se llegó a las siguientes conclusiones. De acuerdo a los objetivos planteados al inicio de esta tesis se cumplieron satisfactoriamente con estos. Se estudiaron los artículos relacionados con la tarea propuesta. Además se diseñó e implementó un modelo para detectar los aspectos en las reseñas proporcionadas, clasificar estos aspectos en entidades predefinidas, con respecto a lo expresado en las reseñas se identificó el atributo sobre cada aspecto identificado y se detectó la polaridad del sentimiento expresado en la reseña para cada aspecto.

Con la implementación del Crawler con los aspectos de entrenamiento en la fase de pre-procesamiento se generaron dos corpus específicos, uno con información relacionada con restaurantes, y otro con datos sobre laptops. **Estos corpus son una aportación de gran utilidad para futuras tareas que necesiten trabajar bajo estos contextos o similares.**

Se generaron las salidas con el formato requerido para participar en la evaluación de la tarea en el marco del SemEval 2016.

Una vez analizados los resultados obtenidos por el modelo se llegaron a las siguientes conclusiones:

6. El modelo se comporta bien en la detección de aspectos, sería de gran utilidad generar un diccionario para mejorar la identificación de aspectos. Este diccionario podría incluir, por ejemplo, nombres de platillos para el caso de restaurantes o nombres de aplicaciones y componentes para el dominio de laptops. Esto debido a que los etiquetadores gramaticales son etiquetadores generales y en ocasiones existen palabras de los contextos (restaurantes y laptops) que no son etiquetados correctamente.
7. Dado que los diccionarios de sentimiento como el utilizado SentiWordNet son diccionarios generales, es necesario generar un diccionario de

sentimientos específico para cada contexto, ya que existen palabras como “hot” o “cold” que bajo el contexto general tienen una polaridad y bajo el contexto de comida y restaurantes tienen otro valor totalmente opuesto.

8. Analizando los resultados obtenidos al utilizar las funciones brindadas por CLiPS, se concluye que esta herramienta es de gran utilidad aunque mejora los resultados al apoyarle de otros etiquetadores como los utilizados en el modelo propuesto.
9. Se debe mejorar en la tarea de la detección de la polaridad. Analizar problemáticas como el uso del sarcasmo en las críticas que invierten la polaridad del sentimiento expresado y estudiar la manera de mejorar los resultados del modelo.
10. La inclusión de la herramienta Word2Vec fue de gran utilidad para mejorar los resultados, ya que proporciona una medida de similitud entre palabras basadas en un contexto dado. Esto mediante el análisis de un corpus especifico. Para mejorar los resultados se podría aumentar el tamaño del corpus de entrada.

Finalmente, se pretende continuar con el trabajo en este modelo para futuras participaciones en las tareas del SemEval. Como trabajo futuro se propone lo siguiente:

- La creación de un diccionario de platillos para el dominio de restaurantes y de aplicaciones y componentes para el dominio de laptops.
- La inclusión de los datos de entrenamiento y pruebas del SemEval 2015 y 2016 para mejorar los diccionarios para la identificación de atributos y entidades y la extracción de aspectos.
- Implementar nuevas características que permitan obtener mejores resultados. Estas pueden ser el uso de aprendizaje automático y similitud entre frases.

Bibliografía

[1] B. Pang, L. Lee, S. Vaithyanathan, “Thumbs up? Sentiment Classification using Machine Learnig Techniques,” in Proc. *EMNLP*., vol. 10, pp. 79-86.

[2] T. Nasukawa, J. Yi, “Sentiment analysis: Capturing favorability using natural language processing,” in Proc of the *K-CAP-03, 2nd International Conference on Knowledge Capture*, 2003, pp. 70-77

[3] M. Hu, B. Liu, “Mining and Summarizing Customer Reviews,” in Proc KDDD-2004 ACM, 2004, pp. 168-177.

[4] S. Kim, E. Hovy, “Extracting opinions, Opinion Holders, and Topics Expressed in Online News Media Text,” in Proc of the Workshop on Sentiment and Subjectivity, 2006, pp. 1-8

[5] G. Ganu, N. Elhadad, A. Marian, “Beyond the Stars: Improving Rating Prediction using Review Text Content,” in Proc of WebDB, 2009, pp. 1-6

[6] M. Tsytsarau, T. Palpanas, “Surver on mining subjective data on the web,” J. Data Mining and Knowledge Discovery, vol 24, 2012, pp. 478-514.

[7] M. Mohammad, X. Zhu, et al, “NRC-Canada-2014: Detecting Aspects and Sentiment in Customers Reviews,” in Proc *SemEval 2014 8th International Workshop on Semantic Evaluation*, Ireland, pp. 437-442.

[8] P. Blinov, E. Kotelnikov, “Blinov: Distributed Representations of Words for Aspect-Based Sentiment Analysis at SemEval 2014,” in *Proc SemEval 2014 8th International Workshop on Semantic Evaluation*, Ireland, pp. 140-144.

[9] K. Schouten, F. Fransincar, F. de Jong, “COMMIT-P1WP3: A Co-ocurrence Based Approach to Aspect-Level Sentiment Analysis,” in Proc SemEval 2014 8th International Workshop on Semantic Evaluation, Ireland,pp. 203-207.

[10] D. K. Gupta, A. Ekbal, “IITP: Supervised Machine Learning for Aspect based Sentiment Analysis,” in Proc SemEval 2014 8th International Workshop on Semantic Evaluation, Ireland, pp. 319-323.

[11] S. Negi, P. Buitelaar, “INSIGHT_Galway: Syntactic and Lexical Features for Aspect Based Sentiment Analysis,” in Proc SemEval 2014 8th International Workshop on Semantic Evaluation, Ireland, pp. 346-350.

[12] B. G. Patra, S. Mandal, et al, “JU_CSE: A Conditional Random Field (CRF) Based Approach to Aspect Based Sentiment Analysis,” in Proc SemEval 2014 8th International Workshop on Semantic Evaluation, Ireland, pp. 370-374.

[13] N. Malhotra, et al, “A Constrained and Supervised Approach for Aspect-Based Sentiment Analysis,” in Proc SemEval 2014 8th International Workshop on Semantic Evaluation, Ireland, pp. 517-521.

[14] V. Pekar, N. Azfal, B. Bohnet, “UBham: Lexical Resources and Dependency Parsing for Aspect-Based Sentiment Analysis,” in Proc SemEval 2014 8th International Workshop on Semantic Evaluation, Ireland, pp. 683-687

[15] T. Brychcín, M. Konkol, J. Steinberger, “UWB: Machine Learning Approach to Aspect-Based Sentiment Analysis,” in Proc *SemEval 2014 8th International Workshop on Semantic Evaluation*, Ireland, pp. 817-822

[16] A. Gelbukh, “Procesamiento de Lenguaje Natural y sus Aplicaciones,“ *Komputer Sapiens*, vol. 1, p. 6 – 11, Enero, 2010

[17] M. Gómez, J. Gutiérrez, “Estado del lArte,“ *Komputer Sapiens*, vol. 1, p. 4 – 5, Enero, 2010

[18] AMPLN, *¿Qué es el Procesamiento del Lenguaje Natural?* [online]. México: Asociación Mexicana para el Procesamiento del Lenguaje Natural, 2009, Disponible en: http://www.ampln.org/pmwiki.php?n=Main.PLN

[19] B. Lui, *Sentiment Analysis and Opinion Mining*, Canada: Morgan & Claypool, 2012

[20] T. De Smedt, W. Daelemans, “Pattern for Python,” *in Proc Journal of Machine Learning Researc*h, 2012, pp. 2063-2067

[21] K. Toutanova, C. Manning, “Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger,” in Proc of *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000)*, 2000, pp. 63-70.

[22] G. Van Rossum, *El tutorial de Python* [online]. 2009, Disponible en: http://docs.python.org.ar/tutorial/pdfs/TutorialPython2.pdf

[23] S. Bird, E. Klein, E. Loper, *Natural Language Processing with Python*, United States Of America: O’Reilly Media, 2009

[24] R. Rehurek, P. Sojka, “Software Framework for Topic Modelling with Large Corpora,” in Proc of the LREC 2010 Workshop on New Callenges for NLP Frameworks, 2010, pp. 45 – 50

[25] T. Mikolov, K. Chen, G. Corrado, J. Dean, “Efficient Estimation of Word Representations in Vector Space,” in Proc of Wordshop at ICLR, 2013

[26] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, “Distributed Representation of Words and Phrases and their Compositionatily,” in Proc of NIPS, 2013.

[27] L. Richardson. (2004). Beautiful Soup. [Online]. Recuperado el 3 de marzo de 2016 de: http://www.crummy.com/software/BeautifulSoup/