Instituto Politécnico Nacional

Unidad Profesional Interdisciplinaria en Ingeniería y Tecnologías Avanzadas

Proyecto Terminal

"Sistema de análisis lingüístico de comentarios para el apoyo del marketing boca a boca"

Que para obtener el título de

"Ingeniero en telemática"

Presenta:

Vanessa San Agustín Gutiérrez

Asesores

Dr. Alexander Gelbukh

M. en C. Miguel Alejandro Martínez Rosales

México D.F. Enero 2016.

Instituto Politécnico Nacional

Unidad Profesional Interdisciplinaria en Ingeniería y Tecnologías Avanzadas

Proyecto Terminal

"Sistema de análisis lingüístico de comentarios para el apoyo del marketing boca a boca"

Que para obtener el título de

"Ingeniero en telemática"

Presenta:

Vanessa San Agustín Gutiérrez

Asesores:

Dr. Alexander Gelbukh

Presidente del Jurado

M. en C. Miguel Alejandro Martínez Rosales

Secretario del Jurado

M. en C. Carlos de la Cruz Sosa

M. en C. Cyntia Eugenia Enríquez Ortíz

México D.F. Enero 2016.

Agradecimientos

Al Instituto Politécnico Nacional y la Unidad Profesional Interdisciplinaria en Ingeniería y Tecnologías Avanzadas que me han dado la oportunidad de poner
“La técnica al servicio de la patria”.

Agradezco también al Laboratorio de cómputo móvil que ha sido el espacio que me ha abierto las puertas de sus instalaciones para el desarrollo del sistema. Los profesores y compañeros que ahí trabajamos somos un gran equipo.

Gracias por el apoyo y confianza de todos ustedes.

Dedicatorias

A Dios porque es él quien ha guiado mi camino, me ha presentado oportunidades maravillosas y siempre está a mi lado.

A mis padres que con su paciencia y trabajo me han ayudado a ser quien soy. Mi padre Antonio San Agustín, que siempre me ha enseñado a lograr mis metas con trabajo y honestidad. Mi madre y amiga incondicional Piedad Gutiérrez, que siempre ha cuidado de mí con amor y me ha acompañado en todo momento enseñándome que nunca me debo rendir. Gracias a ambos por preocuparse por mí, por ser partícipes de este trabajo y darme lo más preciado en este mundo: la vida.

A mis hermanos que siempre se han preocupado por darme un buen ejemplo, sé que siempre podré contar con ellos.

A mis asesores que me han acompañado a lo largo de este proyecto, son un gran ejemplo a seguir, me han contagiado la curiosidad y el ansia de experimentar para descubrir cosas nuevas.

A todos mis profesores, que han formado parte de mi conocimiento y siempre me han motivado con su gran pasión por enseñar.

A todos ustedes GRACIAS.

Vanessa San Agustín Gutiérrez

Índice

RESUMEN ..... **IX**

ABSTRACT ..... **XI**

INTRODUCCIÓN ..... **XIII**

PLANTEAMIENTO DEL PROBLEMA XVI

PROPUESTA DE SOLUCIÓN XVII

ALCANCES XIX

JUSTIFICACIÓN XX

METODOLOGÍA XXIII

OBJETIVO GENERAL ..... **XXV**

OBJETIVOS PARTICULARES XXV

CAPÍTULO 1. MARCO TEÓRICO **1**

1.1. ANÁLISIS DE SENTIMIENTOS 1

1.2. LENGUAJE Y EL PROCESAMIENTO DEL LENGUAJE NATURAL 1

1.3. SEMÁNTICA 2

1.4. ONTOLOGÍAS Y EL CONOCIMIENTO 3

1.5. RDF 4

1.6. OWL 4

1.7. HERRAMIENTAS DEL PROCESAMIENTO DEL LENGUAJE NATURAL 5

1.7.1. Python *5*

1.7.2. Natural Language Toolkit (NLTK) *5*

CAPÍTULO 2. ESTADO DEL ARTE **7**

2.1. "GENERACIÓN AUTOMÁTICA DE RESÚMENES DE OPINIÓN SOBRE ENTIDADES NOMBRADAS EN ESPAÑOL" 7

2.2. "SISTEMA DE ANÁLISIS DEL IMPACTO DE PRODUCTOS A TRAVÉS DE LAS REDES SOCIALES FACEBOOK Y TWITTER DESDE LA PERSPECTIVA DEL USUARIO" 7

2.3. "ARQUITECTURA WEB PARA ANÁLISIS DE SENTIMIENTOS EN FACEBOOK CON ENFOQUE SEMÁNTICO" 8

CAPÍTULO 3. ANÁLISIS **11**

3.1. ANÁLISIS DE REQUERIMIENTOS 11

3.1.1. Propósito *11*

3.1.2. Ámbito del sistema *11*

3.2. DESCRIPCIÓN GENERAL 11
3.2.1. Perspectiva del producto 11
3.2.2. Funciones del producto 12
3.2.3. Características de los usuarios 12
3.2.4. Restricciones 13
3.2.5. Suposiciones y dependencias 13
3.3. REQUERIMIENTOS ESPECÍFICOS 14
3.3.1. Interfaces externas 14
3.3.1.1 Interfaz de hardware 14
3.3.1.2 Interfaz de software 14
3.3.1.3 Interfaz de comunicaciones 19
3.3.2. Funciones 20

CAPÍTULO 4. DISEÑO 23

4.1. DIAGRAMAS DE CASOS DE USO 23
4.2. DIAGRAMAS DE ACTIVIDADES 29
4.3. DIAGRAMAS DE SECUENCIA 33
4.4. DIAGRAMA DE CLASES 36
4.5. DIAGRAMAS PARA BASE DE DATOS 38
4.6. DIAGRAMA DE LA ONTOLOGÍA 39
4.7. INTERFAZ DE USUARIO 44

CAPÍTULO 5. IMPLEMENTACIÓN 47

5.1. EXTRACCIÓN AUTOMÁTICA DE ASPECTOS Y SENTIMIENTOS 47
5.2. ASIGNACIÓN DE CATEGORÍAS UTILIZANDO LA ONTOLOGÍA *CAR* 53
5.3. GENERACIÓN DE ESTADÍSTICAS 56
5.4. ARQUITECTURA GENERAL DEL SISTEMA 58

CAPÍTULO 6. PRUEBAS Y RESULTADOS 63

6.1. ORIGEN DE LOS DATOS PARA LAS PRUEBAS 63
6.2. FUNCIONAMIENTO DE LA PÁGINA DE INICIO DEL SISTEMA 64
6.3. GRÁFICAS GENERADAS PARA CADA ANÁLISIS 66
6.3.1. Gráficas generales 66
6.3.2. Gráfica de todas las partes del auto 66
6.3.3. Tablas de los accesorios 67
6.3.4. Gráfica del chasis 68
6.3.5. Gráfica del sistema eléctrico 68

6.3.6. Gráfica del exterior *69*
6.3.7. Gráfica del interior *69*
6.3.8. Gráfica del tren motor *70*
6.4. PRUEBA DE ANÁLISIS DE LOS COMENTARIOS 70
6.4.1. Resultados del análisis de los comentarios del Jetta 2006 ..... *71*
6.4.1.1 Resultados de los comentarios obtenidos de Cars.com 71
6.4.1.2 Resultados de los comentarios obtenidos de Edmunds.com 74
6.4.1.3 Resultados de los comentarios obtenidos de Kbb.com 76
6.4.2. Resultados del análisis de los comentarios del Spark 2013 ..... *78*
6.4.2.1 Resultados de los comentarios obtenidos de Cars.com.txt 78
6.4.2.2 Resultados de los comentarios obtenidos de Edmunds.com.txt 80
6.4.2.3 Resultados de los comentarios obtenidos de Kbb.com 82
CONCLUSIONES **85**
TRABAJOS A FUTURO **87**
REFERENCIAS **89**

Índice de figuras

FIGURA 1. PUBLICIDAD DE PACKARD MOTORS. XIII
FIGURA 2. INFOGRAFÍA. XV
FIGURA 3. IMAGEN QUE CIRCULÓ POR LAS REDES SOCIALES DE LA CAMISETA DE ZARA. XVI
FIGURA 4. ARQUITECTURA GENERAL DEL SISTEMA. XVII
FIGURA 5. DIAGRAMA DE CASOS DE USO DEL SISTEMA. 23
FIGURA 6. DIAGRAMA DE ACTIVIDADES DEL SISTEMA. 30
FIGURA 7. DIAGRAMA DE ACTIVIDADES DEL ANÁLISIS LINGÜÍSTICO. 31
FIGURA 8. DIAGRAMA DE ACTIVIDADES PARA LA ASIGNACIÓN DE CATEGORÍA. 32
FIGURA 9. DIAGRAMA DE ACTIVIDADES PARA MOSTRAR LAS GRÁFICAS. 33
FIGURA 10. DIAGRAMA DE SECUENCIA GENERAL. 34
FIGURA 11. DIAGRAMA DE SECUENCIA DEL ANÁLISIS LINGÜÍSTICO. 35
FIGURA 12. DIAGRAMA DE SECUENCIA PARA LA ASIGNACIÓN DE CATEGORÍAS. 35
FIGURA 13. DIAGRAMA DE SECUENCIA PARA GENERAR GRÁFICAS. 36
FIGURA 14. DIAGRAMA DE CLASES. 37
FIGURA 15. DIAGRAMA ENTIDAD-RELACIÓN DE LA BASE DE DATOS. 38
FIGURA 16. DIAGRAMA RELACIONAL DE LA BASE DE DATOS. 38
FIGURA 17. DISEÑO DE LA ONTOLOGÍA: JERARQUÍA DE CLASES. 39
FIGURA 18. DISEÑO DE LA ONTOLOGÍA. 40
FIGURA 19. DISEÑO DE LA ONTOLOGÍA PRIMER NIVEL. 41
FIGURA 20. DISEÑO DE LA ONTOLOGÍA SEGUNDO NIVEL: *CHASSIS*. 41
FIGURA 21. DISEÑO DE LA ONTOLOGÍA SEGUNDO NIVEL: *POWERTRAIN*. 42
FIGURA 22. DISEÑO DE LA ONTOLOGÍA SEGUNDO NIVEL: *INTERIOR*. 42
FIGURA 23. DISEÑO DE LA ONTOLOGÍA SEGUNDO NIVEL: *ELECTRICAL*. 43
FIGURA 24. DISEÑO DE LA ONTOLOGÍA SEGUNDO NIVEL: *ACCESSORIES*. 43
FIGURA 25. DISEÑO DE LA ONTOLOGÍA SEGUNDO NIVEL: *EXTERIOR*. 44
FIGURA 26. PÁGINA INICIAL DEL SISTEMA. 44
FIGURA 27. PÁGINA DE RESULTADOS SOBRE EL AUTOMÓVIL EN GENERAL. 45
FIGURA 28. PÁGINA DE RESULTADOS DE ASPECTOS AGRUPADOS. 45
FIGURA 29. DIAGRAMA DE FLUJO DEL ANÁLISIS GRAMATICAL. 52
FIGURA 30. PÁGINA DE RESULTADOS DE ASPECTOS AGRUPADOS. 55
FIGURA 31. RESULTADOS DE ASPECTOS AGRUPADOS. 58
FIGURA 32. ARQUITECTURA CLIENTE-SERVIDOR. 59
FIGURA 33. MODELO-VISTA-CONTROLADOR. 59
FIGURA 34. ORGANIZACIÓN DE LOS ARCHIVOS DEL PROYECTO. 61

FIGURA 35. ESTADÍSTICAS DE LAS QUEJAS DEL JETTA. 63
FIGURA 36. ESTADÍSTICAS DE LAS QUEJAS DEL SPARK. 64
FIGURA 37. PÁGINA INICIAL DEL SISTEMA (1). SELECCIÓN DE UN ARCHIVO (2). NOMBRE DEL ARCHIVO ESPERANDO SER ANALIZADO (3). 65
FIGURA 38. GRÁFICAS GENERALES 66
FIGURA 39. GRÁFICA DE LOS ASPECTOS AGRUPADOS EN CATEGORÍAS DE PRIMER NIVEL. 67
FIGURA 40. TABLAS DE LA CATEGORÍA *ACCESSORIES*. 67
FIGURA 41. GRÁFICA DE RESULTADOS DE LA CATEGORÍA *CHASSIS*. 68
FIGURA 42. GRÁFICA DE RESULTADOS DE LA CATEGORÍA *ELECTRICAL*. 68
FIGURA 43. GRÁFICA DE RESULTADOS DE LA CATEGORÍA *EXTERIOR*. 69
FIGURA 44. GRÁFICA DE RESULTADOS DE LA CATEGORÍA *INTERIOR*. 69
FIGURA 45. GRÁFICA DE RESULTADOS DE LA CATEGORÍA *POWERTRAIN*. 70
FIGURA 46. GRÁFICAS GENERALES DEL ARCHIVO JETTACARS.COM.TXT. 71
FIGURA 47. GRÁFICA DE PRIMER NIVEL DEL ARCHIVO JETTACARS.COM.TXT. 72
FIGURA 48. GRÁFICA DE *POWERTRAIN* DEL ARCHIVO JETTACARS.COM.TXT. 72
FIGURA 49. GRÁFICAS GENERALES DEL ARCHIVO JETTAEDMUNDS.COM.TXT. 74
FIGURA 50. GRÁFICA DE PRIMER NIVEL DEL ARCHIVO JETTAEDMUNDS.COM.TXT. 74
FIGURA 51. GRÁFICAS GENERALES DEL ARCHIVO JETTAKBB.COM.TXT. 76
FIGURA 52. GRÁFICA DE PRIMER NIVEL DEL ARCHIVO JETTAKBB.COM.TXT. 76
FIGURA 53. GRÁFICAS GENERALES DEL ARCHIVO SPARKCARS.COM.TXT. 78
FIGURA 54. GRÁFICAS DE PRIMER NIVEL DEL ARCHIVO SPARKCARS.COM.TXT. 78
FIGURA 55. GRÁFICAS GENERALES DEL ARCHIVO SPARKEDMUNDS.COM.TXT. 80
FIGURA 56. GRÁFICA DE PRIMER NIVEL DEL ARCHIVO SPARKEDMUNDS.COM.TXT. 80
FIGURA 57. GRÁFICAS GENERALES DEL ARCHIVO SPARKKBB.COM.TXT. 82
FIGURA 58. GRÁFICA DE PRIMER NIVEL DEL ARCHIVO SPARKKBB.COM.TXT. 82

Índice de tablas

TABLA 1. COMPARACIÓN DE TRABAJOS. 9
TABLA 2. COMPARACIÓN ENTRE ONTOLOGÍAS. 15
TABLA 3. COMPARACIÓN ENTRE EDITORES DE ONTOLOGÍAS. 17
TABLA 4. COMPARACIÓN ENTRE GESTORES DE BASES DE DATOS. 18
TABLA 5. COMPARACIÓN ENTRE BIBLIOTECAS PARA GENERAR GRÁFICAS. 20
TABLA 6. ESPECIFICACIÓN DEL CASO DE USO SELECCIONAR ARCHIVO. 24
TABLA 7. ESPECIFICACIÓN DEL CASO DE USO ANALIZAR COMENTARIOS. 24
TABLA 8. ESPECIFICACIÓN DEL CASO DE USO ALMACENAR ARCHIVO. 25
TABLA 9. ESPECIFICACIÓN DEL CASO DE USO PRE-PROCESAMIENTO. 25
TABLA 10. ESPECIFICACIÓN DEL CASO DE USO ANALIZAR GRAMÁTICA. 26
TABLA 11. ESPECIFICACIÓN DEL CASO DE USO EXTRAER ASPECTOS. 26
TABLA 12. ESPECIFICACIÓN DEL CASO DE USO EXTRAER SENTIMIENTOS. 27
TABLA 13. ESPECIFICACIÓN DEL CASO DE USO BUSCAR CATEGORÍA. 27
TABLA 14. ESPECIFICACIÓN DEL CASO DE USO ASIGNAR POLARIDAD. 28
TABLA 15. TABLA ESPECIFICACIÓN DEL CASO DE USO GENERAR ESTADÍSTICAS. 28
TABLA 16. TABLA ESPECIFICACIÓN DEL CASO DE USO MOSTRAR RESULTADOS. 29
TABLA 17. ETIQUETAS DE POS. 49
TABLA 18. NOMBRES DE LOS ARCHIVOS Y CANTIDAD DE COMENTARIOS. 70
TABLA 19. RESULTADOS DE LAS ESTADÍSTICAS DEL ARCHIVO JETTACARS.COM.TXT. 73
TABLA 20. RESULTADOS DE LAS ESTADÍSTICAS DEL ARCHIVO JETTAEDMUNDS.COM.TXT. 75
TABLA 21. RESULTADOS DE LAS ESTADÍSTICAS DEL ARCHIVO JETTAKBB.COM.TXT. 77
TABLA 22. RESULTADOS DE LAS ESTADÍSTICAS DEL ARCHIVO SPARKCARS.COM.TXT. 79
TABLA 23. RESULTADOS DE LAS ESTADÍSTICAS DEL ARCHIVO SPARKEDMUNDS.COM.TXT. 81
TABLA 24. RESULTADOS DE LAS ESTADÍSTICAS DEL ARCHIVO SPARKKBB.COM.TXT. 83

Índice de códigos

Código 1. Pre-procesamiento. 48
Código 2. Asignación de polaridad. 51
Código 3. Almacenamiento de categorías desde Java. 54
Código 4. Almacenamiento de categorías (procedimiento almacenado) 54
Código 5. Generar estadísticas de segundo nivel. 57
Código 6. Generación de gráficas con Chart.js. 57

Resumen

Actualmente el análisis de sentimientos ha adquirido gran interés de estudio debido a sus aplicaciones en mercadotecnia, economía y demás áreas. A pesar de que existen herramientas que ayudan a determinar la polaridad de los comentarios, aún falta profundizar en el nivel de detalle que se da como resultado.

Este trabajo tiene como objetivo describir el proceso que se llevó a cabo para desarrollar una herramienta que analiza comentarios para extraer sus aspectos y ser polarizados como positivos o negativos.

Se utilizó la biblioteca NLTK de Python para el procesamiento del lenguaje natural. Además se creó la ontología *Car* que ayuda a clasificar cada aspecto para agruparlos y representarlos en gráficas.

Para esto, se realizó una aplicación Web con Struts2, utilizando una arquitectura de 2 niveles y 3 capas.

Palabras clave: Análisis de comentarios, procesamiento de lenguaje natural, ontología, extracción de aspectos y *marketing*.

Abstract

Nowadays the sentiment analysis has been taken a lot of interest to be applied in many areas like marketing, economy and others. In spite of existing tools which help to assign polarity for comments, the task to obtain more details of the text is hard.

The goal of this work is describe the process implemented to develop a tool which analyze comments to extract aspects and polarize it like positive or negative.

NLTK of Python was used for the natural language processing. Also the *Car* ontology was created to classify each aspect to be grouped and present the results in charts.

For that, a Web application was developed with Struts2, using an architecture in 2 levels and 3 layers.

Keywords: comment analysis, natural language processing, ontology, aspect extraction, and marketing.

Introducción

La mercadotecnia ha tenido que evolucionar gracias a la diversidad de los medios de comunicación que existen actualmente. Mientras algunas empresas continúan inmersas en la publicidad tradicional, se han implementado nuevas formas de persuadir a los clientes como la mercadotecnia boca a boca (*word-of-mouth marketing*).

George Silverman, creador de la mercadotecnia boca a boca, habla sobre un viaje en tres universos distintos para entender la evolución de la mercadotecnia. [14]

El primero es el universo de nuestros abuelos, en el que ansiaban tener información, esperaban los periódicos y hacían sus compras a través de catálogos que conseguían con vecinos, la mercancía permanecía en la bodega hasta efectuar la compra, es decir, no existía una interacción con el producto.

El segundo universo es el de nuestros padres, que se vio afectado por los grandes avances tecnológicos, la radio, la televisión y demás medios, en los cuales los anuncios abundaban, por lo que las personas tenían la necesidad de confirmar la veracidad de la información con otras personas, para saber si realmente lo que decía la publicidad era cierto.

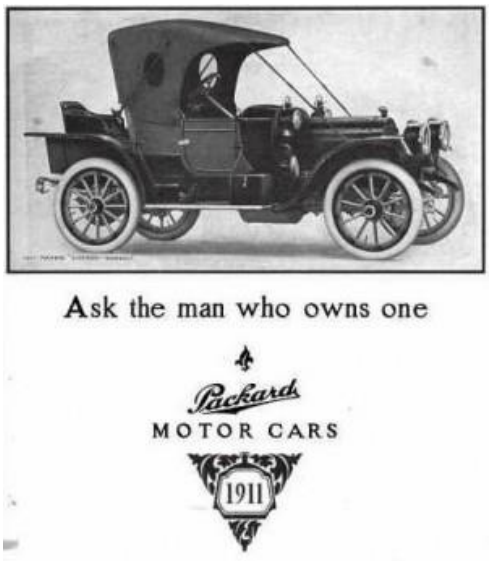

Figura 1. Publicidad de Packard Motors.

Un ejemplo claro es la publicidad de Packard Motors (ver Figura 1), empresa que utiliza como lema "*Ask the man who owns one*", que en español es "Pregúntale al hombre que tiene uno". [14]

Finalmente el universo futuro es el que vivimos actualmente, es una etapa en la que vivimos conectados a pesar de las distancias y nos esforzamos por ignorar la publicidad porque aunque nos persigue no hay tiempo para anuncios tradicionales.

En el estudio "El poder de la recomendación boca a boca en las decisiones de compra del consumidor mexicano" [20], realizado por Leventer Group en 2007, se obtuvieron los siguientes resultados:

- 9 de cada 10 consumidores que eligen por primera vez a una compañía, dicen que los comentarios y recomendaciones tuvieron gran peso en su decisión.
- 7 de cada 10 recomienda un servicio cuando este fue excelente y que la decisión final sobre su compra se sustenta en la frecuencia de comentarios y recomendaciones, por el tono de los comentarios (positivos o negativos) además de la credibilidad del emisor.

Dicho estudio refleja cierto comportamiento de los usuarios en relación con la mercadotecnia boca a boca, donde es importante, la cantidad de comentarios, las calificaciones asignada por personas con alguna experiencia respecto a cierto producto, además de poder confiar en el emisor del comentario.

Es por ello que en la mercadotecnia se crea una nueva estrategia: *digital marketing* o mercadotecnia digital (traducido al español). Según la revista Merca2.0 es de suma importancia debido a que los consumidores están enganchados a las redes sociales, el comercio electrónico, y toda la información que les otorga el Internet a través de sus dispositivos móviles. La mercadotecnia digital se encarga de estudiar y traducir la influencia que tiene *social media* en un determinado producto. Muchas empresas aun contratan personas para que lean los periódicos, vean la televisión o escuchen la radio para conocer qué es lo que se dice de los productos o servicios de su empresa. Esto genera grandes costos y además de que no se escucha la opinión de los consumidores. [24]

Los nuevos medios de comunicación llevaron a la creación de nuevas estrategias de mercadotecnia que a su vez, éstas requieren de nuevas herramientas que apoyen al estudio de sus mercados potenciales o ayuden a vigilar la reputación de las empresas. Para el desarrollo de dichas herramientas es fundamental escuchar a los clientes, es decir, seguirlos y estudiarlos en el medio que se encuentren.

Gracias a las redes sociales las personas tienen un espacio para expresarse, decir lo que sienten, lo que piensan y lo que opinan sobre los productos o servicios que consumen. Las opiniones de los consumidores son muy importantes y han sido estudiadas por áreas como psicología, sociología, computación, etc. Sin embargo, no significa que sea de forma automatizada. Por ejemplo, una infografía creada por Daniel Salgado de Effective Talk (agencia de *marketing* digital), menciona las técnicas que utilizan en las etapas cuantitativas y cualitativas de la escucha del cliente (ver Figura 2). En el artículo [13] que acompaña dicha figura, él dice:

> "*Al contrario de la parte cuantitativa, esta parte (cualitativa) es mucho más manual y la mejor forma de hacerlo es sentarte con una taza de café y ponerte a leer todos los comentarios que la herramienta saca.*"

Figura 2. Infografía.

Con dicha opinión de un mercadólogo, podemos observar que es necesaria la automatización del análisis de los comentarios, ya que mientras más comentarios se analicen más información se extraerá para la posterior toma de decisiones o predecir el éxito/fracaso de una campaña publicitaria. Por ejemplo, la marca Zara, escuchó las opiniones de sus clientes sobre una camiseta (ver Figura 3) que según comentarios de las personas en redes sociales, era parecida al uniforme judío durante la primera guerra mundial. Gracias a que la empresa escucho esas opiniones las camisetas fueron retiradas y se ofreció una disculpa pública.

Figura 3. Imagen que circuló por las redes sociales de la camiseta de Zara.

Como se puede observar, el hecho de escuchar las opiniones es muy importante para las empresas. Es así como inicia el interés de la comunidad científica de procesamiento de lenguaje natural y varias ramas de la inteligencia artificial, por la minería de opinión u *opinion mining* para que el análisis de sentimientos llegue a un mayor nivel de detalle.

Planteamiento del problema

El *marketing* ha evolucionado y nos encontramos en una etapa donde el cliente no necesita que la empresa lo convenza de comprar su producto, sino que desea conocer la opinión de otras personas, sobretodo de aquellas que han tenido alguna experiencia con el producto.

Es fundamental para las empresas escuchar las opiniones de personas que experimentaron su producto o servicio, para conocer qué es lo que sintieron, pues de ellos dependen futuras recomendaciones. Para esto actualmente se han desarrollado herramientas que clasifican los comentarios como positivos, negativos o neutros asignando polaridad al

comentario completo. Esto ayuda a determinar si el producto en generan les agrada o no a los consumidores. Sin embargo no se tienen detalles del porqué les gustó o no, en realidad no se tiene certeza de qué características del producto les agrada o desagrada.

Por ejemplo, el comentario: "*my new Jetta is amazing but ac is a little weak*". Con herramientas simples, se diría que el comentario contiene una palabra positiva (*amazing*) y otra negativa (*weak*) así que promediando dichos resultados se tendría una polaridad neutra. En este caso sería bueno detectar el sentimiento de cada uno de los aspectos del producto para que los mercadólogos observen los aspectos que a las personas no les agradan de su producto.

Es por eso que se plantea la siguiente pregunta de investigación ¿Por qué no desarrollar una herramienta que analice comentarios para conocer los sentimientos del consumidor acerca de cada una de las características del producto?

Propuesta de solución

Ante dicho problema se propone el desarrollo de un sistema que realice el análisis de los comentarios. El propósito no es asignar un sentimiento a cada comentario, sino un sentimiento a cada uno de los aspectos que menciona el cliente.

El sistema contará con los módulos que se observan en la Figura 4.

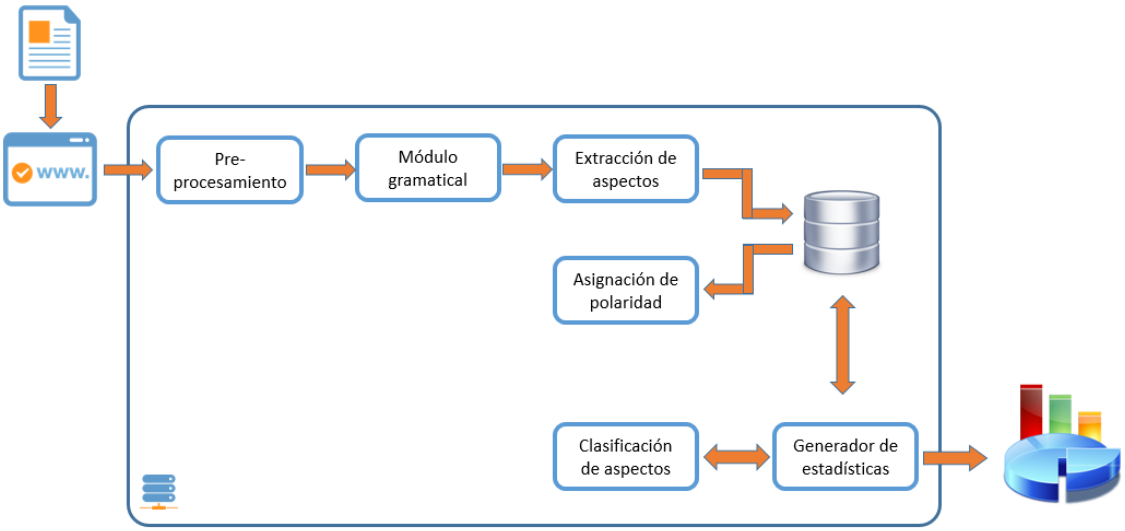

Figura 4. Arquitectura general del sistema.

La entrada al sistema, como se observa en la Figura 4, es un archivo de texto con todos los comentarios a ser analizados, estos deben estar separados por un salto de línea. El archivo se proporciona al sistema a través de una página web que muestra una ventana para navegar en el directorio de archivos y así seleccionar el archivo deseado. Una vez seleccionado el archivo inicia el análisis automático de los comentarios, pasando por cada una de las etapas siguientes:

- **Pre-procesamiento**.- Se hace una normalización de los textos a letras minúsculas para no hacer diferencia con la misma palabra, por ejemplo *the* y *The*. Se eliminan los signos de puntuación, al ser textos que no llevaron algún tipo de revisión no se pueden utilizar signos de puntuación para separar el comentario en oraciones por lo tanto se descartan. Cada comentario se debe separar en palabras (*tokens*) para posteriormente ser etiquetado de acuerdo con su función en la oración (*part of speech*).
- **Módulo gramatical**.- Una vez que se conoce la función de cada palabra se deben analizar oraciones, los signos de puntuación fueron eliminados porque se determinó que no servían para esta tarea. Así que se eligieron dos estructuras gramaticales para encontrar oraciones útiles para el análisis:

  `Sustantivo + Verbo (to be) + Adjetivo`
  `Adjetivo + Sustantivo`

  Así que en este módulo la tarea es determinar si el comentario tiene una o varias veces alguna de las estructuras gramaticales.
- **Extracción de aspectos**.- De acuerdo con las estructuras gramaticales elegidas los sustantivos son los aspectos que critican las personas en sus comentarios. Así la tarea es extraer los sustantivos que estén dentro de una de las estructuras gramaticales mencionadas.
- **Asignación de polaridad**.- Se asigna a cada adjetivo (sentimiento) la polaridad, ya sea positiva o negativa con base en un par de listas de 2005 palabras positivas y 4783 palabras negativas, que fueron elaboradas por el profesor Bing Liu de la Universidad de Illinois [10].

- **Almacenamiento en la base de datos**.- Se construyó una base de datos que relaciona los aspectos, sentimientos y categorías con sus respectivas sentencias y comentarios de donde se extrajeron. Esto con el fin de realizar futuras consultas que serán necesarias para presentar las gráficas al usuario del sistema.
- **Clasificación de aspectos**.- Hasta este punto los aspectos son palabras aisladas que no se relacionan entre sí. Para esto se desarrolló la ontología *Car* cuyos conceptos son partes de un automóvil, cada uno cuenta con sus sinónimos y tiene un padre en la jerarquía que es la categoría a la cual pertenece cada parte del automóvil. Los aspectos son partes del automóvil y tienen una categoría padre que agrupa varias partes, así que se navega la ontología para saber quiénes son las categorías padre de cada aspecto y se guardan en la base de datos para futuras consultas.
- **Generador de estadísticas**.- Se realizan consultas a la base de datos para saber la cantidad de aspectos positivos y negativos que tuvo cada categoría. Se hace comparación entre la cantidad de veces que los autores de los comentarios describieron el auto o una parte de éste. Para mostrar los resultados se utilizan bibliotecas que ayudan a graficar en la página Web.

Alcances

Existen algunas restricciones en desarrollo del sistema:

- Todos los comentarios del archivo hablarán de un automóvil en especial.
- Los comentarios deben ser extraídos de un sitio en el que escriban personas que son o fueron dueñas del automóvil.
- El idioma de los comentarios debe ser inglés.
- Los comentarios son seleccionados por el usuario del sistema, éstos deben tener al menos una oración con las estructuras gramaticales:

```
Sustantivo + Verbo (to be) + Adjetivo
Adjetivo + Sustantivo
```

- Los comentarios deben estar en un archivo de texto, cada comentario deberá estar separado por un salto de línea.

- No se atiende a la conjugación de verbos.

El prototipo final es un sistema Web que muestra gráficas de la cantidad de aspectos positivos y negativos, agrupados en las distintas categorías a las cuales pertenecen las partes del automóvil. Estas gráficas son el resultado de las consultas realizadas a la base de datos, que es donde se almacenan todos los datos resultantes de la navegación en la ontología y el análisis lingüístico de los comentarios.

Justificación

Ya que el análisis manual de los comentarios es una tarea muy tardada, se han desarrollado diversas herramientas que ayudan a decidir la polaridad de los comentarios. Una forma sencilla es contar las palabras definidas como positivas o negativas, obtener un promedio y decir si es un comentario positivo, negativo o neutro. Sin embargo este método no relaciona los sentimientos con un objeto. Por ejemplo, suponiendo que una persona comenta un aspecto bueno y un aspecto malo de un mismo producto, si se usa una herramienta como la anteriormente mencionada no se obtiene información valiosa de cada uno de los aspectos descritos en el comentario. Es decir, el resultado sería neutro y probablemente el comentario sería ignorado a pesar de contar con información valiosa.

Para la toma de decisiones al redirigir una campaña publicitaria no basta con saber que a las personas les gusta o no el producto, también es importante conocer qué sienten respecto a cada aspecto del producto, por ejemplo, saber si el nuevo color o diseño les gustó. Sin embargo esta información no solo es útil para las empresas, también es de gran utilidad para personas que están en búsqueda de datos que ayuden a la toma de decisión de la futura compra de un producto.

Al inicio puede parecer que la información es escaza, pero cuando se navega en el gran mundo del Internet ésta es interminable. Por esta razón y para facilitar la búsqueda de información útil, ha sido de gran interés en varias ramas de la inteligencia artificial, analizar automáticamente grandes cantidades de texto.

Para extraer información de un texto no basta conocer vocabulario o reglas gramaticales, también es necesario que las palabras se encuentren relacionadas, palabras aisladas no ayudan cuando queremos comunicarnos, simplemente se aíslan porque no tienen algún significado que se relacione con otras palabras. Por ejemplo, cuando se quiere añadir una palabra nueva al vocabulario es necesario aprenderla tantas veces sea necesaria hasta adquirir un concepto nuevo, así ese concepto se relacionará con otros que ya se tienen almacenados.

Los aspectos que se mencionan en determinado comentario están relacionados con otros aspectos más generales que no se mencionan, sin embargo la computadora no lo sabe. Por ello es necesario proporcionar un conjunto de conceptos relacionados, para que además de responder con los conceptos que se dan explícitamente, se realicen inferencias dando como resultado aspectos más generales que no se encontraban en el texto (aspectos implícitos).

Para extraer aspectos implícitos es necesario tener un experto que pueda inferirlos dadas las características descritas por el autor del comentario. Esto refleja que el sistema debe trabajar específicamente con una rama de la lingüística: la semántica. La semántica se trabaja en diversas áreas de la computación, con reglas, diccionarios y ontologías. Siendo éstas últimas muy populares, debido a su amplio uso en la Web Semántica. Las ontologías que se desarrollan son sobre un tema en específico y todos los conceptos así como las relaciones deben estar validadas por un grupo de expertos en el área.

Ya que las ontologías son especializadas para un tema específico, el trabajo debe limitarse a un solo tema, así que con estudios realizados por la AMIPCI fue posible elegir el tema central del cual debían buscarse los comentarios a ser analizados. Uno de los estudios más relevantes para la búsqueda del tema es “Estudio de Marketing Digital y Social Media 2014” [16], el cual fue un análisis realizado a las 100 empresas más grandes de México de acuerdo con el ranking de la revista Expansión 2014. En dicho estudio el 80% de las empresas tiene presencia en *Twitter* y/o *Facebook*. Por ejemplo, en *Facebook* los giros con mayor cantidad de fans son: bebidas y cervezas, productos de consumo y armadoras. Por otro lado, los giros con más seguidores en *Twitter* son: aerolíneas, logística

y transporte, telecomunicaciones y armadoras. Con estos resultados se observa que las empresas armadoras están interesadas en conocer los comentarios de los consumidores.

Por otro lado, en noviembre de 2013 *Google Think Insights* publicó los resultados de su estudio [21] "¿Cómo deciden los mexicanos comprar un auto nuevo?". En dicho estudio muestran los siguientes resultados relevantes:

- 74% de los consumidores investiga a través de buscadores para saber que auto comprar.
- 61% opina que los buscadores les ayudaron a descubrir nuevas marcas.
- 68% dijo que los videos online les facilito encontrar información sobre autos.
- 65% dijo que los videos online les ayudaron a descubrir marcas que nunca habían escuchado.
- 68% de las visitas a concesionarias se hacen a la mitad y al final del proceso de compra.
- 80% de los compradores llegan a la concesionaria con información en mano.

Este estudio muestra que al conseguir información para la adquisición de un auto, los consumidores no la obtienen (en su mayoría) a través de vendedores, sino de la información que está en buscadores y redes sociales.

México es un país en el que se encuentran diversas armadoras y marcas de automóviles en el mercado, sin embargo no es el país con la mayor cantidad de ventas. China y Estados Unidos están en los primeros lugares con ventas de más de 12 millones de dólares en 2011, así que este mercado es realmente extenso y con grandes ganancias.

La información anterior define que la industria automotriz es increíblemente amplia y por lo tanto es un tema sumamente relevante. Pero ¿qué ocurre con los consumidores? ¿en realidad necesitan o quieren conocer los productos de estas empresas?, la respuesta es que sí. Existen diversos programas de radio, columnas en periódico y revistas dónde expertos de automóviles, describen, critican, comparan y evalúan automóviles, todo con el fin de ayudar a los futuros propietarios de un automóvil a realizar una buena elección. Sin embargo, los consumidores no se limitan a leer o escuchar los comentarios de expertos,

sino que necesitan conocer las experiencias de distintas personas que son o fueron dueñas del que podría ser su futuro automóvil. Es por ello que la Web se ha llenado de infinidad de foros que se dedican a recopilar reseñas que ayuden a los lectores en su compra.

Metodología

Con base en el libro "Cómo elaborar y asesorar una investigación de tesis" [12], se determinó que la guía para el desarrollo del proyecto es una metodología de investigación tecnológica, cuyas etapas son:

- Identificación del problema.
- Formulación del problema de investigación.
- Identificación del objeto de estudio.
- Formulación del objetivo de la investigación.
- Definición de hipótesis o propuesta tecnológica.
- Estado del arte.
- Identificación del campo específico de la investigación.
- Elaboración del modelo teórico conceptual.
- Liberación y producción masiva.
- Formulación y difusión de resultados.

Dicha metodología fue seleccionada porque la investigación tecnológica es la aplicación de un conjunto de técnicas, conocimientos y procesos generados por la ciencia para resolver problemas de la sociedad. Se resalta que en esta metodología el diseño inicial no es único ni definitivo, porque no existe una solución única al problema pues siempre es posible mejorar y modificar el diseño constantemente, debido a que las técnicas y métodos están en constante evolución renovándose para adaptarse a los conocimientos científicos.

Esta metodología fue de gran utilidad para obtener el prototipo final. Como se mencionó en el párrafo anterior, no existe una solución única al problema pues siempre es posible mejorar y modificar el diseño constantemente, esto se vio reflejado principalmente en las herramientas para el procesamiento de lenguaje natural, la ontología que tuvo que ser

creada, el diseño de la base de datos y los algoritmos a ser programados así como el lenguaje de programación del cual se desconocía el paradigma y su sintaxis.

Objetivo general

Desarrollar un sistema que analice comentarios para conocer los sentimientos expresados en dichos comentarios sobre un producto.

Objetivos particulares

- Investigar y seleccionar los algoritmos y técnicas de minería de opinión a utilizar.
- Seleccionar los tópicos de los comentarios con los cuales se trabajará y los sitios de donde se extraerán estos.
- Investigar e implementar las reglas de los análisis léxico y semántico.
- Evaluar y clasificar los sentimientos expresados en los comentarios, con el algoritmo y/o técnicas seleccionadas de minería de opinión.
- Desarrollar la aplicación para mostrar los resultados del análisis de los comentarios.

Capítulo 1. Marco teórico

1.1. Análisis de sentimientos

El análisis de sentimientos y la minería de opinión es el campo de estudio que analiza las opiniones, sentimientos, valoraciones, actitudes y emociones de la lengua escrita de las personas. Es una de las áreas de investigación más activas en el procesamiento del lenguaje natural y también es ampliamente estudiada en la minería de datos, minería Web, y la minería de texto.

De hecho, esta investigación se ha extendido fuera de la informática a ciencias de la gestión y ciencias sociales, debido a su importancia para los negocios y la sociedad en conjunto. La creciente importancia del análisis de los sentimientos coincide con el crecimiento de los medios de comunicación sociales, foros de discusión, *blogs*, *micro-blogs*, *Twitter* y las redes sociales. Por primera vez en la historia humana, se cuenta con un enorme volumen de datos de opinión grabados en formato digital para su análisis. [11]

Los sistemas de análisis de sentimiento se aplican en casi todos los ámbitos empresariales y sociales porque las opiniones son fundamentales para casi todas las actividades humanas y son importantes factores de influencia de nuestros comportamientos. Nuestras creencias, percepciones de la realidad y las decisiones que tomamos, están condicionadas en gran medida de la forma en que otros ven y evalúan el mundo. Por esta razón, cuando tenemos que tomar una decisión con frecuencia se buscan las opiniones de los demás. Y esto es cierto no sólo para los individuos sino también para las organizaciones. [11]

1.2. Lenguaje y el procesamiento del lenguaje natural

Los humanos poseemos una forma muy eficaz de comunicarnos: usamos el lenguaje natural. Incluso, para nosotros no es solamente la herramienta de comunicación, es un modo de pensar. [4]

Otro punto importante del lenguaje natural es su habilidad de representar el conocimiento. La mayor parte del conocimiento tiene una forma simbólica, es decir, la forma de texto escrito. Así que, si las computadoras entendieran el lenguaje natural, serían unos ayudantes mucho más valiosos y efectivos. [4]

Entonces, ¿cómo hacer que las computadoras entiendan lo que nosotros dominamos con una gran facilidad: la lengua natural? Para ello se estudia desde el punto de vista de la ciencia del lenguaje humano: la lingüística. La lingüística construye modelos del lenguaje o de diferentes fenómenos relacionados con el lenguaje. Así que la lingüística computacional trabaja con modelos que deben ser entendibles para las computadoras, esto es, modelos que tienen un grado adicional de precisión y formalización. Un modelo bueno debe explicar, predecir, ser elegante y simple; no introducir conceptos innecesarios. [4]

Para poder desarrollar modelos es preferible dividir el objeto en partes y construir modelos más pequeños, para esto se usa el concepto de niveles del lenguaje. Tradicionalmente, el lenguaje natural se divide en seis niveles:

- Fonética/fonología
- Morfología
- Sintaxis
- Semántica
- Pragmática
- Discurso

No existen criterios exactos para la separación de cada uno de los niveles; de hecho, las diferencias entre los niveles se basan en el enfoque de análisis de cada uno. Por eso pueden existir traslapes entre niveles sin presentar contradicción alguna. [4]

1.3. Semántica

La semántica es uno de los niveles de la lingüística. El propósito de la semántica es “entender” la frase. Para entender hay que saber el sentido de todas las palabras e interpretar las relaciones sintácticas. Los investigadores están más o menos de acuerdo con los resultados del análisis semántico deben ser redes semánticas, donde se representan

todos los conceptos y las relaciones entre ellos. Otra posible representación son los grafos conceptuales. [4]

1.4. Ontologías y el conocimiento

Partiendo del criterio existente sobre el papel principal del conocimiento en la solución de problemas, resulta claro que es necesario considerar cuidadosamente como realizar el almacenamiento del mismo. Para realizar el almacenamiento del conocimiento se determina su representación, lo cual no es más que el proceso de estructurar el conocimiento sobre un dominio de aplicación de modo que los problemas sean más fáciles de resolver. Esta representación debe expresar con claridad tanto los hechos del dominio, como las relaciones entre los mismos. No existe actualmente una forma de representación del conocimiento general capaz de ser usada en todo tipo de aplicación con éxito; las formas disponibles están limitadas a más o menos un dominio específico. [15]

La Web Semántica fue propuesta (actualmente en desarrollo) por Tim Berners-Lee, creador de la Web y fundador del *World Wide Web Consortium*. El objetivo es filtrar la información que actualmente se encuentra en la Web para dar información relevante al usuario cuando la necesite. La Web Semántica utiliza ontologías para lograr su objetivo y ha desarrollado lenguajes para su modelado. Es costoso crearlas porque es necesario que los conceptos sean revisados por un grupo de expertos en el tema, sin embargo, otorgan buenos resultados ya que realizan inferencias y son capaces de responder a los usuarios a preguntas no exactas.

El término ontología proviene de la filosofía, donde es el estudio de la existencia. Sin embargo, actualmente es un término muy importante en las tecnologías semánticas. En ciencias de la computación, una ontología es una descripción del conocimiento sobre un dominio de interés, cuyo núcleo es una especificación procesable por una máquina, con un significado definido formalmente. [9]

En la Web Semántica, los vocabularios definen conceptos y relaciones (también conocidos como "términos") utilizados para describir y representar un área de ocupación. Los vocabularios se utilizan para clasificar los términos que se pueden utilizar en una

aplicación particular, caracterizar posibles relaciones, y definir posibles restricciones sobre el uso de esos términos. En la práctica, los vocabularios pueden ser muy complejos (con varios miles de términos) o muy simples (que describen sólo uno o dos conceptos).

La tendencia es a usar la palabra "ontología" para vocabularios más complejos, y la recolección bastante formal de términos. Los vocabularios son los bloques básicos de construcción para las técnicas de inferencia.

El papel de los vocabularios es para ayudar a la integración de datos, cuando por ejemplo, existen ambigüedades en los términos utilizados en diferentes conjuntos de datos, o cuando un poco de conocimiento adicional puede conducir al descubrimiento de nuevas relaciones. La complejidad del vocabulario dependerá de la aplicación.

Para satisfacer estas diferentes necesidades, W3C ofrece una amplia gama de técnicas para describir y definir diferentes formas de vocabularios en un formato estándar, además de mostrar listas de algunas herramientas útiles para su edición y visualización con su respectiva liga en Internet. [27]

1.5. RDF

RDF (*Resourcer Description Framework*) es un lenguaje formal para describir información estructurada. Su objetivo es establecer aplicaciones para intercambiar datos en la Web mientras se prevalece el significado original. En comparación con HTML y XML la principal intención no es mostrar documentos correctamente, sino permitir procesamiento adicional y recombinar la información contenida. En consecuencia RDF es visto con frecuencia como el formato de representación básica para desarrollar la Web Semántica. [27]

1.6. OWL

OWL (*Ontology Web Language*) es un estándar recomendado por la W3C desde 2004 para modelar ontologías. El objetivo principal en su diseño era encontrar un equilibrio razonable entre la expresividad del lenguaje y el razonamiento eficiente. [27]

Esto fue con el fin de ser usado en aplicaciones que necesitan procesar el contenido de la información en lugar de únicamente representar información para los humanos. OWL facilita un mejor mecanismo de interoperabilidad de contenido Web que los mecanismos admitidos por XML, RDF, y esquema RDF (RDF-S) proporcionando vocabulario adicional junto con una semántica formal. OWL tiene tres sub-lenguajes, con un nivel de expresividad creciente: OWL Lite, OWL DL, y OWL Full. [28]

1.7. Herramientas del procesamiento del lenguaje natural

Para que cualquier usuario pueda interactuar con un sistema en su lenguaje natural debe haber todo un trabajo detrás. Por ello existen lenguajes de programación que cuentan con bibliotecas especiales para trabajar con diversos idiomas, etiquetadores y demás herramientas que apoyan el desarrollo del sistema.

1.7.1. Python

De acuerdo con la página oficial, Python es un lenguaje de programación que permite integrar los sistemas más efectivamente y trabajar más rápido [26]. Es también un lenguaje amigable y fácil de aprender. Su sintaxis hace que por naturaleza los códigos sean ordenados y por lo tanto fácilmente entendibles. Existen tipos de dato que no se encuentran en otros lenguajes como Java o C. Permite al programador enfocarse en el desarrollo de un algoritmo sin la necesidad de atender a la declaración de variables. Todas esas características se equilibran con el hecho de que es un lenguaje interpretado y por lo tanto la velocidad de ejecución del programa es más lento.

Debido a la gran flexibilidad de Python, existen bibliotecas para diversas tareas como *scraping*, graficación, desarrollo de juegos, animación en 3D y por su puesto procesamiento de lenguaje natural.

1.7.2. Natural Language Toolkit (NLTK)

Como lo mencionan en su página oficial NLTK “es una biblioteca increíble para jugar con el lenguaje natural”. [25] Contiene corpus y recursos léxicos como *WordNet* [29] así como bibliotecas para la clasificación y etiquetado. Es de código abierto, está disponible

para los sistemas operativos Windows, Mac OS y Linux, existe un foro de discusión activa y hay documentación además de un libro escrito por los creadores.

Es un conjunto de bibliotecas muy completo que permiten elegir entre diferentes algoritmos para implementar y se encuentra en constante actualización.

Capítulo 2. Estado del arte

A continuación se menciona trabajos relacionados con el sistema desarrollado, éstos fueron realizados por alumnos del IPN.

2.1. "Generación automática de resúmenes de opinión sobre entidades nombradas en español" [7]

Este trabajo se desarrolló en el año 2013 y trata sobre el desarrollo de un sistema que genera el resumen de comentarios representativos de un conjunto de opiniones clasificadas como positivas, negativas, neutras e informativas acerca de un producto, para saber lo que piensa la gente que se encuentra en redes sociales. Este sistema va dirigido a los usuarios que necesitan información de un producto. Se llevó a cabo obteniendo publicaciones de Facebook que en su contenido tuvieran las entidades "*tablet*" o "*iPad*". Los comentarios se clasificaron con el algoritmo *Naive Bayes* y se obtuvo un resumen utilizando heurísticas y frecuencias.

Algunas características importantes son:

- Sus usuarios son los consumidores de productos, que buscan opiniones de diferentes personas en Facebook.
- Trabaja con grandes cantidades de comentarios pero no se especifica si son comentarios elegidos al azar.
- Utiliza únicamente dos entidades "*tablet*" e "*iPad*".

2.2. "Sistema de análisis del impacto de productos a través de las redes sociales Facebook y Twitter desde la perspectiva del usuario" [5]

Es un trabajo que fue desarrollado en el 2014 y tiene como objetivo crear un sistema para extraer comentarios de las redes sociales *Facebook* y *Twitter*, dichos comentarios se analizan con algoritmos de clasificación y heurísticas para identificar la aceptación de productos específicos. Es un sistema que va dirigido a personas en el área de

mercadotecnia. Al inicio del trabajo mencionan que trabajarán solo con un tema, así que sus pruebas fueron realizadas respecto al producto "*iPhone*".

El usuario observa como resultado gráficas que muestran el análisis estadístico que realizaron para mostrar los resultados obtenidos de su análisis de los comentarios.

Algunas características importantes son:

- Sus usuarios son personas del área de mercadotecnia.
- Trabaja con grandes cantidades de comentarios extraídos de *Facebook* y *Twitter*.
- Utilizan un tema en especial, es decir, un producto. Sus pruebas fueron realizadas con la entidad "*iPhone*"

2.3. "Arquitectura Web para análisis de sentimientos en Facebook con enfoque semántico" [1]

Este trabajo tiene como objetivo desarrollar una arquitectura de software que clasifique publicaciones de *Facebook* en tres emociones: alegría, enojo y tristeza. Su sistema lleva a cabo la extracción de publicaciones usando el núcleo de la interfaz de programación de *Facebook*, pre-procesamiento de palabras, clasificador base semántico, clasificador *Naive Bayes*, funciones de administración Web y despliegue de resultados. La sección de clasificación está dividida en dos: primero se emplea un recurso que contiene conceptos de sentimientos ponderados basada en *WordNet Affect*[1] ya que los conceptos se encuentran organizados y enriquecidos semánticamente como una ontología; después se ejecuta un clasificador *Naive Bayes* que fue seleccionado debido a los resultados de publicaciones anteriores.

Algunas características importantes son:

- Es un trabajo que implementa un clasificador supervisado y además crean una jerarquía basada en *WordNet Affect* que contiene conceptos con distintos niveles de intensidad emocional.

[1] WordNet Affect es una extensa base de datos léxica que relaciona los conceptos de palabras afectivas. [29][30]

- Está dirigido a usuarios mexicanos ubicados en la Ciudad de México y área metropolitana.
- El español es el idioma de las publicaciones a ser analizadas y se asume que tienen buena ortografía.
- Las publicaciones que se analizan son las del perfil del usuario que inició sesión en el sistema.

En la Tabla 1 se muestra la comparación de algunas características importantes de los trabajos mencionados con el sistema desarrollado.

Tabla 1. Comparación de trabajos.

Parámetro	Sistema de análisis lingüístico de comentarios para el apoyo del marketing boca a boca. (Proyecto desarrollado)	Generación automática de resúmenes de opinión sobre entidades nombradas en español.	Sistema de análisis del impacto de productos a través de las redes sociales Facebook y Twitter desde la perspectiva del usuario.	Arquitectura Web para análisis de sentimientos en Facebook con enfoque semántico
Propósito	Asignar polaridad a cada sentimiento de los aspectos descritos en los comentarios de personas que hablan de un mismo producto	Generar resúmenes de opiniones clasificadas como positiva, negativa o neutra	Extraer comentarios de las redes sociales para analizarlos e identificar la aceptación de productos específicos	Clasificación de publicaciones de perfil de Facebook en alegría, enojo y tristeza.
Usuario	Mercadólogos en el área de *marketing* digital o consumidores interesados en la búsqueda de información para la compra de un producto	Consumidores de productos que buscan opiniones en Facebook	Personas interesadas en conocer la tendencia en redes sociales así como las opiniones de los consumidores	Usuarios mexicanos con una cuenta de Facebook.
Tema	Automóviles	Utilizan entidades: “Tablet ” e “iPad”	Un tema: “iPhone” de acuerdo con las pruebas	Ninguno en específico
Entrada	Comentarios relacionados con un solo automóvil proporcionados por el usuario	Buscan comentarios en Facebook	Buscan comentarios en Facebook y Twitter	Al iniciar sesión en el sistema se extraen las publicaciones del perfil.
Salida	Gráficas que muestran las proporciones de polaridad positiva o negativa de aspectos implícitos de los comentarios, así como las de los aspectos que se mencionan en el comentario	Comentarios representativos de un conjunto de opiniones	Gráficos que muestran el análisis estadísticos del análisis de comentarios	Gráfica que muestra la cantidad de publicaciones que fueron relacionadas con cada emoción y aquellas que fueron ambiguas.
Algoritmos, técnicas y/o herramientas	Herramientas de PLN y ontología	Naive Bayes, heurísticas y frecuencias	Naive Bayes, árboles de decisión J48 y K-means (Utilizando WEKA)	Naive Bayes y una jerarquía basada en WordNet Affect.

Capítulo 3. Análisis

3.1. Análisis de requerimientos

3.1.1. Propósito

En esta sección de análisis se muestran los requerimientos que fueron extraídos al analizar el problema planteado, que sirvieron como base para elegir las herramientas y llevar a cabo la implementación de la solución propuesta. El Instituto de Ingeniería Eléctrica y Electrónica (IEEE) estableció el estándar IEEE-830 para la especificación de requerimientos de software, por lo que se hizo uso de éste. [22]

3.1.2. Ámbito del sistema

El sistema desarrollado tiene como objetivo analizar comentarios proporcionados por el usuario, para extraer los aspectos (características) del producto del cual se habla, así como asignar una polaridad al sentimiento que describe cada aspecto. Los aspectos son clasificados para presentar gráficas en una página Web que muestran la agrupación de los aspectos respecto a su categoría padre.

No se lleva a cabo la recolección automática de comentarios relevantes como entrada al sistema.

3.2. Descripción general

Algunos factores que afectan a la toma de decisiones al elegir las herramientas son: el idioma y el tema específico con el que se trabajará. Dichos factores y algunas otras consideraciones se describen en las siguientes subsecciones.

3.2.1. Perspectiva del producto

El sistema en desarrollado no tiene relación con algunos otros trabajos desarrollados, es decir, no es un trabajo a futuro propuesto por algún trabajo descrito en el estado del arte.

Existen varias herramientas en el mercado que sirven para extraer comentarios, sin embargo no se utilizan. Así que como entrada se solicita un archivo de texto con los comentarios seleccionados por el usuario. Para la salida del sistema se utiliza una herramienta que ayuda a mostrar gráficas para el usuario.

3.2.2. Funciones del producto

El producto final de este proyecto es una aplicación Web, donde el usuario puede realizar funciones como:

- Proporcionar su archivo de comentarios con un formato específico.
- Observar las gráficas con los resultados del análisis de los comentarios contenidos en el archivo proporcionado.

Para poder entregar gráficas como resultado para el usuario es necesario que el sistema lleve a cabo las siguientes tareas:

- Pre-procesar los comentarios
- Analizar los comentarios gramaticalmente
- Extraer aspectos
- Asignar polaridad a los sentimientos extraídos
- Agrupar los aspectos
- Almacenar todos los datos obtenidos
- Generar estadísticas

3.2.3. Características de los usuarios

El prototipo desarrollado es un sistema dirigido a personas que se encuentra en el área de mercadotecnia digital, es decir, mercadólogos que se encuentren llevando a cabo análisis de *social media* [2]. La información obtenida del análisis también es útil para futuros consumidores de un producto que están en búsqueda de opiniones de otras personas que han tenido algún tipo de experiencia con el producto descrito en sus reseñas.

[2] *Social media.* Medio por el cual las empresas o la marca obtiene retroalimentación constante sobre su producto, gustos y preferencias del público.

3.2.4. Restricciones

De acuerdo con el análisis de las herramientas se encontraron restricciones para el sistema:

- El idioma de los comentarios es el inglés.
- Al trabajar con ontologías debe elegirse un dominio pues no se cuenta actualmente con una ontología que contenga el conocimiento de todos los dominios existentes. Así que con estudios analizados en la justificación de este trabajo se llegó a la conclusión de trabajar con automóviles.
- El archivo de entrada debe contener comentarios de un solo coche, debe tener extensión `TXT`.
- Los comentarios deben estar separados por saltos de línea.
- Se descartarán los *emoticonos* que existan en los comentarios.
- Los sentimientos extraídos se buscarán en listas de palabras positivas y negativas. Estas listas son tomadas del sitio Web del Profesor Bing Liu del Departamento de Ciencias de la Computación de la Universidad de Illinois en Chicago (UIC). Por lo que, si los sentimientos de los comentarios no se encuentran en las listas, no serán tomados en cuenta. [10]
- Los lenguajes de programación son Python y Java. Python es un lenguaje muy utilizado por la comunidad científica para el PLN y Java tiene APIs que apoyan el desarrollo del sistema, como Jena para navegar la ontología y puede trabajar con *frameworks* como Struts2 que obliga a realizar una separación entre la vista y el control del sistema para que la codificación sea ordenada.

3.2.5. Suposiciones y dependencias

Para que el sistema funcione es necesario tener conexión a Internet, pues la forma en la que interactúa el usuario con el sistema es a través de una página Web.

3.3. Requerimientos específicos

Esta sección es importante ya que los requerimientos se describen a un mayor nivel de detalle para permitir llevar a cabo el diseño del sistema. A continuación se mencionan cada uno de ellos clasificados en tres las interfaces externas.

3.3.1. Interfaces externas

3.3.1.1 Interfaz de hardware

Se utilizó un equipo de cómputo proporcionado por el laboratorio de cómputo móvil de la unidad académica. Dicho equipo cuenta con las siguientes características:

- Procesador: AMD E1-1200 APU 1.40 GHz
- Memoria RAM: 16GB
- Capacidad de disco duro: 400 GB de espacio libre disponible
- Conexión a internet: Tarjeta de red inalámbrica
- Sistema operativo: Windows 7 Ultimate (Service Pack 1) de 64 bits

3.3.1.2 Interfaz de software

En esta sección se describe el análisis realizado para seleccionar cada una de las herramientas a utilizar. Se comparan diferentes herramientas y se eligen aquellas con características que se adapten para dar solución al problema.

A. Ontología

Se dio preferencia a buscar una ontología en español, debido al idioma que se habla en el país que se desarrolló el trabajo, sin embargo, las ontologías encontradas están en inglés. En la Tabla 2 se muestra la comparación entre las ontologías encontradas para determinar si alguna de ellas era de utilidad.

Así mismo, no se ha seleccionado alguna ontología de acuerdo con sus características comparadas, así que se describe brevemente cada una:

Tabla 2. Comparación entre ontologías.

Parámetro	Volkswagen Vehicles Ontology	Car Options Ontology	Vehicle Sales Ontology	The Automotive Ontology	DBpedia Ontology
Descripción	Vocabulario para describir las funciones específicas de los automóviles Volkswagen.	Vocabulario para exponer las opciones de configuración disponibles para los modelos de automóviles.	Vocabulario Web para describir coches, barcos, bicicletas y otros vehículos para el comercio electrónico.	Conceptos y diseño de una ontología en desarrollo, para representar conocimiento semántico sobre el usuario, el vehículo y la situación de conducción.	Es el esfuerzo de una comunidad para extraer información estructurada de Wikipedia y ponerla disponible en la Web.
Nivel de profundidad	Solo es útil el primer nivel de clasificación de aspectos.	Profundiza demasiado en conceptos como el color, pero no en los componentes del auto.	Al ser generalizada a varios vehículos, no profundiza en las partes de un automóvil.	No se involucra con las partes del auto.	Contiene conceptos de partes de automóviles pero no tiene gran nivel de detalle.
Desarrollador	Dr. Martin Hepp	Dr. Martin Hepp, Hepp Research GmbH	Dr. Martin Hepp	The Automotive Ontology Working Group	Proyecto de comunidad abierta
Disponibilidad	Si, bajo Creative Commons Attribution 3.0 license.	Si, bajo Creative Commons Attribution 3.0 license.	Si, bajo Creative Commons Attribution 3.0 license.	Si, uniéndose al equipo de trabajo.	Si, uniéndose al equipo de trabajo.
Idioma	Inglés	Inglés	Inglés	Inglés	Inglés
Lenguajes	OWL y RDF	OWL y RDF	OWL y RDF	No se menciona	OWL
Revisada por expertos	Si	Si	Si	Todas las personas interesadas pueden intervenir	La información que organizan está en Wikipedia
Elección					

Volkswagen Vehicles Ontology. Fue una ontología que parecía útil para trabajar, porque contiene conceptos sobre las funciones del automóvil, está revisada por expertos y puede usarse bajo la licencia *Creative Commons Attribution 3.0* para copiarla, distribuirla y transmitir el trabajo. Sin embargo, al revisarla a detalle, se observó que al agrupar los

aspectos, éstos quedarían en un solo nivel, es decir "*engine*" tiene conceptos hijos que agrupan a más conceptos, pero esta ontología no cuenta con el nivel de profundidad que se requiere.

Car Options Ontology. Tiene muy buenas características sin embargo no es suficiente para cubrir los conceptos que se necesitan pues es una ontología dedicada al uso de personas que se encuentran en el área de ventas de automóviles Volkswagen, para realizar inferencias y saber si al modelo del automóvil se le pueden agregar o quitar características como el color, el quemacocos, etc.

Vehicle Sales Ontology. Tiene como objetivo describir autos, botes, bicicletas y otros vehículos para *e-commerce*. La información acerca de autos no es tan detallada como en otras ontologías, por lo que no es seleccionada.

The Automotive Ontology. No se eligió debido a que abarca demasiada información. A pesar de incluir información del automóvil, toma en cuenta aspectos como los datos del conductor y el entorno en el que se conduce, como los viajes que realiza. Además se encuentra en desarrollo colaborativo por un equipo de personas y compañías.

DBpedia Ontology. Se dedica a organizar la información de "Wikipedia", una enciclopedia libre donde todas las personas interesadas pueden editarla. La información contenida en dicho sitio no está aprobada por expertos por lo que no fue seleccionada.

Finalmente se tomó la decisión de crear una ontología llamada *Car*, con ayuda de conceptos de libros y páginas de mecánica automotriz. Además de contar con la supervisión de un asesor experto en el área.

B. Editor ontológico

El *World Wide Web Consortium* (W3C) menciona 12 editores de ontologías (*Protégé*, NeOn Toolkit, Neologism, TopBraid Composer, Vitro, Knoodl, Anzo for Excel, OWLGrEd, Fluent Editor, Semantic Turkey y VocBench) que pueden ser utilizados cuando se requiere construir, visualizar o modificar ontologías. En la Tabla 3 se muestra la comparación de algunos de los editores mencionados en la página Web de W3C.

Tabla 3. Comparación entre editores de ontologías.

Característica	*Protégé*	NeOn Toolkit	SWOOP	Anzo for Excel	TopBraid Composer	Semantic Turkey
Descripción	Plataforma que ofrece un conjunto de herramientas para la construcción de modelos de dominio y aplicaciones basadas en el conocimiento con ontologías	Editor de ontologías esta especialmente equipado para proyectos grandes y pesados (multimodular, multilenguaje o integración de ontologías)	Pequeño y simple editor y navegador de ontologías	Genera una ontologia inicial basada en datos de hojas de cálculo y estructuras	Editor de semántica Web multipropósito	Herramienta de desarrollo de ontologías. Permite hacer seguimiento de los sitios web.
Desarrollador	Standford University	NeOn Foundation	University of Maryland	Cambridge Semantics Inc.	TopQuadrant	University of Rome Tor Vergata
Licencia	Software libre	Plugins libres	Código abierto	No	Si	Código abierto
Lenguaje de programación	Java	--	Java	Javascript, Java y.Net	Java	JavaScript, XUL, XBL
Compatibilidad con herramientas	PROMPT, OKBC, JESS, FaCT y Jena	--	No	No	Sesame, Jena y AllegroGraph	No
Visualización gráfica	Sí	No	Sí	Si	Sí	Con OntoGraph View
Razonador	RacerPro	--	Pellet y RDF-like	SPARQL	OWLIM, Pellet, Jena Rules, Oracle Rules y SPARQL Rules	Reasoner
Entorno de trabajo	Web y Standalone	Eclipse	Web	Excel	Standalone	Navegador Mozilla FireFox
Requisitos	Tomcat y mongoDB (Web) JVM (Standalone)	Software Eclipse	JVM	Excel 2003, Excel 2007 o Excel 2010 y Anzo Server	JVM	JRE y Mozilla FireFox
Fuentes de datos y lenguajes	XML (S), RDF (S), OWL, HTML	RDF y OWL	RDF y OWL	Excel, RDFS y OWL	SPIN, RDF, WOL, XML(S)	RDF, OWL, SKOS,SKOS-XL, SPARQL

Tabla 3. Comparación entre editores de ontologías (continuación).

Característica	*Protégé*	NeOn Toolkit	SWOOP	Anzo for Excel	TopBraid Composer	Semantic Turkey
Última modificación	2015	2011	--	--	2015	2014
Documentación	Sí	Sí, poca	Si, poca	Si	Si	Grupo de discusión
Selección	X					

Con base en los datos anteriores se llegó a la conclusión de utilizar el editor de ontologías *Protégé* pues actualmente es el que tiene mayor soporte sin estar bajo licencia. Se pueden utilizar los lenguajes OWL y RDF los cuales son indispensables para el uso de la ontología desarrollada. Además de que incluye un razonador propio para realizar inferencias, además de que está en constante actualización. [17]

C. Gestor de base de datos relacional

Como se muestra en la propuesta de solución, es necesario almacenar datos relevantes después de haber realizado la extracción de aspectos y asignación de polaridad a los sentimientos. No obstante es importante recalcar que es un modelo relacional simple con el que se trabaja, pues no hay demasiadas entidades ni relaciones complejas entre ellos.

En la Tabla 4 se muestran los gestores de bases de datos SQL Server, MySQL, Oracle, PostgreSQL, para realizar una comparación entre ellos y elegir el adecuado para el sistema.

Tabla 4. Comparación entre gestores de bases de datos.

Parámetro	SQL Server	MySQL	Oracle	PostgreSQL
Desarrollador	Microsoft	Oracle Corporation	Oracle Corporation	PostgreSQL Global Development Group
Licencia	Si	Si, licencia comercial y GPL[3]	Si	Libre

[3] General Public License

Tabla 4. Comparación entre gestores de bases de datos (continuación).

Parámetro	SQL Server	MySQL	Oracle	PostgreSQL
Convenio con el instituto	Si, hasta la versión SQL Server 2014 Professional Edition	Versión con licencia – No	Si, hasta la versión Oracle Database 12c	No aplica
Uso	Aplicaciones para empresas de gran tamaño	Gestión de bases de datos de aplicaciones Web.	Aplicaciones para empresas de gran tamaño	Diseñada para entornos con altos volúmenes de transacciones
Lenguajes soportados	Java, Ruby, Python, VB, .Net, and PHP	C, C#, C++, Java, Python, Ruby, Objective C y más	C, C#, C++, Java, Ruby, Objective C y más	C/C++, Java, .Net, Perl, Python, Ruby, Tcl, ODBC y más
Sistema operativo	Windows	Windows, Linux y OS X	Windows, Linux y OS X	Windows, Linux y OS X
Selección		X		

En este caso no se trabaja con bases de datos distribuidas ni de múltiples usuarios, no se requieren datos espaciales y el esquema relacional es muy simple. Así que debido al uso que se le da a MySQL como gestor de bases de datos para aplicaciones Web, se elige trabajar con la versión de GPL del mismo.

3.3.1.3 Interfaz de comunicaciones

Un aspecto muy importante es la comunicación con el usuario y las gráficas son una forma muy eficiente para comunicar información visualmente, es por ello que es necesario utilizar herramientas que permitan realizar gráficos en un sitio Web. En la Tabla 5 se comparan algunas de ellas.

Tabla 5. Comparación entre bibliotecas para generar gráficas.

Parámetro	Chart.js	Google Chart Tools	HighCharts
Tipo de gráficas	Barras, pastel, donas y líneas en 2D	Barras, pastel, donas, líneas, líneas de tiempo, árboles y más en 2D	Gran variedad de gráficas de barras, pastel, dona, línea y más en 2D y 3D
Plataforma o Lenguaje	JavaScript	JavaScript	JavaScript
Renderización	Canvas	SVG y VML	Canvas
Soporte en HTML5	Si	Si, a partir de las últimas versiones (año 2015) de navegadores: IE, Firefox, Safari, Chrome y Opera.	Si
Facilidad de uso	Fácil	Medio	Medio
Licencia	Libre	Libre	Libre para fines educativos
Selección	X		

Los tipos de gráficas a utilizar en el sistema para tener interacción con el usuario son barras. Como se observa en la Tabla 5, todas las herramientas proporcionan ese tipo de gráficas, sin embargo un aspecto importante de selección es que sea fácil de usar, así que Chart.js se elige debido a que cuenta con esta característica.

3.3.2. Funciones

Las funciones a desarrollar a lo largo de la implementación para obtener el sistema son las siguientes:

1) **Pre-procesamiento**: Este módulo es el encargado de normalizar los comentarios, quitar signos de puntuación y realizar una separación en *tokens* para que sean etiquetados con su respectiva función que ocupa en la oración.
2) **Módulo gramatical**: En este módulo se buscan las estructuras gramaticales que ayudan a determinar que hay una posible oración útil para el análisis.

3) **Extracción de aspectos**: Se utilizarán resultados del módulo gramatical para extraer sustantivos de las oraciones que son útiles para el análisis. A su vez los adjetivos se extraen como sentimientos que describen el aspecto.
4) **Asignación de polaridad**: Utilizando las listas de palabras clasificadas como positivas o negativas, se asigna una polaridad a cada uno de los sentimientos extraídos.
5) **Clasificación de aspectos**: Una vez que se encontraron los aspectos, se obtiene la categoría a la que pertenece. El objetivo de esto es determinar cuál es el aspecto implícito que están describiendo las características que califica el autor del comentario, para que en futuros módulos estos resultados ayuden a la generación de estadísticas agrupando los aspectos.
6) **Almacenamiento en la base de datos**: Los comentarios, sentencias, sentimientos, categorías y aspectos son datos importantes. Se espera hasta el término del análisis de todos los comentarios para dar resultados al usuario, mientras tanto los datos no pueden perderse y por eso deben guardarse en una base de datos.
7) **Generador de estadísticas**: Una vez que el análisis termina, este módulo realiza consultas a la base de datos y se generan gráficas con herramientas que facilitan esta tarea, para así mostrarlas en la página Web y el usuario del sistema observe los resultados del análisis de los comentarios contenidos en el archivo que dio como entrada.

Capítulo 4. Diseño

De acuerdo con el análisis realizado en la sección anterior, se llevó a cabo el diseño del sistema, para ello se siguió el estándar UML 2.0 descrito en el libro "El lenguaje unificado de modelado". [3]

4.1. Diagramas de casos de uso

A continuación se muestra el diagrama de casos de uso (ver Figura 5), así como la especificación de cada uno (ver Tablas 6-16).

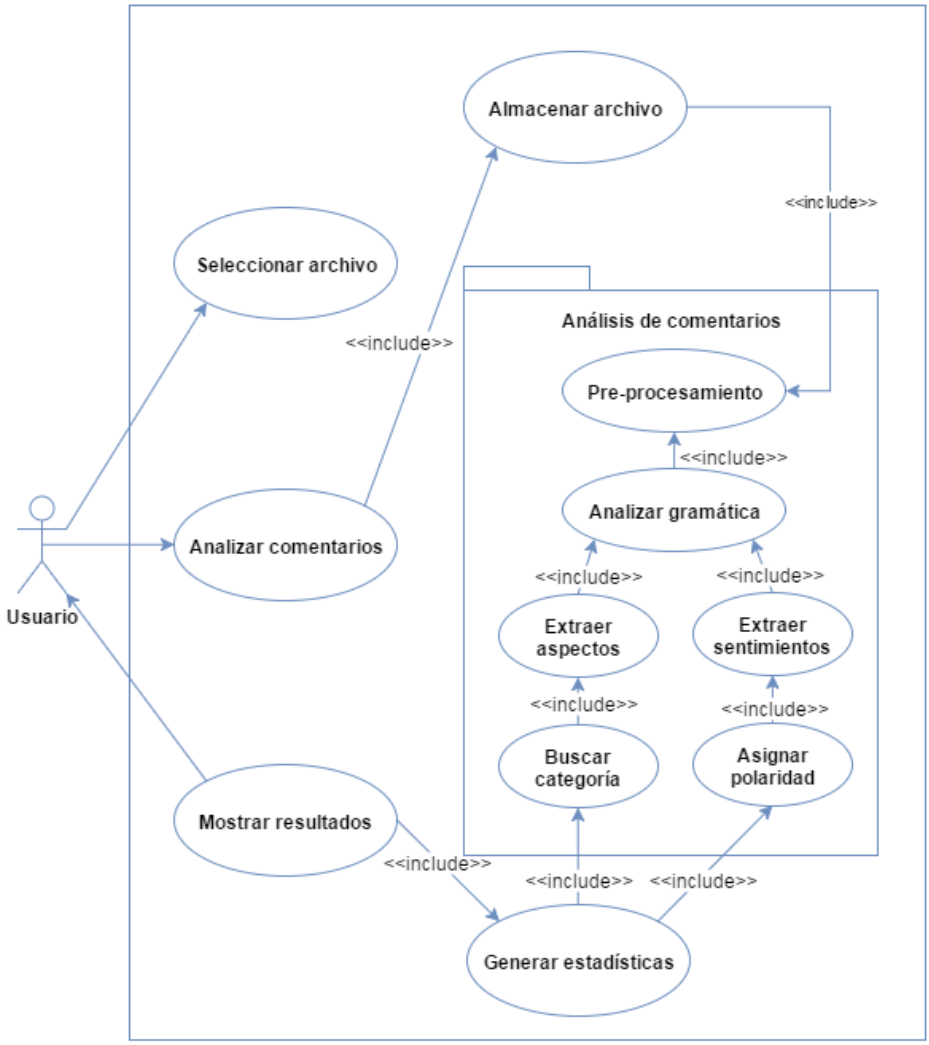

Figura 5. Diagrama de casos de uso del sistema.

Tabla 6. Especificación del caso de uso seleccionar archivo.

Caso de uso	Seleccionar archivo
Actor primario	Usuario
Descripción	Seleccionar un archivo para ser almacenado en el servidor.
Disparador	El usuario da clic en botón buscar archivo.
Prioridad	Alta
Pre condiciones	El usuario necesita que el sistema analice los comentarios de su archivo.
Post condiciones	Mostrar una ventana para que el usuario navegue en su directorio.
Flujo normal	1. El usuario da clic en botón buscar. 2. Se muestra ventana para la navegación del directorio. 3. Se elige un archivo. 4. Se da clic en abrir. 5. Cargar archivo.

Tabla 7. Especificación del caso de uso analizar comentarios.

Caso de uso	Analizar comentarios
Actor primario	Usuario
Descripción	El usuario quiere comenzar con el análisis de los comentarios.
Disparador	El usuario da clic en el botón analizar.
Prioridad	Alta
Pre condiciones	El archivo debe estar almacenado en el servidor.
Post condiciones	Inicia el análisis de comentarios con el módulo de pre-procesamiento.
Flujo normal	1. El usuario da clic en el botón analizar y si se ha seleccionado un archivo comienzan a ejecutarse los siguientes pasos: 1.1. Pre-procesamiento. 1.2. Análisis de gramática. 1.3. Extracción de aspectos y sentimientos. 1.4. Asignación de polaridad y búsqueda de categorías. 2. Si no hay un archivo seleccionado se envía un mensaje de error.

Tabla 8. Especificación del caso de uso almacenar archivo.

Caso de uso	Almacenar archivo
Actor primario	Ninguno
Descripción	Almacenar en el servidor el archivo de comentarios proporcionado por el usuario.
Disparador	El usuario selecciona el archivo a ser analizado
Prioridad	Alta
Pre condiciones	El usuario da clic en botón buscar archivo y selecciona el archivo que desea analizar.
Post condiciones	El archivo ya almacenado se utiliza para comenzar con el análisis de los comentarios.
Flujo normal	1. El usuario selecciona su archivo a cargar. 2. El archivo se carga al servidor.

Tabla 9. Especificación del caso de uso pre-procesamiento.

Caso de uso	Pre-procesamiento
Actor primario	Ninguno
Descripción	Normalizar, quitar signos de puntuación, separación de comentarios en *tokens* y su etiquetado con PoS[4].
Disparador	Caso de uso analizar comentarios.
Prioridad	Alta
Pre condiciones	El archivo de comentarios debe estar almacenado en el servidor.
Post condiciones	Se realiza el análisis gramatical de los comentarios pre-procesados.
Flujo normal	1. Se eliminan los signos de puntuación de todos los comentarios. 2. Se normalizan los caracteres a minúsculas. 3. Cada comentario se fragmenta en *tokens*. 4. A cada *token* le es asignado una etiqueta con su función de la oración.

[4] PoS.- *Part of Speech* (función de las palabras en la oración).

Tabla 10. Especificación del caso de uso analizar gramática.

Caso de uso	Analizar gramática
Actor primario	Ninguno
Descripción	Buscar estructuras gramaticales en los comentarios para cortar sentencias que podrían ser útiles al contener aspectos y sentimientos.
Disparador	Caso de uso analizar comentarios.
Prioridad	Alta
Pre condiciones	El pre-procesamiento de los comentarios fue terminado.
Post condiciones	Extraer aspectos y sentimientos de los comentarios.
Flujo normal	1. Se toma un comentario. 2. Se busca si hay alguna de las estructuras gramaticales previamente seleccionadas. 3. Se hace un corte de la sentencia que por su estructura parece ser útil. 4. Se almacena el comentario y las sentencias en la base de datos. 5. Regresar al paso 1 hasta terminar con todos los comentarios.

Tabla 11. Especificación del caso de uso extraer aspectos.

Caso de uso	Extraer aspectos
Actor primario	Ninguno
Descripción	Una vez detectadas las sentencias útiles se sabe que hay aspectos, así que son extraídos de la sentencia.
Disparador	Caso de uso analizar comentarios.
Prioridad	Alta
Pre condiciones	El análisis gramatical debe haber terminado.
Post condiciones	Búsqueda de la categoría a la cual pertenece.
Flujo normal	1. Tomar una sentencia útil. 2. Extraer el sustantivo de la oración. 3. Almacenar los aspectos en la base de datos. 4. Regresar al paso 1 hasta terminar todas las sentencias.

Tabla 12. Especificación del caso de uso extraer sentimientos.

Caso de uso	Extraer sentimientos
Actor primario	Ninguno
Descripción	Una vez detectadas las sentencias útiles se sabe que hay sentimientos, así que son extraídos de la sentencia.
Disparador	Caso de uso analizar comentarios.
Prioridad	Alta
Pre condiciones	El análisis gramatical debe haber terminado.
Post condiciones	Asignación de polaridad a cada sentimiento.
Flujo normal	1. Tomar una sentencia útil. 2. Extraer el adjetivo de la oración. 3. Almacenar los sentimientos en la base de datos. 4. Regresar al paso 1 hasta terminar todas las sentencias.

Tabla 13. Especificación del caso de uso buscar categoría.

Caso de uso	Buscar categoría
Actor primario	Ninguno
Descripción	Buscar la categoría a la cual pertenece cada aspecto.
Disparador	Caso de uso analizar comentarios.
Prioridad	Alta
Pre condiciones	Todos los aspectos deben estar en la base de datos.
Post condiciones	Generar estadísticas, haciendo consultas a la base de datos.
Flujo normal	1. Se toma un aspecto de la base de datos. 2. Se navega la ontología para encontrarlo. 3. Al ser igual con algún concepto se busca quien es el concepto padre que lo agrupa con otros aspectos. 4. Se almacena el concepto padre (categoría) en la base de datos. 5. Se busca el concepto padre del concepto padre el concepto 6. Si no existe el concepto padre en la base de datos como una categoría se almacena. 7. Se repiten los pasos 5 y 6 hasta llegar al concepto más general que agrupa a todos los aspectos. 8. Regresar al paso 1 hasta terminar con todos los aspectos en la base de datos.

Tabla 14. Especificación del caso de uso asignar polaridad.

Caso de uso	Asignar polaridad
Actor primario	Ninguno
Descripción	Asignar polaridad a cada sentimiento encontrado.
Disparador	Caso de uso analizar comentarios.
Prioridad	Alta
Pre condiciones	Todos los sentimientos deben estar en la base de datos.
Post condiciones	Generar estadísticas, haciendo consultas a la base de datos.
Flujo normal	1. Se busca un sentimiento en la base de datos. 2. Se busca en las listas de palabras positivas y negativas. 2.1. Si está en la lista de palabras positivas, tendrá una polaridad positiva 2.2. Si está en la lista de palabras negativas, tendrá una polaridad negativa 3. Se almacena la polaridad del sentimiento en la base de datos. 4. Regresar al paso 1 hasta terminar con todos los sentimientos de la base de datos.

Tabla 15. Tabla Especificación del caso de uso generar estadísticas.

Caso de uso	Generar estadísticas
Actor primario	Ninguno
Descripción	Consultas a la base de datos para contar la cantidad de aspectos que son positivos o negativos de acuerdo con las categorías que los agrupan.
Disparador	Caso de uso analizar comentarios.
Prioridad	Alta
Pre condiciones	Tener todos los datos en la base de datos. Los comentarios, las sentencias útiles, los aspectos con categoría asignada y los sentimientos con polaridad asignada.
Post condiciones	Generar gráficas que presenten los resultados al usuario.
Flujo normal	1. Realizar consultas a la base de datos. 1.1. Cuando el aspecto es del auto en general. 1.2. Cuando el aspecto es de una parte del auto.

Tabla 16. Tabla Especificación del caso de uso mostrar resultados.

Caso de uso	Mostrar resultados
Actor primario	Ninguno
Descripción	Generar gráficas para la visualización de los resultados del análisis de los comentarios en la página Web.
Disparador	Caso de uso analizar comentarios.
Prioridad	Alta
Pre condiciones	Las consultas de las estadísticas ya están creadas y están arrojando los resultados.
Post condiciones	El usuario interactúa con la página para observar los resultados.
Flujo normal	1. Se toman los resultados de las consultas a la base de datos. 2. Se leen y almacenan temporalmente. 3. Se utiliza la herramienta seleccionada para graficar en la página Web.

4.2. Diagramas de actividades

En esta sección se describen los diagramas de actividades que muestran el flujo de los módulos que conforman la aplicación Web.

En la Figura 6 se encuentra el diagrama de actividades que describe el flujo general del sistema.

En la Figura 7 se muestra el diagrama de actividades del pre-procesamiento y el análisis lingüístico para extraer los aspectos y sentimientos de los comentarios.

El diagrama de actividades para la asignación de categorías se muestra en la Figura 8.

Y finalmente en la Figura 9 se describe el diagrama de actividades para la generación de gráficas, las cuales muestran los resultados del análisis al usuario.

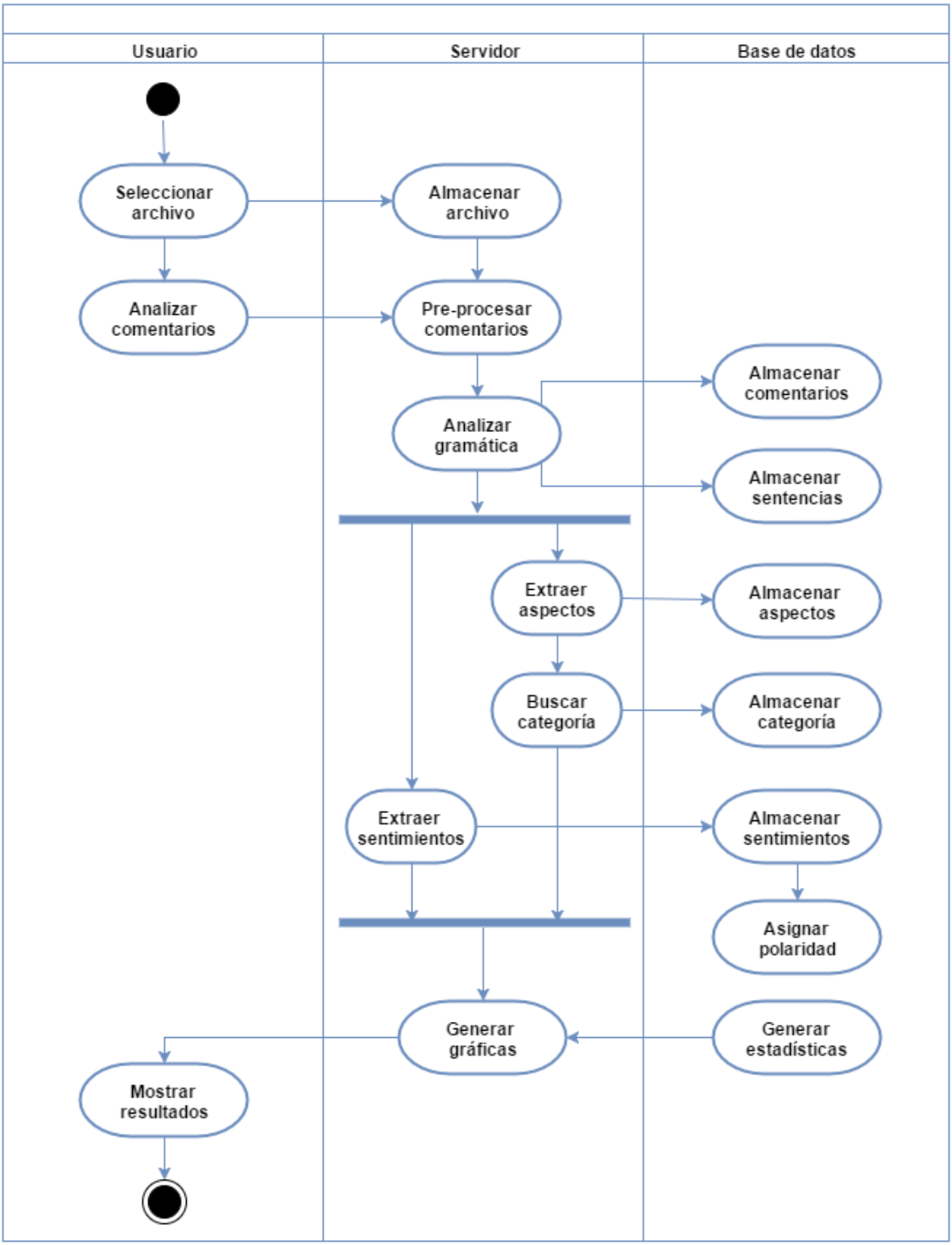

Figura 6. Diagrama de actividades del sistema.

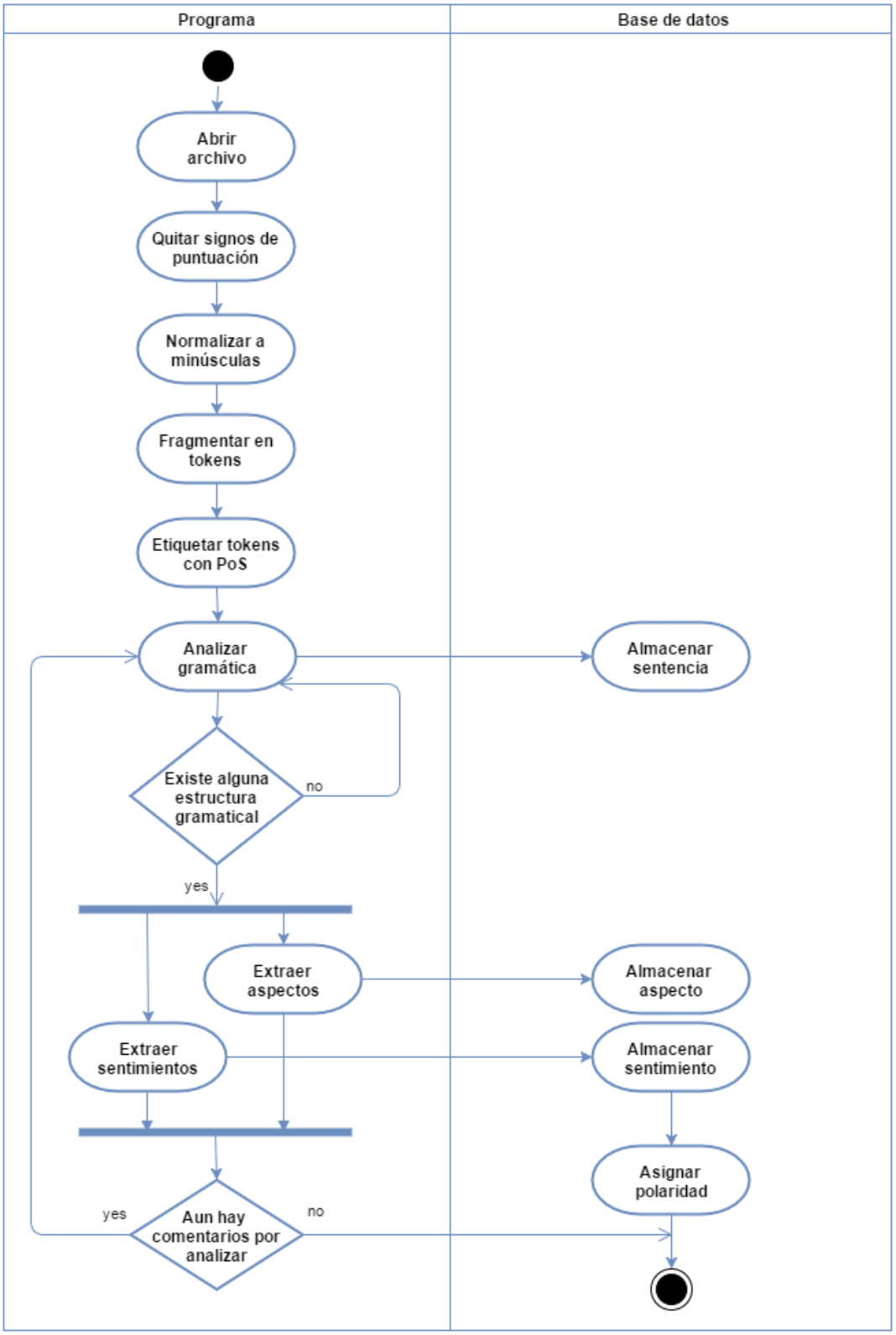

Figura 7. Diagrama de actividades del análisis lingüístico.

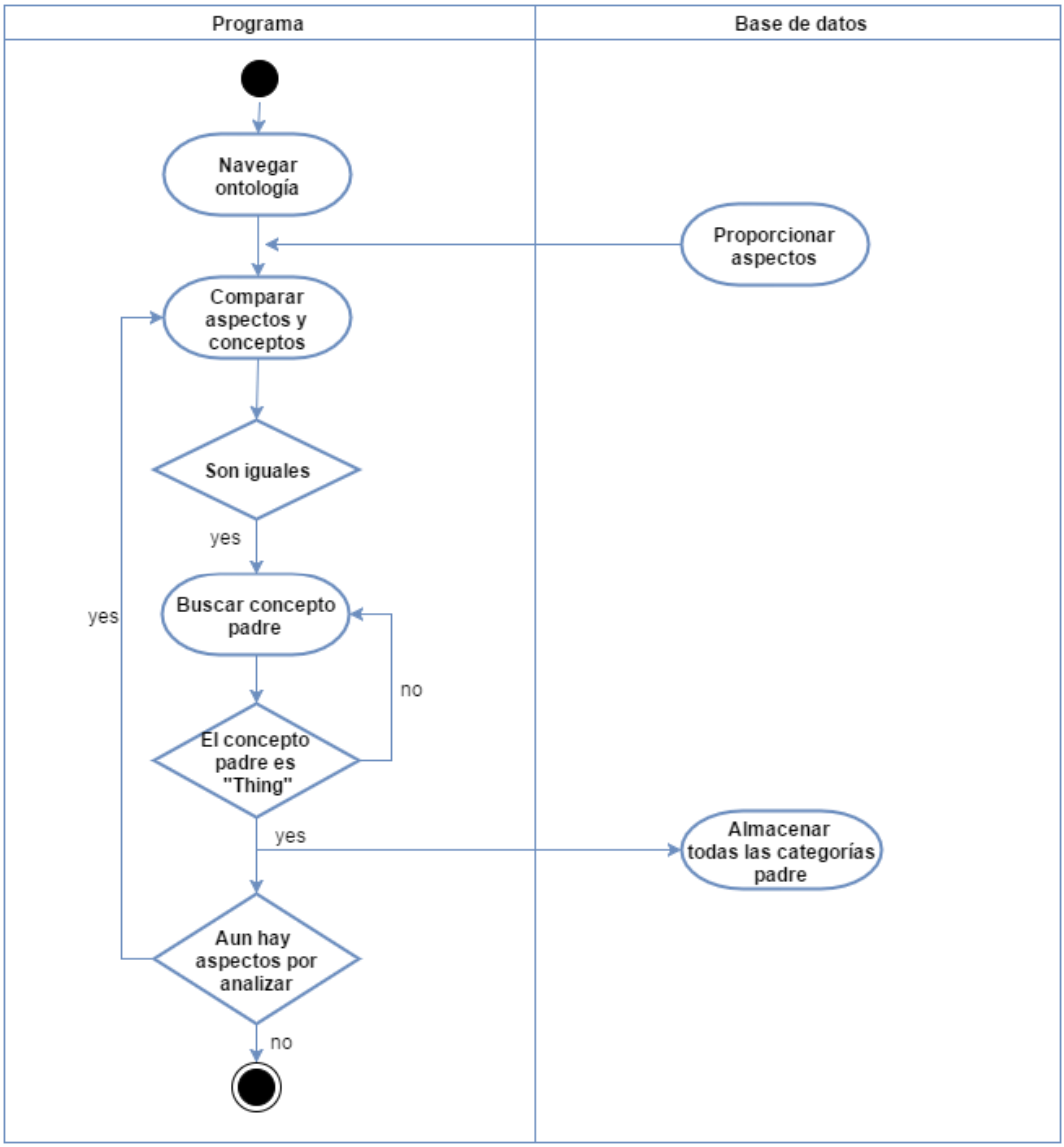

Figura 8. Diagrama de actividades para la asignación de categoría.

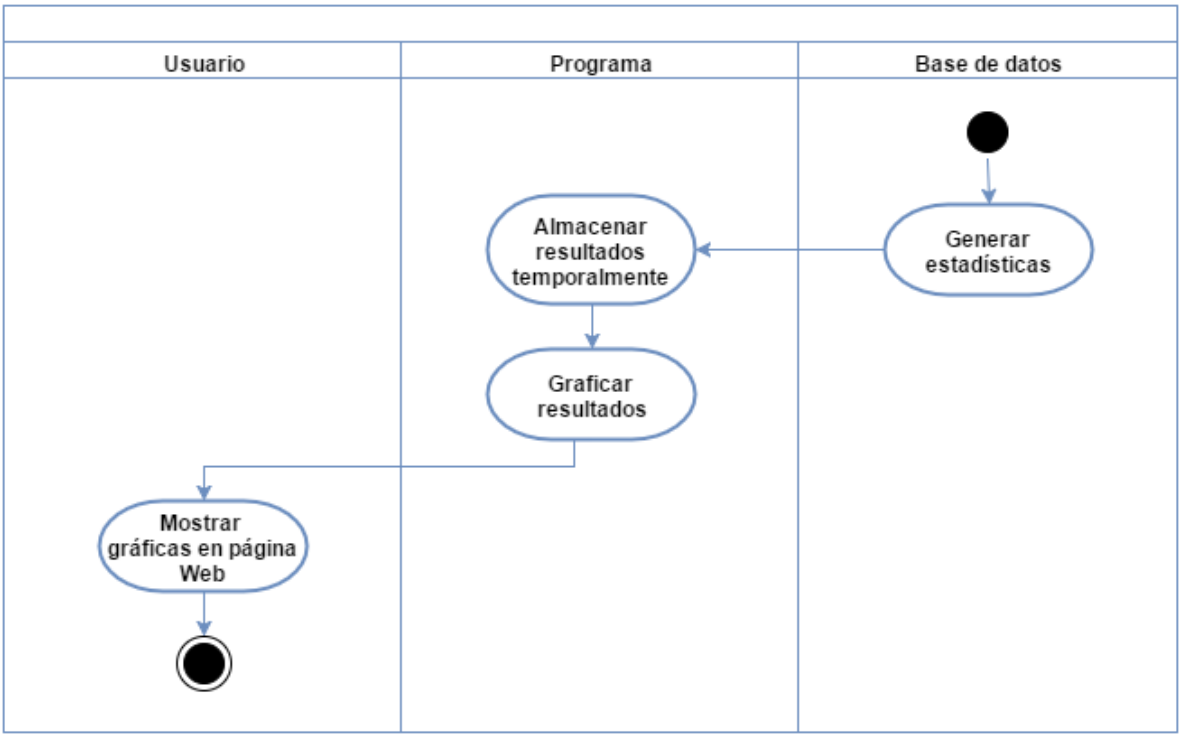

Figura 9. Diagrama de actividades para mostrar las gráficas.

4.3. Diagramas de secuencia

En la Figura 10 se observa el flujo general del sistema, para enfatizar la interacción de los objetos que participan y el orden cronológico en el que ocurren las tareas.

En la Figura 11 se muestra el diagrama de secuencia de las tareas a realizar para el análisis lingüístico. En este diagrama se incluyen tareas del pre-procesamiento como quitar signos de puntuación, normalización de los comentarios a letras minúsculas, la fragmentación de los comentarios en *tokens* y el etiquetado de *tokens*.

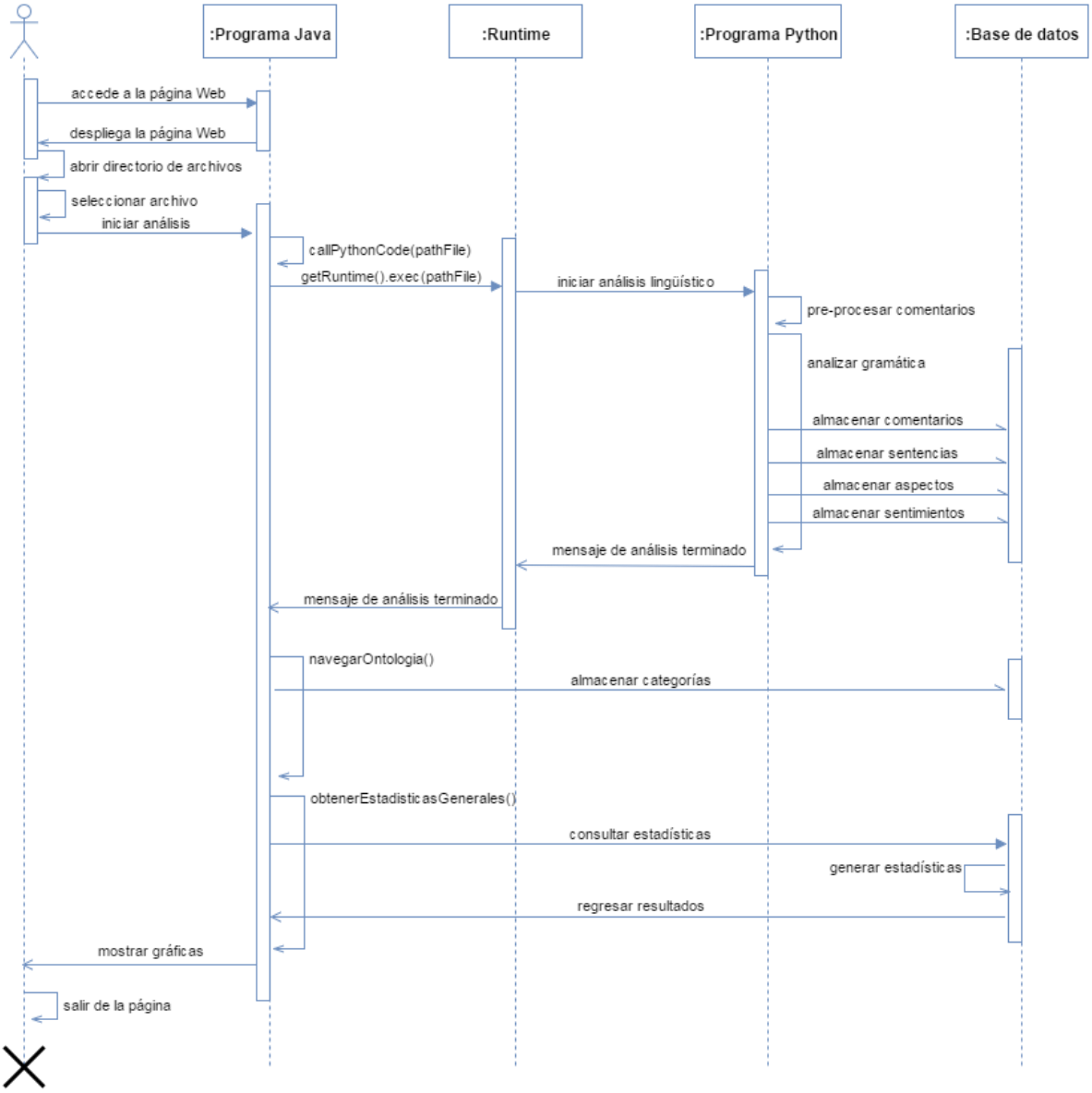

Figura 10. Diagrama de secuencia general.

La Figura 12 muestra la interacción entre el programa y la base de datos para llevar a cabo la asignación de categorías de los aspectos previamente almacenados en la base de datos.

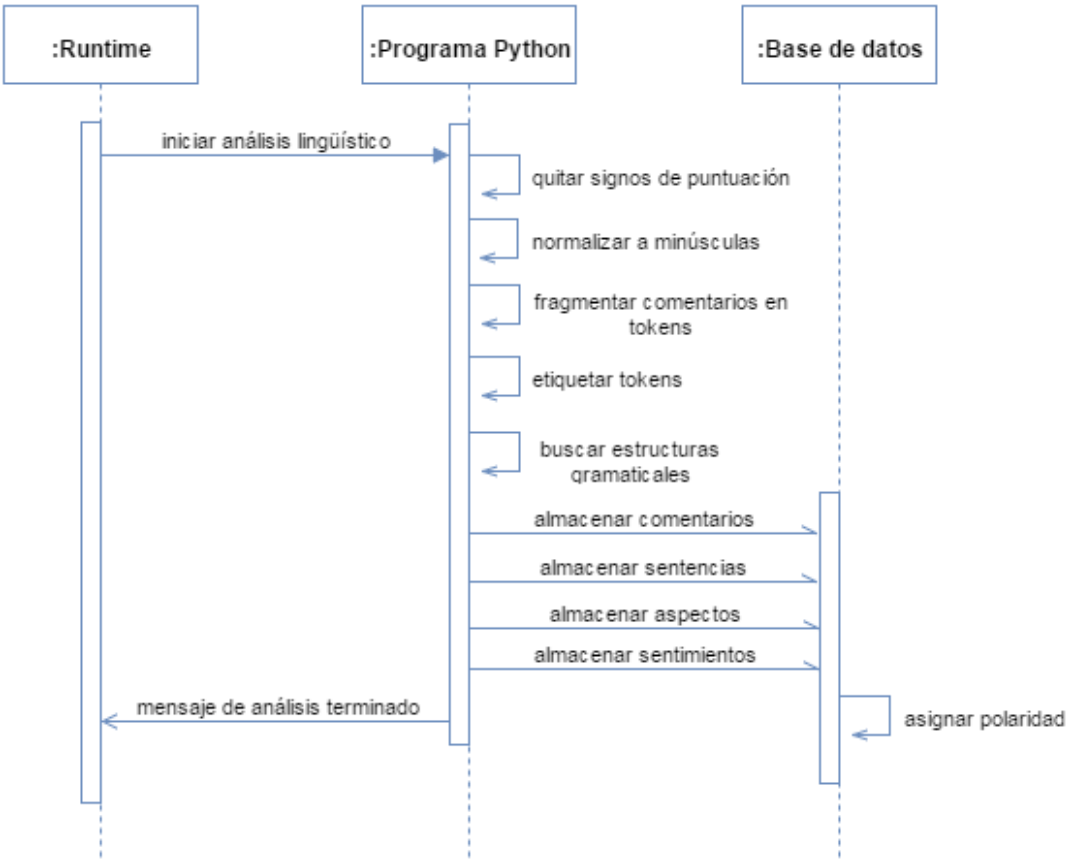

Figura 11. Diagrama de secuencia del análisis lingüístico.

Figura 12. Diagrama de secuencia para la asignación de categorías.

El diagrama de secuencia de la Figura 13 muestra el orden cronológico para obtener las gráficas en la página Web.

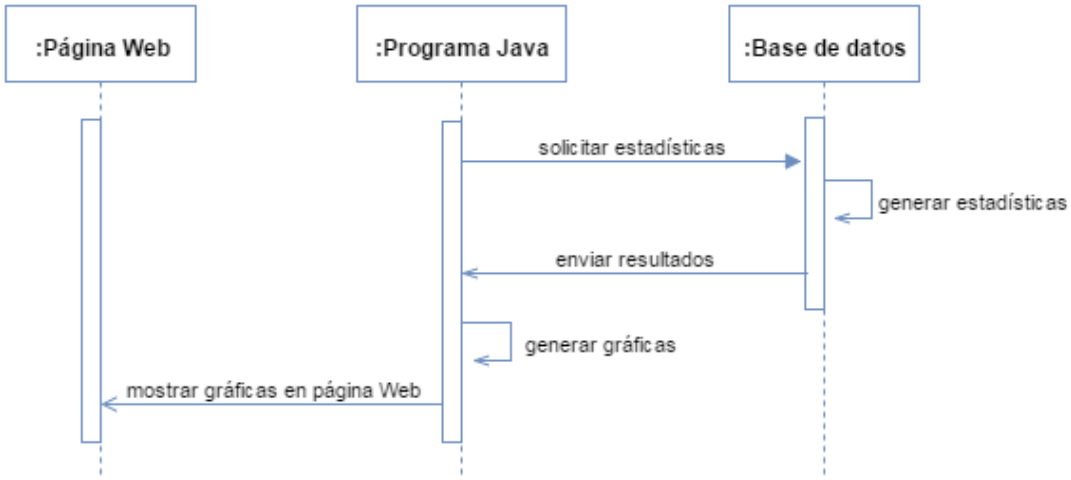

Figura 13. Diagrama de secuencia para generar gráficas.

4.4. Diagrama de clases

Parte del proyecto fue desarrollado con el lenguaje de programación Java, por lo que se utilizó el IDE NetBeans 8.0.2 para facilitar la visualización del código y la implementación de bibliotecas. Con ayuda de dicho IDE se codificaron las clases que se muestran en el diagrama de la Figura 14.

Las clases que terminan en la palabra `Action` son clases que interactúan directamente con los JSP a través de Struts2, para intercambiar información.

`PythonProcess.java` es una clase que contiene el método `callPythonCode()` que llama al código programado en Python para realizar el análisis lingüístico.

Las clases DAO son aquellas que contienen métodos que conocen los nombres de los procedimientos almacenados para hacer consultas a la base de datos.

`Conexion.java` es una clase estática, que sirve para establecer conexión con la base de datos, es la única que contiene datos como el usuario, contraseña y nombre de la base de datos.

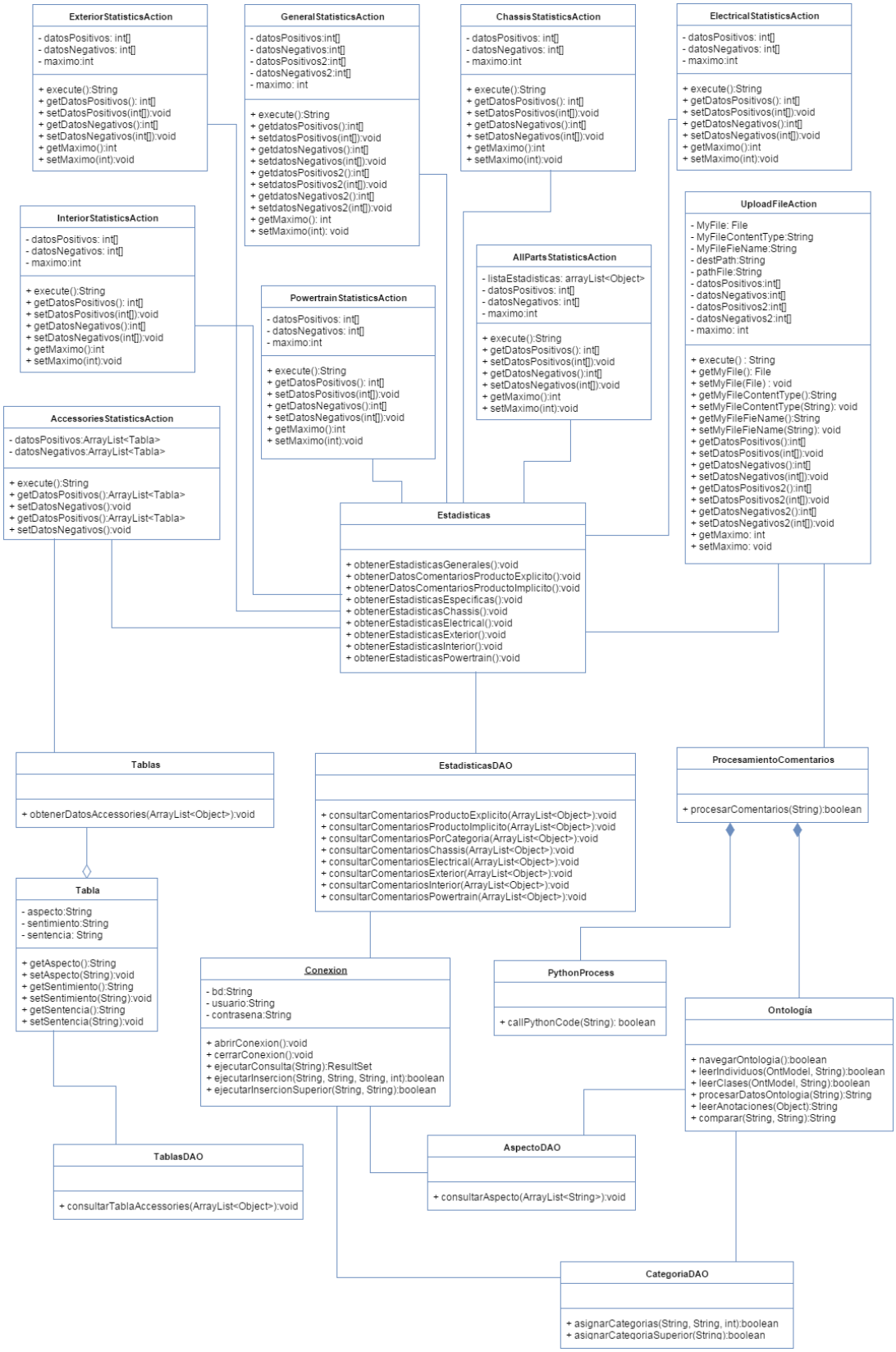

Figura 14. Diagrama de clases.

4.5. Diagramas para base de datos

En las Figura 15 y 16 se muestran los diagramas de entidad-relación y relacional respectivamente, para poder crear la base de datos.

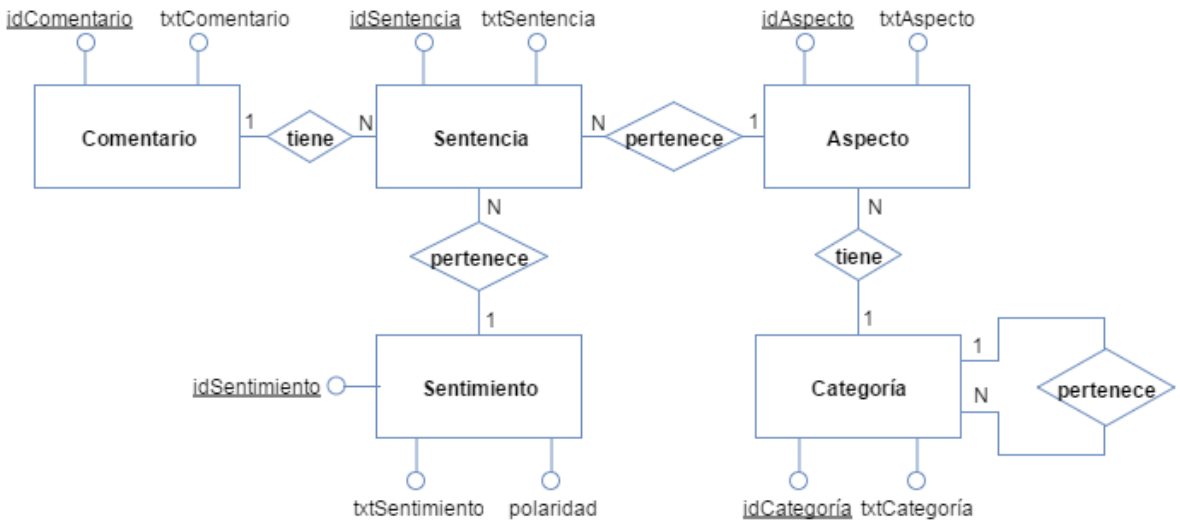

Figura 15. Diagrama entidad-relación de la base de datos.

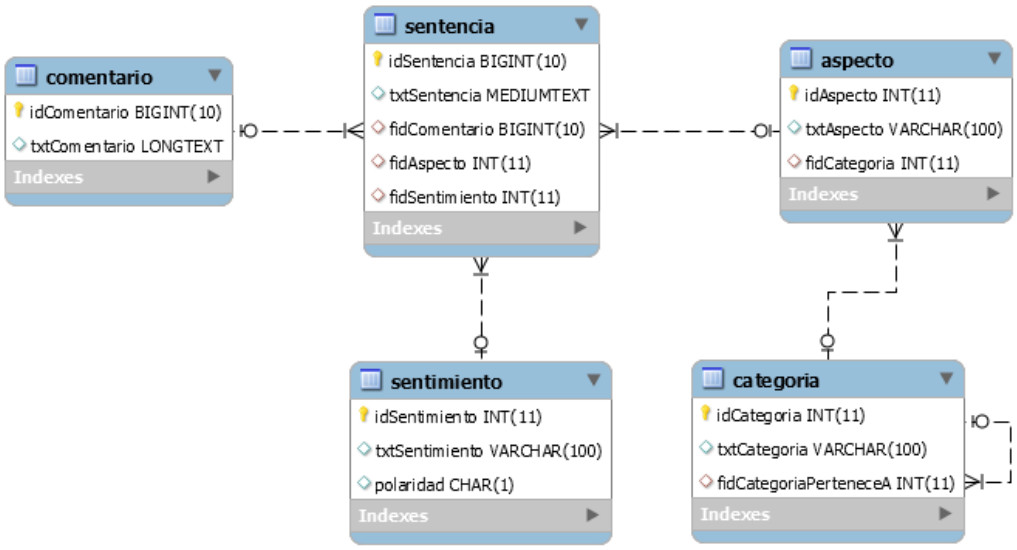

Figura 16. Diagrama relacional de la base de datos.

4.6. Diagrama de la ontología

A continuación se presenta el diseño de la ontología desarrollada para su implementación en la asignación de categorías de los aspectos. De acuerdo con el análisis de la interfaz de software realizado en la sección de análisis, la herramienta para trabajar con la ontología fue *Protégé*, por esta razón los diagramas se realizaron con dicha herramienta. En la Figura 17 se muestra la jerarquía de clases de la ontología.

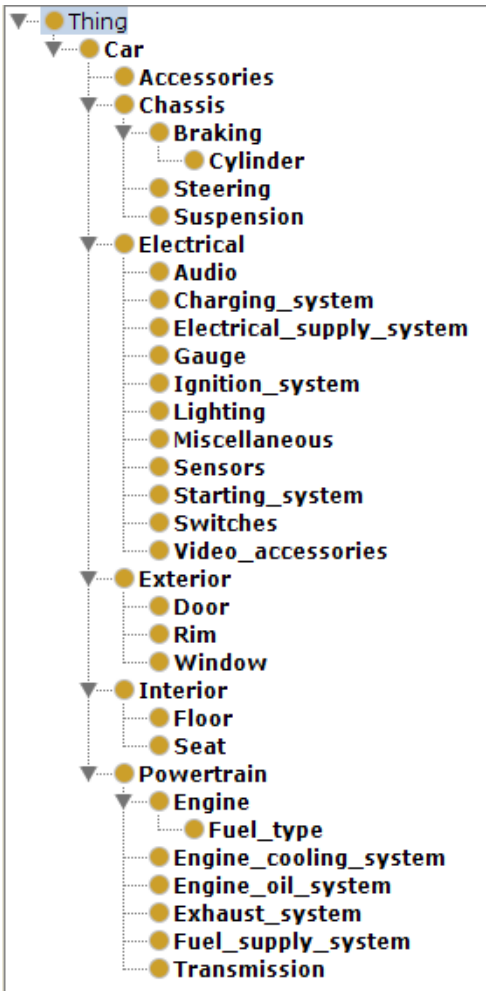

Figura 17. Diseño de la ontología: jerarquía de clases.

En la Figura 18 se muestra el diseño final de la ontología desarrollada para este proyecto. Como se puede apreciar, en el primer nivel se encuentran las clases: *electrical*, *accessories*, *powertrain*, *interior*, *exterior* y *chassis*. Así como los hijos de cada una de estas clases.

Figura 18. Diseño de la ontología.

En la Figura 19 se muestran únicamente los conceptos del primer nivel.

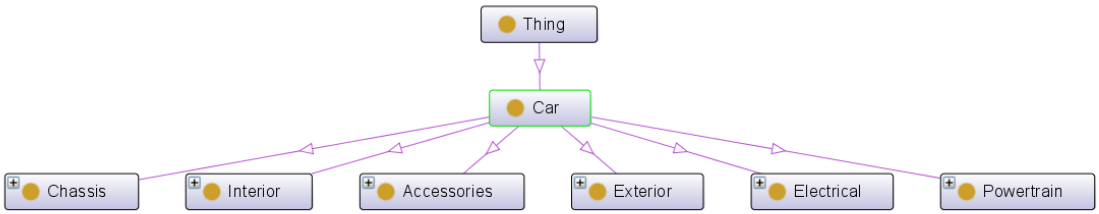

Figura 19. Diseño de la ontología primer nivel.

En la Figura 20 se muestran los conceptos hijo del concepto ***chassis***.

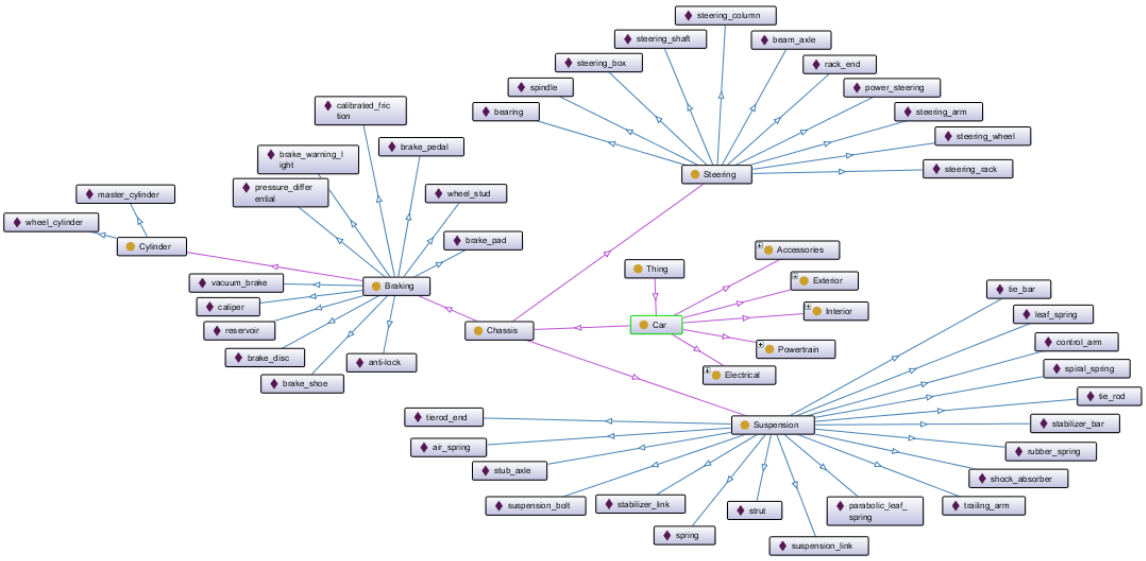

Figura 20. Diseño de la ontología segundo nivel: *chassis*.

En la Figura 21 se muestran los conceptos hijo del concepto ***powertrain***.

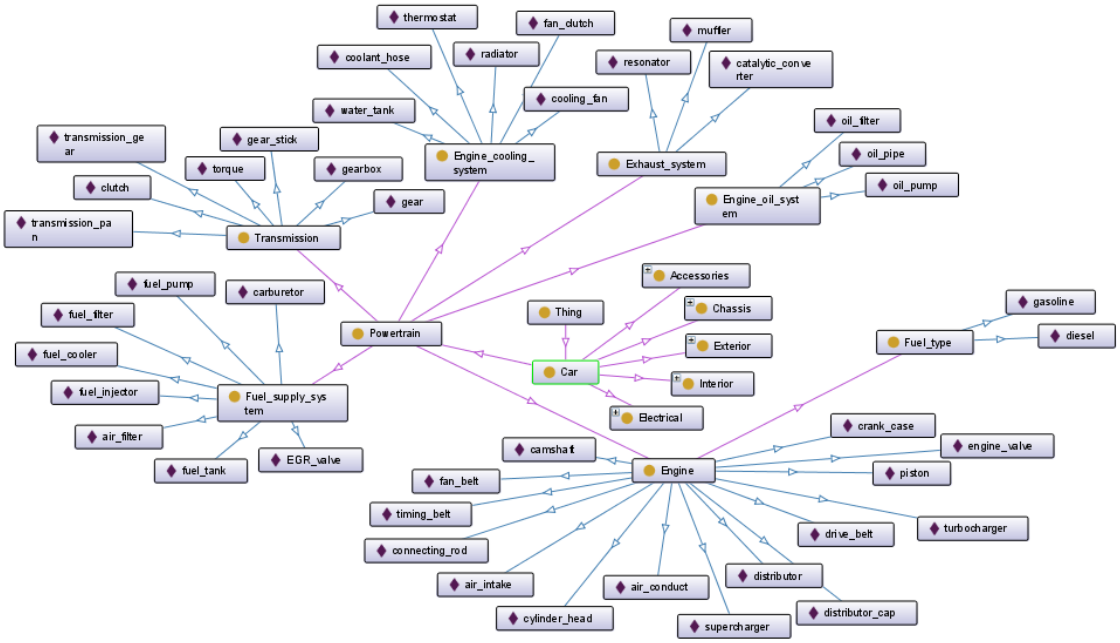

Figura 21. Diseño de la ontología segundo nivel: *powertrain*.

En la Figura 22 se muestran los conceptos hijo del concepto ***interior***.

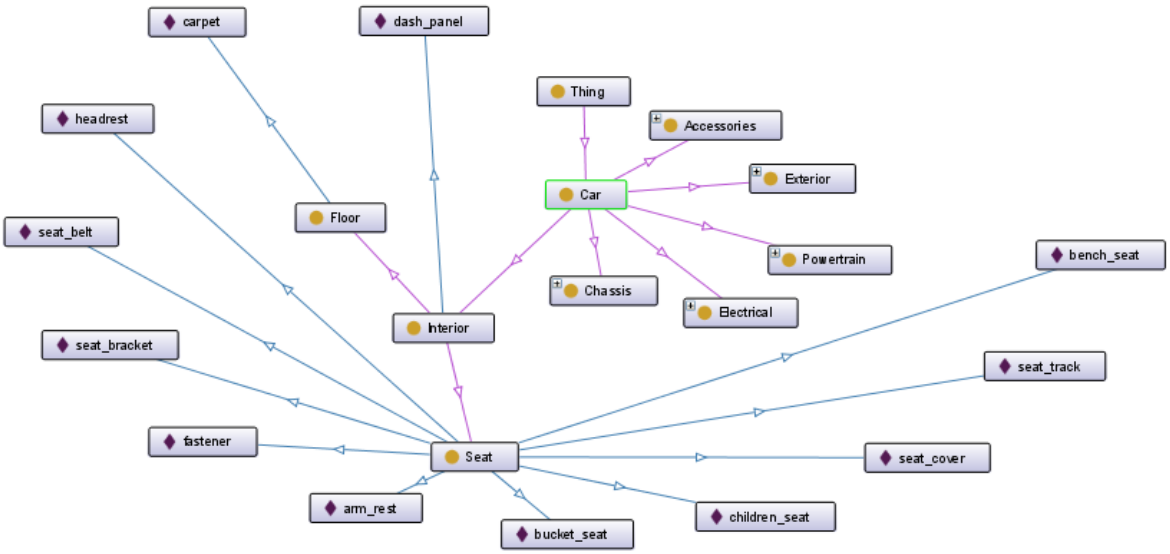

Figura 22. Diseño de la ontología segundo nivel: *interior*.

En la Figura 23 se muestran los conceptos hijo del concepto ***electrical***.

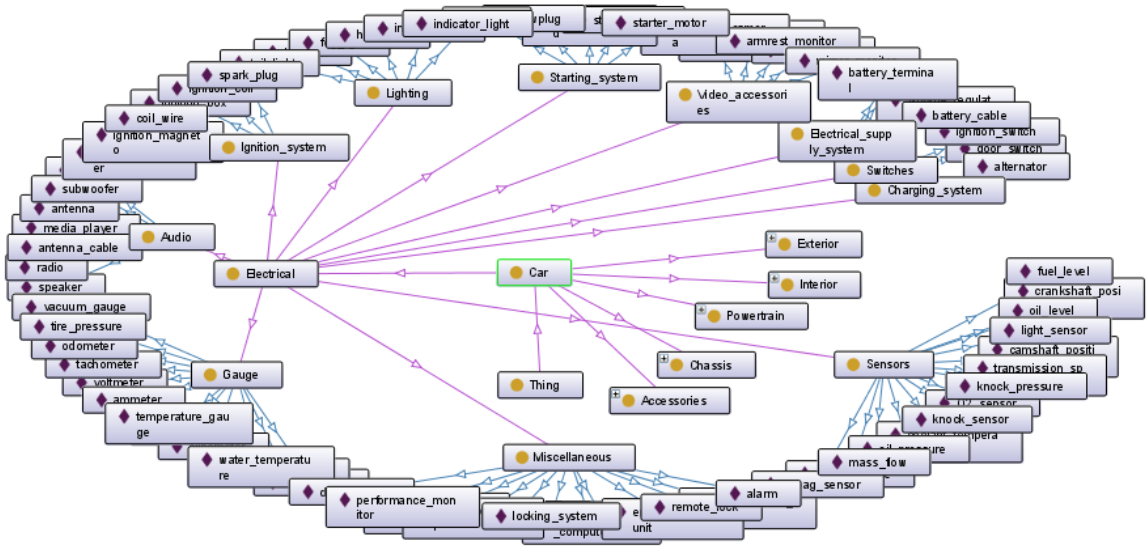

Figura 23. Diseño de la ontología segundo nivel: *electrical*.

En la Figura 24 se muestran los conceptos hijo del concepto ***accessories***.

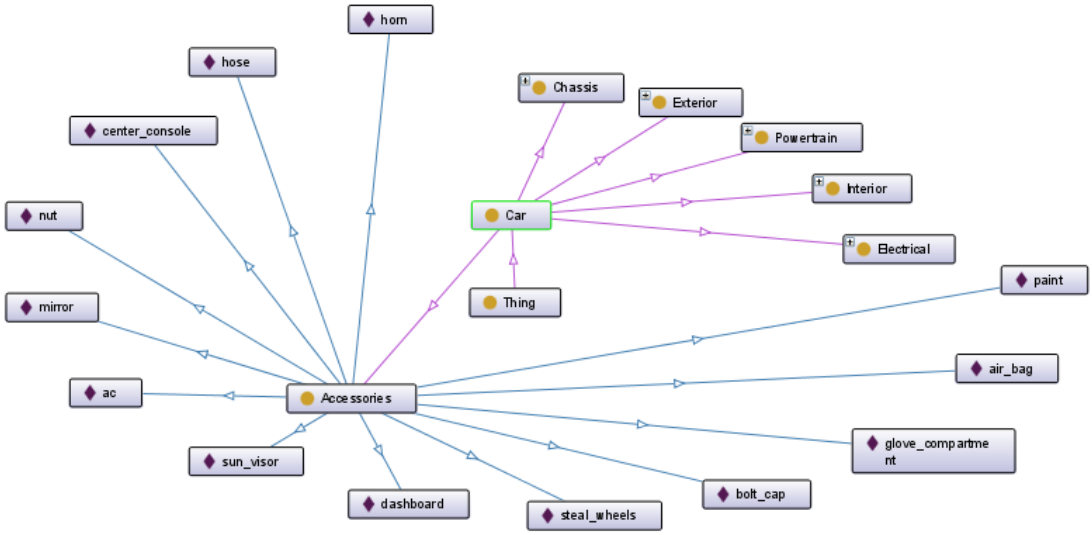

Figura 24. Diseño de la ontología segundo nivel: *accessories*.

En la Figura 25 se muestran los conceptos hijo del concepto ***exterior***.

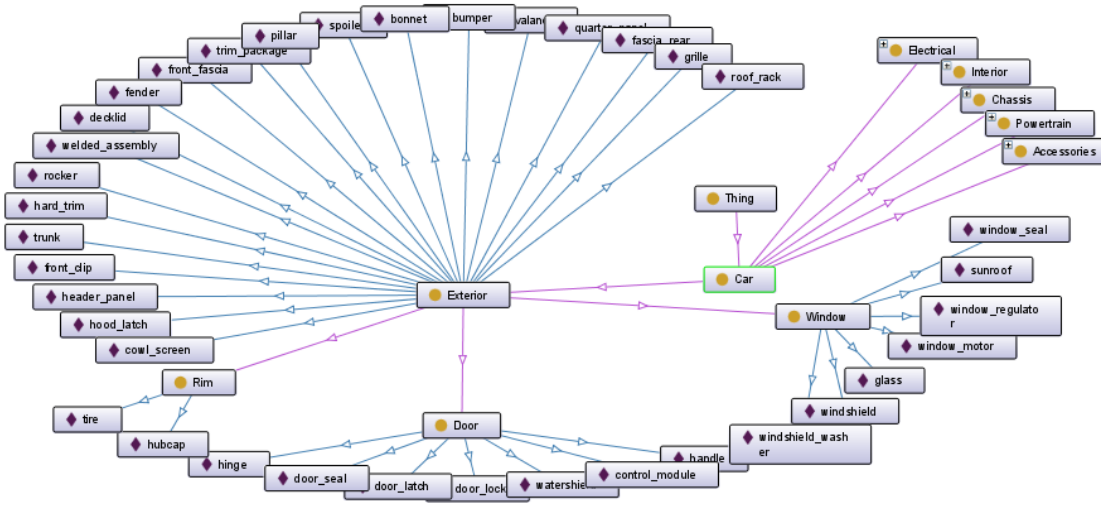

Figura 25. Diseño de la ontología segundo nivel: *exterior*.

4.7. Interfaz de usuario

En las siguientes figuras se muestran las capturas de pantalla de la interfaz gráfica con la que interactúa el usuario.

En la Figura 26 se observa la página inicial del sistema. Cuenta con dos botones, uno para buscar el archivo y otro para comenzar con el análisis.

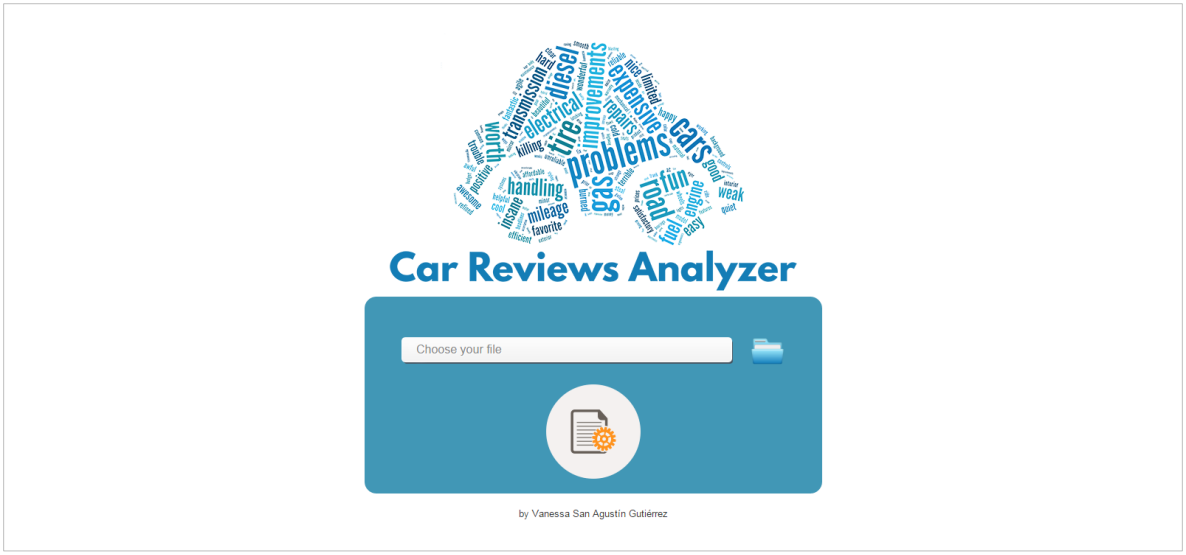

Figura 26. Página inicial del sistema.

En la Figura 27 se muestra la página que contiene los resultados, en ella aparecen un par de gráficas de barras donde se observa la cantidad de aspectos que hablan positiva y negativamente del automóvil en general, así como los que hablan de alguna parte en especial.

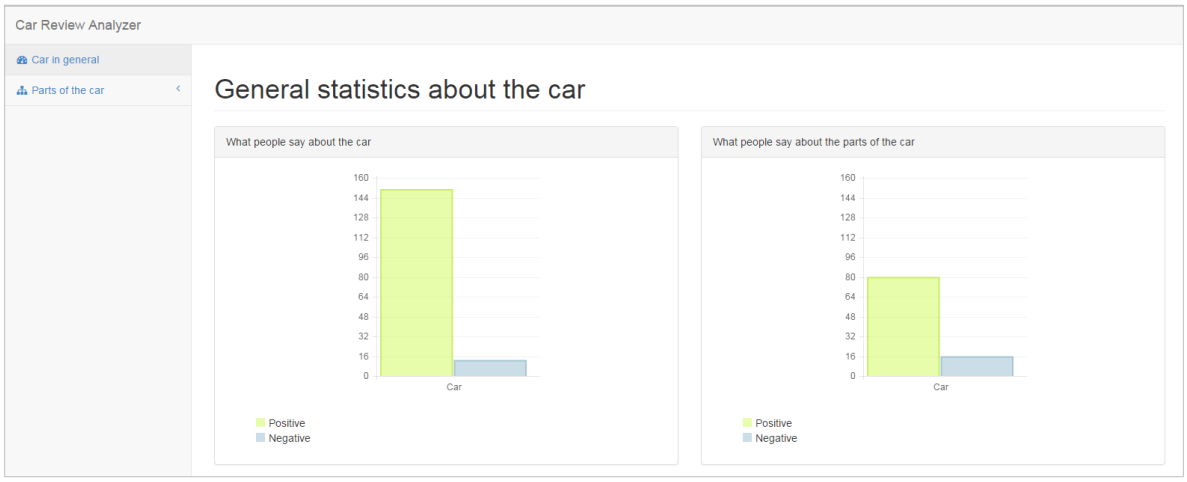

Figura 27. Página de resultados sobre el automóvil en general.

En la Figura 28 se encuentra otra gráfica, pero esta vez, los aspectos se encuentran agrupados en sus respectivas categorías.

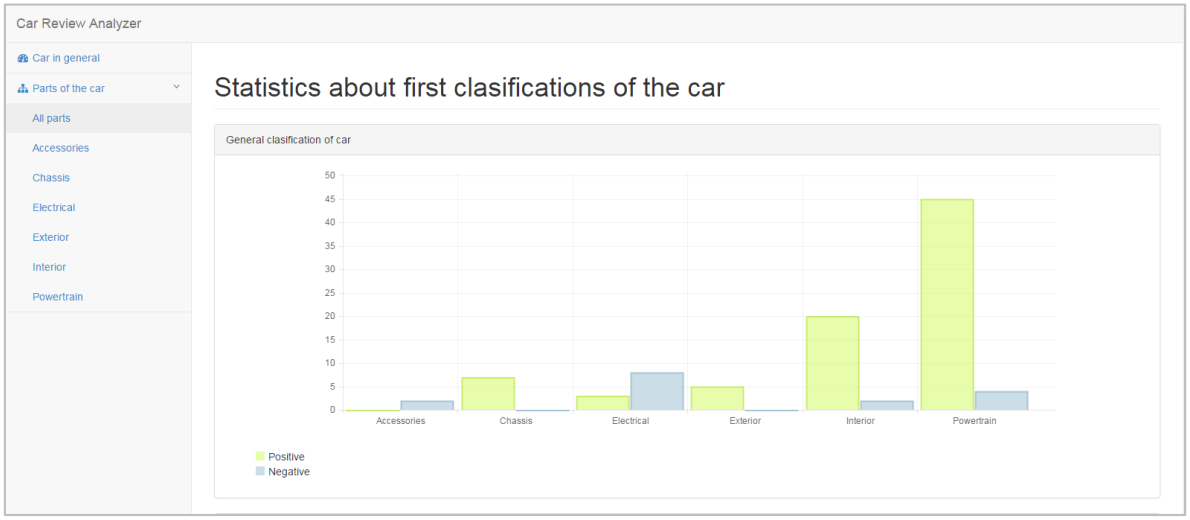

Figura 28. Página de resultados de aspectos agrupados.

Capítulo 5. Implementación

El sistema fue desarrollado en varios módulos los cuales se mencionan a continuación, con imágenes, diagramas y fragmentos de código.

5.1. Extracción automática de aspectos y sentimientos

El lenguaje natural es tan normal en nuestra vida cotidiana que el reto en esta área es que las computadoras puedan entenderlo. El gran reto en este proyecto es que los comentarios que se desean analizar, son reseñas de personas que tienen o han tenido el automóvil que describen, expresando las emociones que ese producto les provocó así como los problemas que tuvieron. Un ejemplo de estos comentarios es el siguiente:

"I bought my Jetta S 2.5 a few months ago and i am still i love the only issues i have with it are the drivers door sticks a little and the steal wheels but that is to be expected with the basic model, it is so much fun to drive and it looks great. also the ac is a little weak but i was told that is common with volkwagens. i plan to buy another when i pay this one off"

Este tipo de comentarios o "reseñas", son muy útiles cuando se desea tener información de la experiencia de otras personas y sobre todo con productos tan costosos como un auto, pues no es tan fácil comprar uno para probarlo, después regresarlo y comprar otro. Puede ser que el auto esté equipado con el mejor motor y el diseño de la carrocería sea el de moda o en general ser el mejor auto (según la empresa) pero si una persona opina que las bolsas de aire no funcionan y muchas más que el estéreo está defectuoso, el futuro comprador dudaría en realizar la compra.

Dicho lo anterior es muy importante saber qué parte del automóvil en especial fue de agrado o no para los usuarios. Como parte del reto se tiene que los comentarios no son escritos con buena ortografía, ni una estructura gramatical adecuada, los signos de puntuación no se usan en lugares que corresponden, son textos muy largos que contienen varios sentimientos, algunos sobre características buenas y otros sobre características malas.

Todos los problemas anteriores no son más que características naturales de una reseña pero hay otro problema con el cual lidiar, pues las herramientas actuales del procesamiento del lenguaje natural aun no pueden detectar automáticamente los aspectos (los componentes del auto) que describe una persona a través de sus comentarios.

Para atacar todos estos problemas se utilizó el NLTK para Python 2.7 que proporciona herramientas muy útiles para dar solución al problema general del trabajo, encontrar automáticamente los aspectos y los sentimientos de comentarios.

Primero se llevó a cabo un pre-procesamiento, que se muestra en el Código 1, las funciones utilizadas se encuentran documentadas en el libro "*Natural Language Processing with Python*" [2].

```
#Quitar signos de puntuacion

comentariosSinPuntuacion = [re.sub('[%s]' % re.escape(string.punctuation), ' ',c)
for c in comentarios]

#Normalizar a letras minúsculas

comentariosMin = [c.lower() for c in comentariosSinPuntuacion]

#Fragmentar comentarios en tokens y etiquetarlos

comentariosEtiquetados = [nltk.pos_tag(nltk.word_tokenize(comentario)) for
comentario in comentariosMin]
```

Código 1. Pre-procesamiento.

Ya que no todos los signos de puntuación se encuentran en un lugar correctamente escrito, se decide descartarlos porque no ayudan a realizar una fragmentación en oraciones en las cuales podrían ayudar los puntos y coma, las comas y los puntos. En este paso se pensó en dejar aquellos signos que aportaran alguna emoción (emoticonos) pero se llegó a la conclusión de que las emociones que transmiten no están relacionadas directamente con algún aspecto, así que signos como :) , :(, :'(, etc. fueron descartados.

La normalización a letras minúsculas es necesaria ya que en futuros pasos es indispensable una búsqueda léxica y es mejor que el texto esté preparado para evitar distinciones entre *"the"* y *"The"*.

La siguiente tarea es realizar una *tokenization*, es decir, fragmentar los comentarios en *tokens* para que cada uno de éstos sean etiquetados con un PoS (*Part of Speech*) de la Tabla 17. La función `pos_tag()` ejecuta un etiquetador de máxima entropía que de acuerdo con [6] tiene una efectividad del 97.6% con los corpus que se ha probado. Con estos *tokens* etiquetados se hace la búsqueda de aspectos y sentimientos. El ejemplo de uno de los comentarios al terminar el pre-procesamiento es el siguiente:

```
[('i', 'PRP'), ('bought', 'VBD'), ('my', 'PRP$'), ('jetta', 'NN'),
('s', 'NNS'), ('2.5', 'CD'), ('a', 'DT'), ('few', 'JJ'), ('months',
'NNS'), ('ago', 'RB'), ('and', 'CC'), ('i', 'PRP'), ('am', 'VBP'),
('still', 'RB'), ('i', 'PRP'), ('love', 'VBP'), ('the', 'DT'),
('only', 'JJ'), ('issues', 'NNS'), ('i', 'PRP'), ('have', 'VBP'),
('with', 'IN'), ('it', 'PRP'), ('are', 'VBP'), ('the', 'DT'),
('drivers', 'NNS'), ('door', 'NN'), ('sticks', 'NNS'), ('a', 'DT'),
('little', 'RB'), ('and', 'CC'), ('the', 'DT'), ('steal', 'NN'),
('wheels', 'NNS'), ('but', 'CC'), ('that', 'IN'), ('is', 'VBZ'),
('to', 'TO'), ('be', 'VB'), ('expected', 'VBN'), ('with', 'IN'),
('the', 'DT'), ('basic', 'JJ'), ('model', 'NN'), ('it', 'PRP'), ('is',
'VBZ'), ('so', 'RB'), ('much', 'RB'), ('fun', 'VBN'), ('to', 'TO'),
('drive', 'VB'), ('and', 'CC'), ('it', 'PRP'), ('looks', 'VBZ'),
('great', 'JJ'), ('also', 'RB'), ('the', 'DT'), ('ac', 'NN'), ('is',
'VBZ'), ('a', 'DT'), ('little', 'RB'), ('weak', 'JJ'), ('but', 'CC'),
('i', 'PRP'), ('was', 'VBD'), ('told', 'JJ'), ('that', 'IN'), ('is',
'VBZ'), ('common', 'JJ'), ('with', 'IN'), ('volkwagens', 'NNS'), ('i',
'PRP'), ('plan', 'VBP'), ('to', 'TO'), ('buy', 'VB'), ('another',
'DT'), ('when', 'WRB'), ('i', 'PRP'), ('pay', 'VBP'), ('this', 'DT'),
('one', 'CD'), ('off', 'RP')]
```

Tabla 17. Etiquetas de PoS.

Etiqueta PoS	Descripción	Etiqueta PoS	Descripción
CC	coordinating conjunction	**PRP$**	possessive pronoun
CD	cardinal number	**RB**	adverb
DT	determiner	**RBR**	adverb, comparative
EX	existential there	**RBS**	adverb, superlative
IN	preposition/subordinating conjunction	**TO**	to

Tabla 17. Etiquetas de PoS (continuación).

Etiqueta PoS	Descripción	Etiqueta PoS	Descripción
FW	foreign word	**RP**	particle
JJ	adjective	**UH**	interjection
JJR	adjective, comparative	**VB**	verb, base form
JJS	adjective, superlative	**VBD**	verb, past tense
LS	list marker	**VBG**	verb, gerund/present participle
MD	modal	**VBN**	verb, past participle
NN	noun, singular or mass	**VBP**	verb, sing. present, non-3d
NNS	noun plural	**VBZ**	verb, 3rd person sing. present
NNP	proper noun, singular	**WDT**	wh-determiner
NNPS	proper noun, plural	**WP**	wh-pronoun
PDT	predeterminer	**WP$**	possessive wh-pronoun
POS	possessive ending	**WRB**	wh-abverb
PRP	personal pronoun		

Una vez terminadas las funciones de pre-procesamiento se siguió el algoritmo descrito en el diagrama de flujo de la Figura 29. Se observa que independientemente de que si existe alguna sentencia útil se almacenan uno a uno los comentarios. Para ello primero se sustituye la ‘ por \\’ porque hay problemas al insertar el comentario en la base de datos. Después se va evaluando si en el comentario existe alguna sentencia útil; si esto ocurre se almacenan datos como la sentencia, el aspecto y el sentimiento.

La sentencia es un corte del comentario que consta de 11 palabras, en el centro está el verbo *to be*, a la derecha 5 palabras y 5 más a la izquierda. Se realizaron pruebas con sentencias de 5, 7, 9 y 11 palabras, el resultado fue que en promedio la estructura gramatical `sustantivo + verbo (to be) + adjetivo` se encuentra entre las 8 palabras centrales, así que para dar un poco de holgura se utilizan 11 en total. Tal vez sea algo confuso decir que esa estructura se encuentra entre 8 palabras si solo se necesitan a lo

más 4 (dos etiquetadas como sustantivos, una como verbo y una como adjetivo). El problema es que no siempre se encuentra la estructura tal cual porque entre el sustantivo, verbo y adjetivo existen adverbios, artículos y demás que dan pie a estructuras gramaticales más específicas.

"looks great also the ac is a little weak but i"

En la sentencia anterior *is* se encuentra en el centro, a la izquierda está el sustantivo y a la derecha se encuentra el artículo *a* seguido del adverbio *little*. Si se omite el artículo y el adverbio no se pierde el sentido "*ac is weak*". La idea inicial era eliminar todas aquellas palabras que no fueran adjetivos, sustantivos y verbos pero al realizar eso se pierde por completo el sentido del comentario. Por esto es tan importante encontrar sentencias útiles que probablemente contengan algún aspecto y sentimiento.

El almacenamiento de los datos se lleva a cabo a través de procedimientos almacenados por lo que el programa escrito en Python para esta parte del sistema, no conoce el diseño de la base de datos. Todo con el fin de hacer una separación y tener una mejor organización del código.

La asignación de polaridades se realiza mediante la ejecución del procedimiento almacenado `asignarPolaridadesSP()` que se muestra en el Código 2.

```
CREATE PROCEDURE `asignarPolaridadSP`()
BEGIN
START TRANSACTION;

UPDATE analisis_comentarios.sentimiento
INNER JOIN sentimientos_polaridad.sentimientoPolarizado
    ON analisis_comentarios.sentimiento.txtSentimiento =
sentimientos_polaridad.sentimientoPolarizado.palabraSentimiento
SET analisis_comentarios.sentimiento.polaridad =
sentimientos_polaridad.sentimientoPolarizado.polaridadSentimiento;

COMMIT;
END
```

Código 2. Asignación de polaridad.

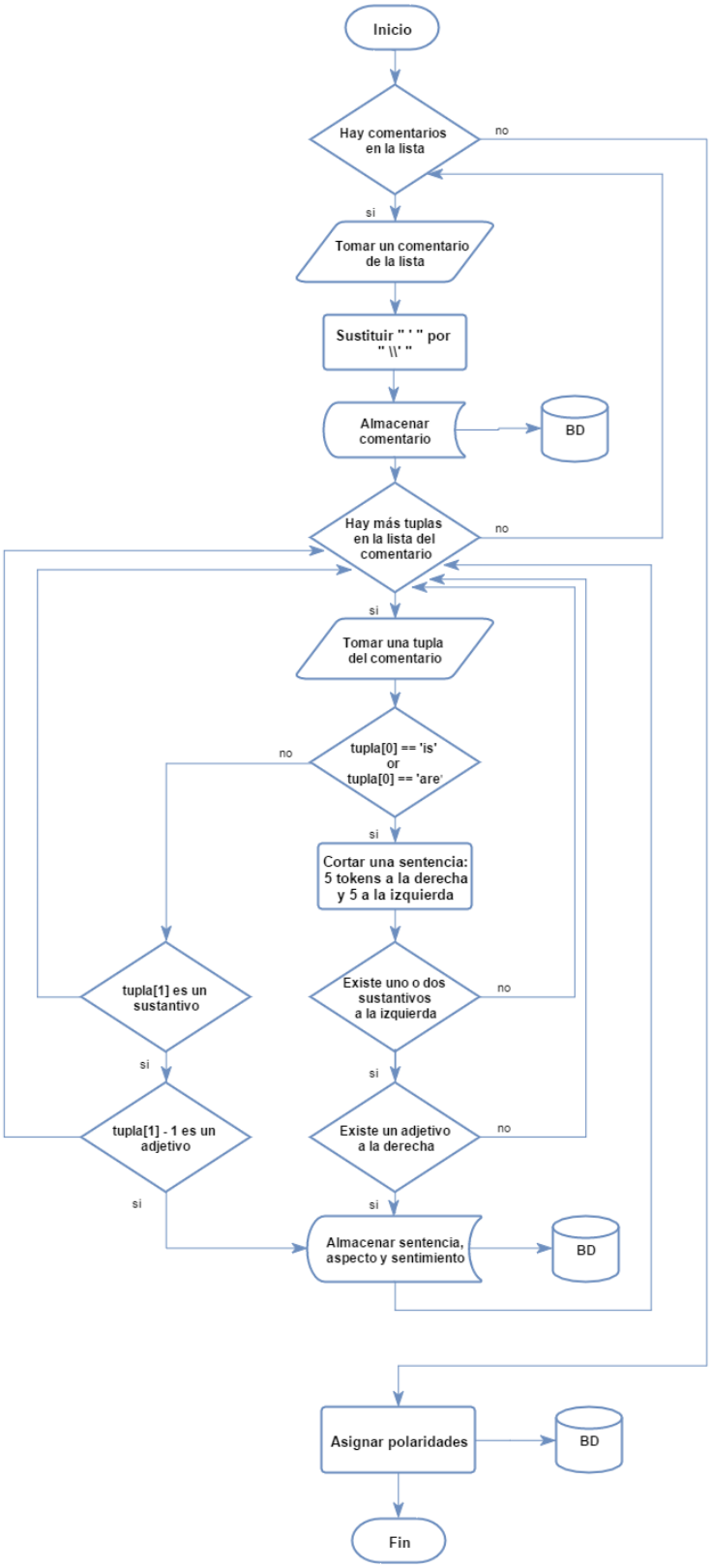

Figura 29. Diagrama de flujo del análisis gramatical.

5.2. Asignación de categorías utilizando la ontología *Car*

Una vez que se encontraron los aspectos es momento de asignar su categoría. El objetivo de que cada aspecto tenga su categoría es que se puedan agrupar. Si una persona conoce sobre autos y realiza su reseña diciendo algo malo sobre la banda de sincronía, probablemente alguien que no conoce que dicha pieza se encuentra en el motor, ignoraría esa información importante.

Como se mencionó anteriormente, el usuario del sistema es o no conocedor del área automotriz, por esta razón se muestran gráficas con diferentes niveles de profundidad. El problema es saber cómo agruparlos, por ejemplo, no se pude decir que el tren motor es malo porque la pintura, el aire acondicionado y los asientos no funcionan. Así que ¿cómo saber qué aspectos se relacionan entre sí? La respuesta que se da a este problema es tener un experto para responder que ciertos aspectos pertenecen a alguna categoría, ese experto en este caso es la ontología *Car*.

En esta tarea es indispensable navegar la ontología para comparar que el aspecto existe en alguno de los conceptos y así preguntar quién es el concepto padre hasta llegar al concepto *Thing*. Por ello se utilizó el *framework* Apache Jena [17], con el cual se buscan las clases e individuos de la ontología para ser comparados con los aspectos que se encuentran en la base de datos. En la comparación si alguno de los aspectos no coincide con algún concepto, no le es asignado ninguna categoría y por lo tanto no se considera en la generación de estadísticas. Estos puntos y la secuencia para encontrar las categorías se muestran en el diagrama de flujo de la Figura 30.

De igual manera que en el programa en Python para el análisis lingüístico, el almacenamiento de datos se hace a través de procedimientos almacenados. A continuación se muestra en el Código 3 la forma en que se almacenan las categorías desde el programa en Java.

```
if(comparar(sinonimos,concepto) != null){
       respuesta = true;
       clasePadre = individuo.getOntClass();
       categoriaDAO.asignarCategorias(procesarDatosOntologia(clasePadre.getLocalNa
       me()), concepto, 0);
```

```
        while(clasePadre.getLocalName().compareTo("Car")!=0){
              temp = procesarDatosOntologia(clasePadre.getLocalName());
              clasePadre = clasePadre.getSuperClass();
              categoriaDAO.asignarCategorias(procesarDatosOntologia(clasePadre.get
              LocalName()),temp,1);
        }
}
```

Código 3. Almacenamiento de categorías desde Java.

Como se observa, la forma de enviar las categorías a la base de datos es una a una, lo cual facilita la codificación en el programa de Java, sin embargo, la complejidad para almacenar los datos correctamente a través del procedimiento almacenado aumenta, como se muestra en el Código 4.

```
CREATE PROCEDURE `asignarCategoriaSP`(IN _conceptoPadre VARCHAR(100), IN
_conceptoHijo VARCHAR(100), IN _bandera INT)
BEGIN
DECLARE idAux INT DEFAULT 0;
START TRANSACTION;

   IF _bandera = 0 THEN
      IF (SELECT idCategoria FROM categoria WHERE txtCategoria = _conceptoPadre)
IS NULL THEN #si el padre no existe
        INSERT INTO categoria (txtCategoria) VALUES (_conceptoPadre);
      END IF;

      UPDATE aspecto SET fidCategoria = (SELECT idCategoria FROM categoria WHERE
txtCategoria = _conceptoPadre)
      WHERE txtAspecto = _conceptoHijo;
   ELSE
      IF (SELECT idCategoria FROM categoria WHERE txtCategoria = _conceptoPadre)
IS NULL THEN #si el padre no existe
        INSERT INTO categoria (txtCategoria) VALUES (_conceptoPadre);
      END IF;

         IF (SELECT idCategoria FROM categoria WHERE txtCategoria =
_conceptoHijo) IS NOT NULL THEN #si el hijo existe
        SELECT idCategoria into idAux FROM categoria WHERE txtCategoria =
_conceptoPadre;
        UPDATE categoria SET fidCategoriaPerteneceA = idAux WHERE txtCategoria =
_conceptoHijo;
         ELSE
        INSERT INTO categoria(txtCategoria, fidCategoriaPerteneceA) VALUES
(_conceptoHijo, (SELECT idCategoria FROM categoria WHERE txtCategoria =
_conceptoPadre));
         END IF;
   END IF;
COMMIT;
END
```

Código 4. Almacenamiento de categorías (procedimiento almacenado)

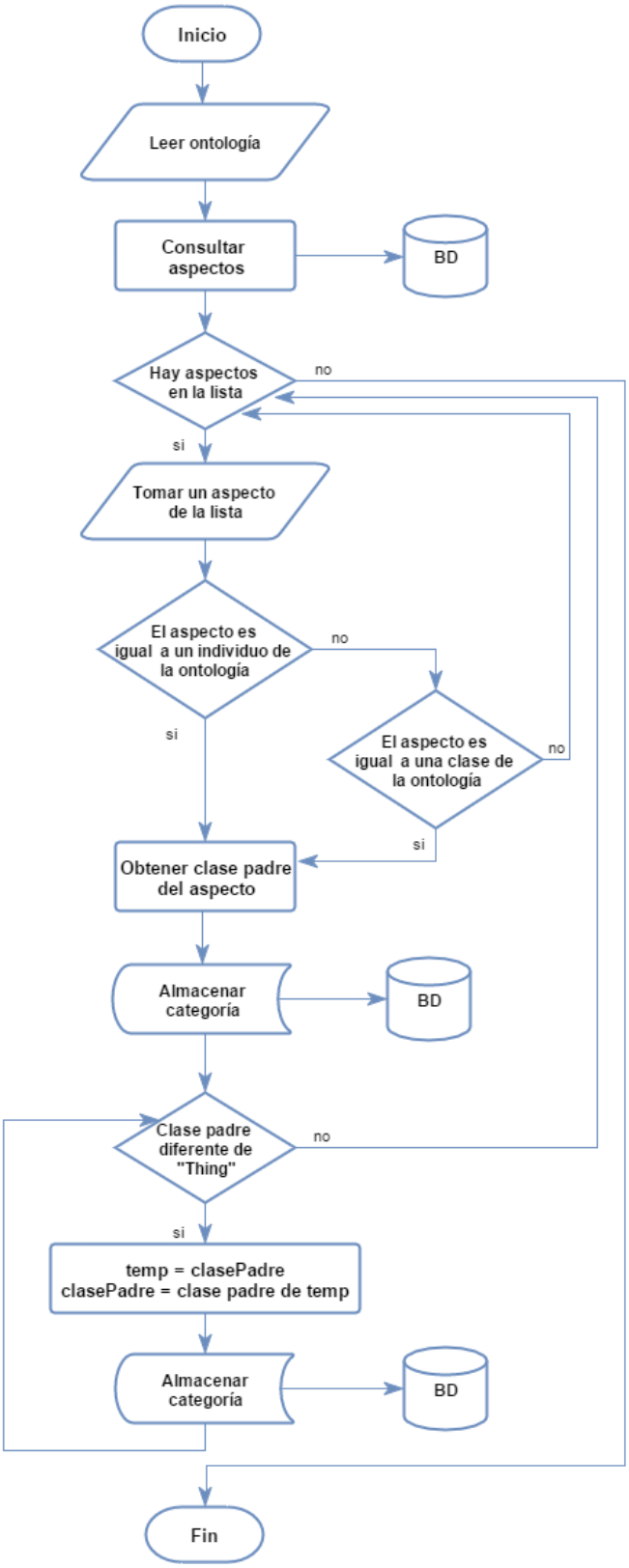

Figura 30. Página de resultados de aspectos agrupados.

5.3. Generación de estadísticas

Las gráficas son una forma visual de presentar los resultados. Son fáciles de interpretar y además existen herramientas que ayudan a presentarlas en páginas Web de una manera sencilla. Para lograr llegar a visualizar los resultados es necesario generar las estadísticas, éstas se obtienen a través de diferentes procedimientos almacenados, uno por cada una de las categorías de segundo nivel y del producto en general. En el Código 5 se muestra el código para generar las estadísticas de primer nivel de detalle en el que se agrupan los aspectos obtenidos del análisis.

```
DECLARE cursorFila CURSOR FOR

       SELECT  idCategoria, polaridad, txtAspecto
       FROM sentencia, aspecto, sentimiento,categoria,comentario
       WHERE fidAspecto = idAspecto AND aspecto.fidCategoria IS NOT NULL
       AND fidSentimiento = idSentimiento AND sentimiento.polaridad IS NOT NULL
       AND fidCategoria = idCategoria AND fidComentario = idComentario
       AND txtCategoria != 'thing';

START TRANSACTION;
OPEN cursorFila;
ciclo: LOOP
       SET contador = contador +1;
       IF contador > numero THEN
              LEAVE ciclo;
       END IF;

FETCH cursorFila INTO cursor_idCategoria,cursor_polaridad,cursor_txtAspecto;
SET idCat = cursor_idCategoria;
SET idCategoriaPadre = idCat;
SET txtCategoriaPadre = null;

SELECT idCategoria INTO idCar FROM Categoria WHERE txtCategoria = 'car';
WHILE (idCategoriaPadre != idCar) DO
       SET idCat= idCategoriaPadre;
       SELECT   fidCategoriaPerteneceA,   txtCategoria   INTO   idCategoriaPadre,
       txtCategoriaPadre FROM categoria WHERE idCategoria = idCat;
END WHILE;

CASE txtCategoriaPadre
       WHEN 'accessories' THEN
              BEGIN
              IF cursor_polaridad = 'p' THEN
                     SET contador_acc_P = contador_acc_P + 1;
              ELSE
                     SET contador_acc_N = contador_acc_N + 1;
              END IF;
       END;

       # … demás conteo de registros de acuerdo a su categoría

END CASE;
END LOOP ciclo;
CLOSE cursorFila;
```

```
SELECT
contador_acc_P,contador_acc_N,contador_cha_P,contador_cha_N,contador_ele_P,
contador_ele_N,contador_ext_P,contador_ext_N,contador_int_P,contador_int_N,
contador_pow_P,contador_pow_N;

COMMIT;
```

Código 5. Generar estadísticas de segundo nivel.

Como se observa, lo que entrega esta consulta son 12 resultados de tipo entero, a cada categoría le corresponde un par, es decir, un resultado de aspectos contados como positivos y para el otro de aspectos negativos. Estos resultados se presentan en gráficas de barras con la herramienta Chart.js, por esto es que en el Código 6 hay un fragmento de cómo crearlas con lenguaje JavaScript.

```
var steps = 10;
var max = maxvalue.innerHTML;
var p = new Array();
var n = new Array();
p = arregloDatosPositivos.innerHTML.split(",");
n = arregloDatosNegativos.innerHTML.split(",");
var canvas = document.getElementById('statisticsSpecific');

var barChartData = {
labels:
["Accessories","Chassis","Electrical","Exterior","Interior","Powertrain"],
datasets: [{
       label: "Positive",
       fillColor: "rgba(208,250,88,0.5)",
       strokeColor: "rgba(191,231,82,0.8)",
       highlightFill: "rgba(208,250,88,0.75)",
       highlightStroke: "rgba(191,231,82,1)",
       data: p
}, {
       label: "Negative",
       fillColor: "rgba(151,187,205,0.5)",
       strokeColor: "rgba(151,187,205,0.8)",
       highlightFill: "rgba(151,187,205,0.75)",
       highlightStroke: "rgba(151,187,205,1)",
       data: n
}]
}
var bar = new Chart(canvas.getContext('2d')).Bar(barChartData,{
       scaleOverride: true,
       scaleSteps: steps,
       scaleStepWidth: Math.ceil(max / steps),
       scaleStartValue: 0
});

document.getElementById('legendSpecific').innerHTML = bar.generateLegend();
```

Código 6. Generación de gráficas con Chart.js.

De los arreglos `arregloDatosPositivos` y `arregloDatosNegativos` se obtienen los resultados a graficar (aquí los resultados de la consulta a la base de datos ya están separados). `Label` son las etiquetas que llevará la gráfica en el eje de las abscisas y las ordenadas se definen tomando el número máximo de los arreglos, se divide entre 10 escalones y se redondea. Después de asignar los datos anteriores así como los colores y datos que deberán graficar, se crean las gráficas en 2D además de la leyenda indicando el color de las barras positivas y las negativas.

Finalmente la gráfica generada con las consultas y el código de JavaScript se muestra en la Figura 31.

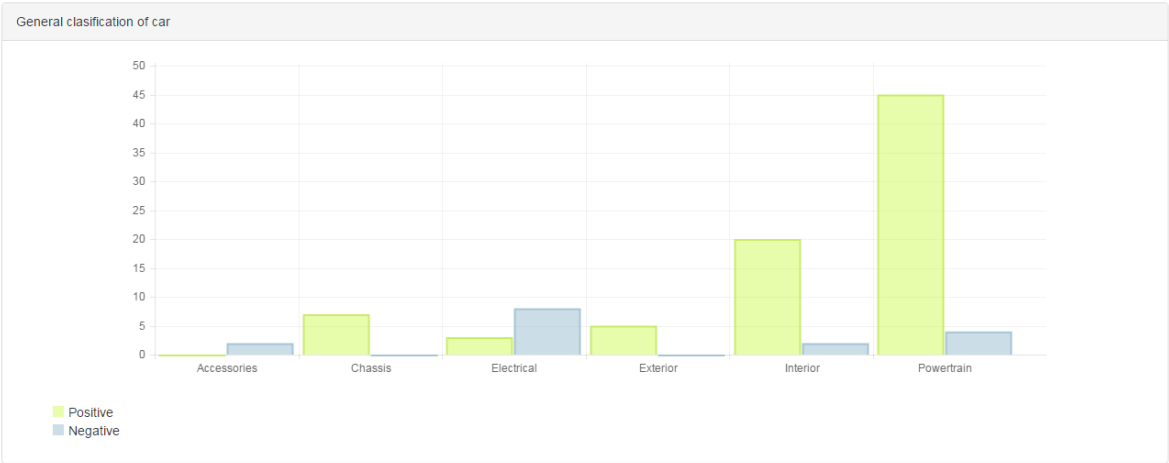

Figura 31. Resultados de aspectos agrupados.

5.4. Arquitectura general del sistema

Una vez que han sido explicados los módulos más sobresalientes, se describe la arquitectura utilizada para el desarrollo del sistema.

Se utiliza una arquitectura Cliente - Servidor de 2 niveles, esto debido a que el servidor responde al cliente sin necesidad de consultar a otro servidor. La Figura 32 muestra la separación física del sistema.

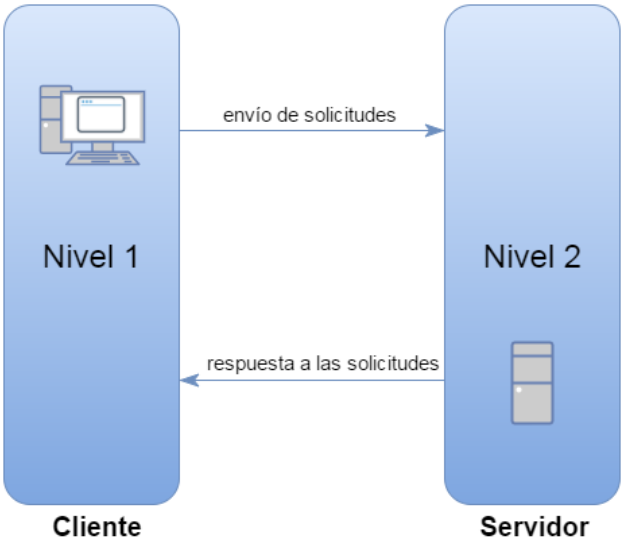

Figura 32. Arquitectura cliente-servidor.

Para la separación lógica en el servidor, se utiliza una arquitectura de 3 capas: modelo-vista-controlador (MVC), para facilitar la visualización del código y tener una mejor organización pues es indispensable manipular dos lenguajes de programación, bibliotecas para navegar la ontología y generar gráficas, así como realizar constantes consultas a la base de datos.

En la Figura 33 se muestran gráficamente las capas del patrón MVC.

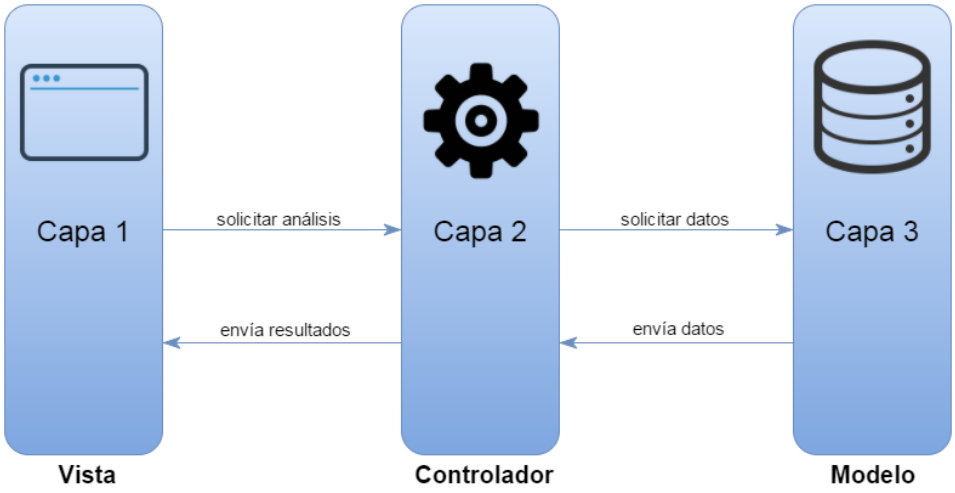

Figura 33. Modelo-Vista-Controlador

Para tener una separación entre la vista y el controlador se utilizó el *framework* Struts2 para que los JSP que se muestran en la página Web no tengan ni una línea de código Java. Los JSP programados solo contienen código de HTML5 y la siguiente línea que indica, el uso de etiquetas de Struts2.

```
<%@ taglib prefix="s" uri="/struts-tags" %>
```

En la vista se utilizó HTML5 para modelar las páginas Web (código contenido en los JSP), hojas de estilo (CSS) para dar formato a dichas páginas, códigos en JavaScript para la visualización de gráficas, así como imágenes y tipografías.

El modelo se mantiene separado del controlador al contener los procedimientos almacenados con los que puede interactuar para enviar los resultados. En ningún momento se muestran las tablas ni atributos que las componen, es decir, no hay ningún mapeo de la base de datos en el controlador. Los códigos en lenguajes Java y Python solo conocen datos como el nombre de la base, el usuario y la contraseña para poder establecer una conexión.

En el controlador se utiliza el lenguaje java para controlar todas las tareas a realizar. Hay clases como `GeneralStatisticsAction.java` que interactúan indirectamente con la vista a través del *framework*. Para interactuar con el modelo, están las clases DAO que conocen los nombres de los procedimientos almacenados, solo en estas clases se utiliza el tipo de dato `ResultSet` pues los métodos leen los resultados obtenidos de la base de datos y los convierten en cadenas, enteros, arreglos o demás. También existe una clase estática `Conexión.java`, ésta es la única que conoce los datos necesarios para establecer la conexión.

En la Figura 34 se muestra la separación de las clases utilizadas para el desarrollo del sistema.

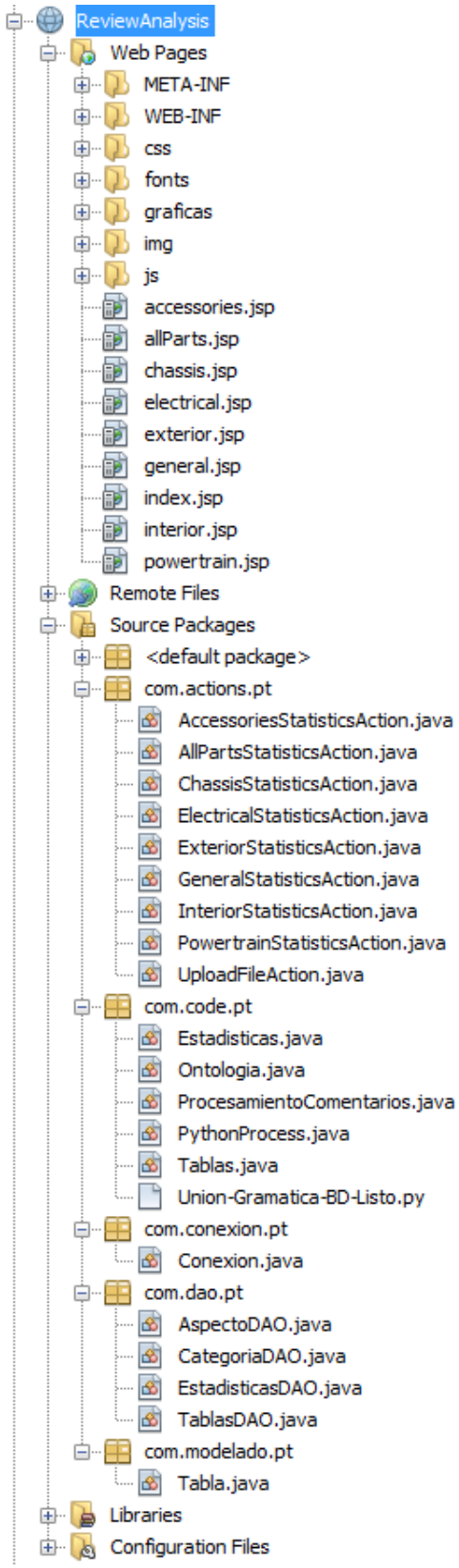

Figura 34. Organización de los archivos del proyecto.

Capítulo 6. Pruebas y resultados

6.1. Origen de los datos para las pruebas

Para probar el sistema es necesario formar archivos de comentarios que hablen de un auto en especial. Así que se buscaron foros de personas que escriben reseñas de su experiencia con su automóvil, éstas reseñas están en inglés y son textos extensos comparados por ejemplo con un *Tweet*[5]. Las direcciones Web de los sitios donde se extrajeron los comentarios son las siguientes:

- http://www.cars.com/ [18]
- http://www.edmunds.com/ [19]
- http://www.kbb.com/ [23]

En estas páginas existen reseñas de muchos modelos y marcas de automóviles. Para seleccionar qué automóviles analizar se consultó la página CarComplaints.com porque ofrece estadísticas por años de los automóviles que más quejas ha tenido, esta evaluación se hace con la puntuación que los usuarios (dueños de los automóviles) asignan al modelo. Por ejemplo para el auto Volkswagen Jetta se tiene la gráfica de la Figura 35, donde se muestra que el modelo 2006 es el que tiene una mayor cantidad de quejas.

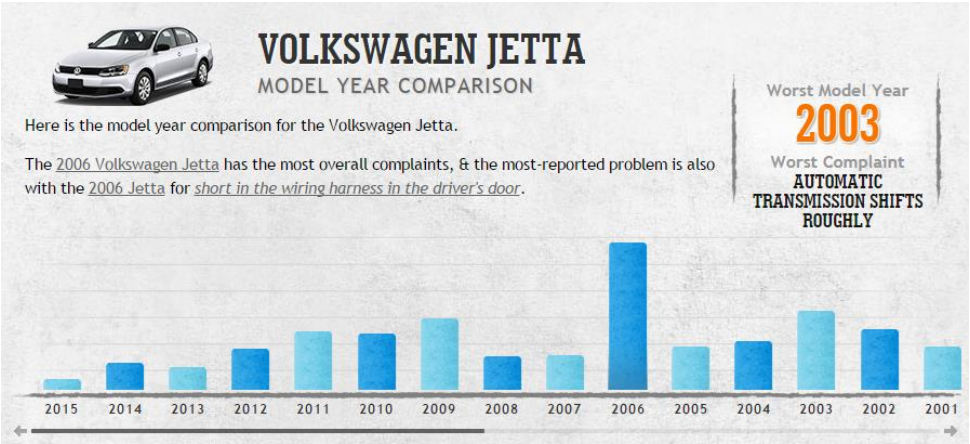

Figura 35. Estadísticas de las quejas del Jetta.

[5] Tweet es un texto corto de 140 caracteres, publicado en la red social Twitter.

Para hacer pruebas con otro automóvil se revisaron las estadísticas en CarComplains.com del Chevrolet Spark, en la Figura 36, se muestra que el modelo 2013 es el que cuenta con más quejas.

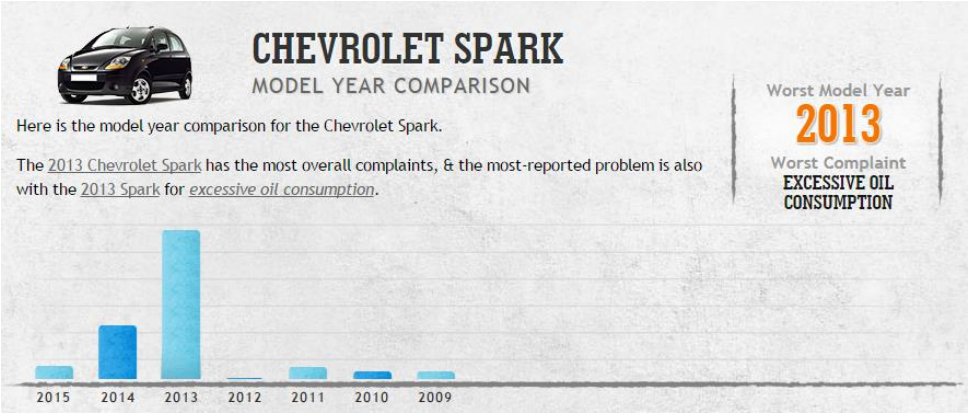

Figura 36. Estadísticas de las quejas del Spark.

Con las estadísticas anteriores se llegó a la conclusión de buscar comentarios sobre los automóviles Volkswagen Jetta 2006 y Chevrolet Spark 2013. Se seleccionaron dos marcas diferentes porque la ontología se especializa en aspectos de los automóviles de la marca Volkswagen. El objetivo de esto es observar que a pesar de realizar el análisis con un modelo que no sea de dicha marca se obtendrá un resultado.

Una vez que los archivos de los comentarios están listos, es decir, en un archivo de texto separando cada comentario con un salto de línea, comienza la interacción entre el usuario y el sistema.

6.2. Funcionamiento de la página de inicio del sistema

En la Figura 37 se muestra la página principal (número 1), el icono de la carpeta es el botón para seleccionar un archivo mientras la imagen del archivo procesándose, da inicio al análisis.

En el número 2 de la Figura 37 se muestra como se abre la ventana para buscar el archivo deseado en el directorio. Finalmente en el número 3 de dicha figura se muestra como se ve la página al haber seleccionado un archivo.

Figura 37. Página inicial del sistema (1). Selección de un archivo (2). Nombre del archivo esperando ser analizado (3).

6.3. Gráficas generadas para cada análisis

Una vez terminado el análisis se muestran las siguientes gráficas.

6.3.1. Gráficas generales

El objetivo del sistema es mostrar la cantidad de aspectos buenos y malos de categorías específicas del auto, es decir, no evaluar al producto como bueno o malo en general. Sin embargo, se observó que en muchos comentarios se tienen opiniones como:

"The VW Jetta is a nice car"

En ese tipo de sentencias se está describiendo explícitamente al producto en general además que gramaticalmente es correcto, por esta razón se grafican la cantidad de aspectos evaluados como positivos y negativos sobre el auto en general (ver Figura 38, gráfica de la izquierda). Sin embargo también se gráfica la suma de todos los aspectos de los cuales se hablaron, esto se muestra en la gráfica de la derecha en la Figura 38.

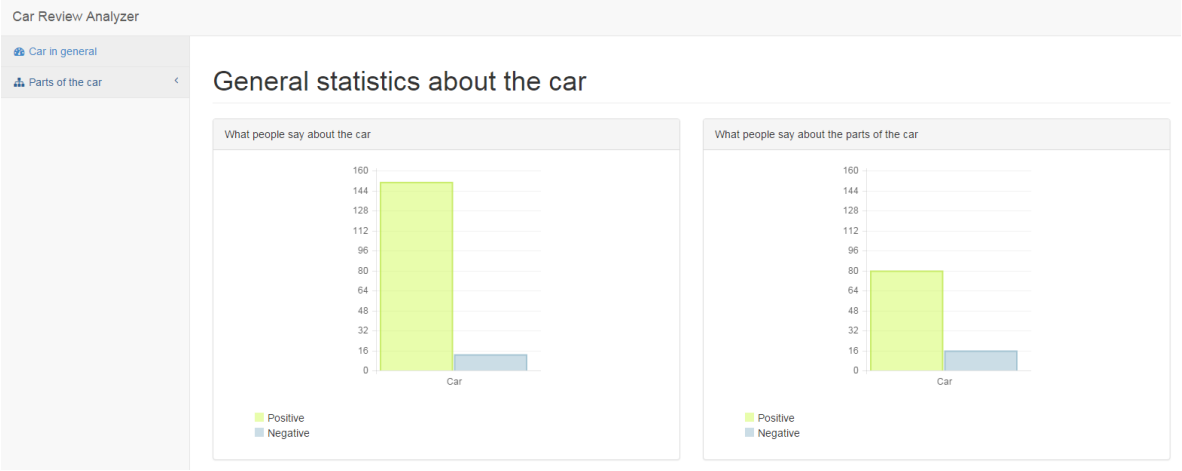

Figura 38. Gráficas generales

6.3.2. Gráfica de todas las partes del auto

En esta gráfica ya se pueden observar los resultados del primer nivel de clasificación de las partes del auto. De acuerdo con la ontología diseñada, un automóvil se divide en: accessories, chassis, electrical, exterior, interior y powertrain, por eso la gráfica tiene esas etiquetas en el eje de las abscisas como en la Figura 39. Cada etiqueta tiene una barra verde

que indica la cantidad de aspectos evaluados como positivos y una azul para los aspectos con calificación negativa.

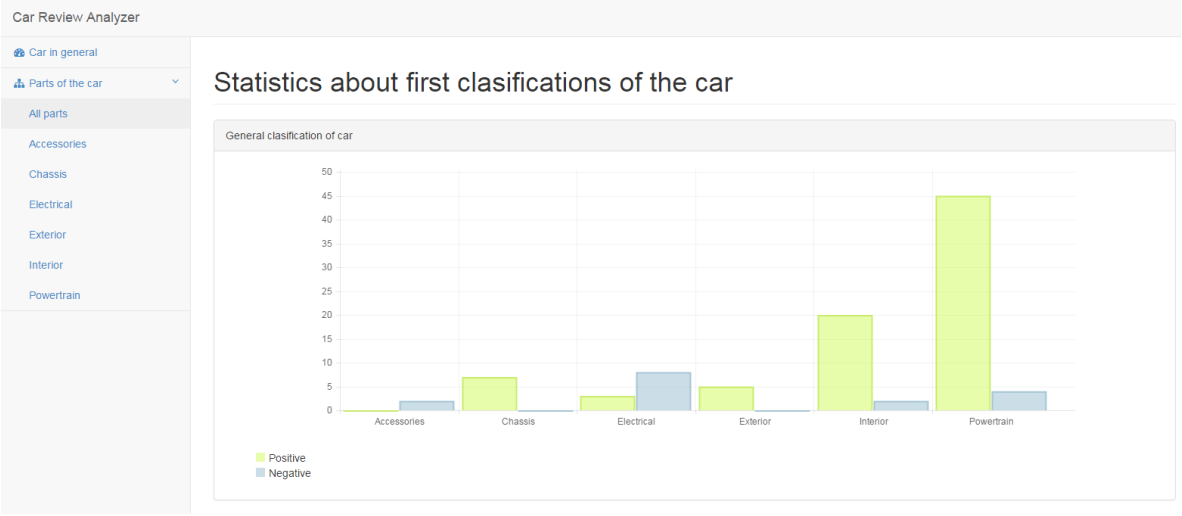

Figura 39. Gráfica de los aspectos agrupados en categorías de primer nivel.

6.3.3. Tablas de los accesorios

La categoría de accesorios no contiene más categorías pues todos los conceptos hijo son de *accessories*, por eso no puede presentarse una gráfica agrupando los aspectos y en su lugar se encuentran un par de tablas, en la primera se muestran los aspectos, sentimientos y la sentencia de donde se extrajeron los datos para aspectos calificados como positivos, en la segunda tabla se muestran los mismos datos para la aspectos negativos. (Ver Figura 40)

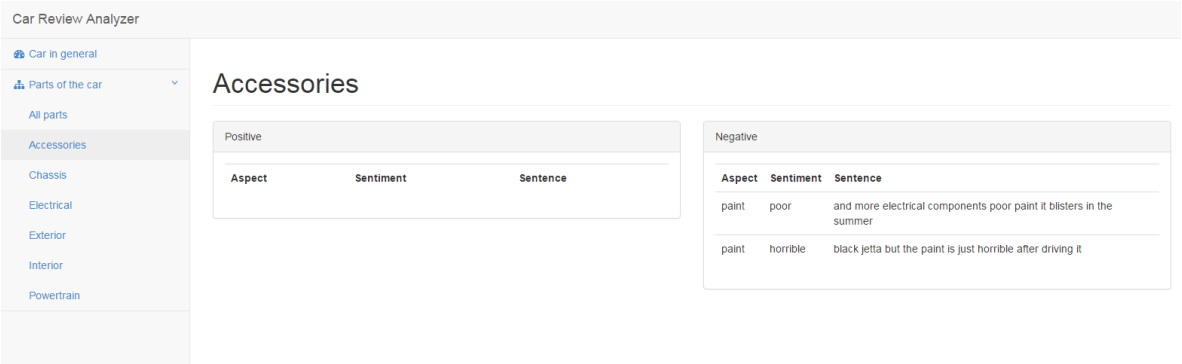

Figura 40. Tablas de la categoría *accessories*.

6.3.4. Gráfica del chasis

La categoría *chassis* está dividido en *braking*, *steering* y *suspension* así que es posible mostrar una gráfica de los aspectos positivos y negativos que pertenecen a esas categorías como se observa en la Figura 41.

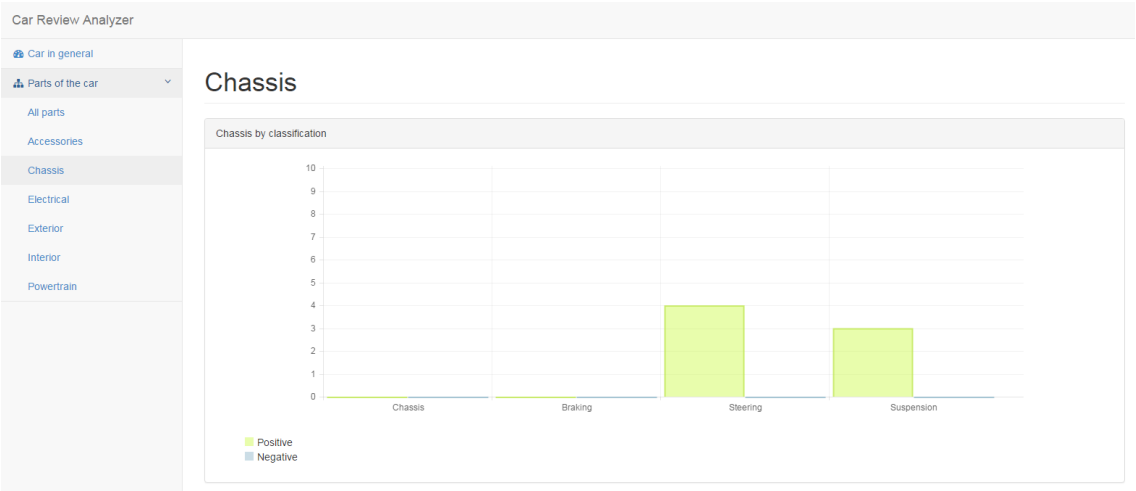

Figura 41. Gráfica de resultados de la categoría *chassis.*

6.3.5. Gráfica del sistema eléctrico

De acuerdo con la ontología *electrical* está dividido en 8 categorías: *audio, charging system, electrical supply system, gauge, ignition system, lighting, miscellaneous, sensors, starting system, switches* y *video accessories.* (Ver Figura 42).

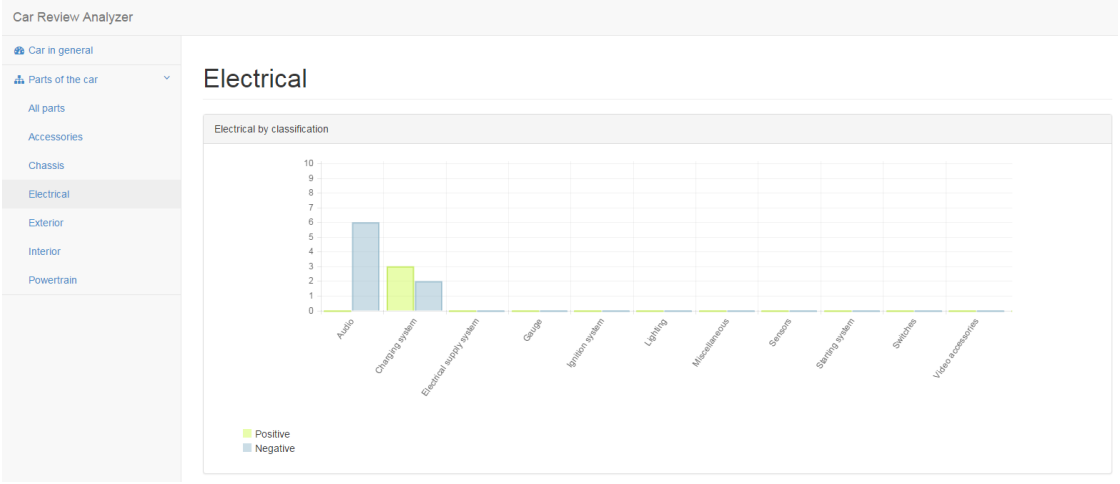

Figura 42. Gráfica de resultados de la categoría *electrical.*

6.3.6. Gráfica del exterior

La categoría *exterior* se puede dividir en *door*, *rim* y *window*, sin embargo, en muchas ocasiones las opiniones son respecto a la categoría *exterior* (ver Figura 43), por eso se muestra un par de barras para expresiones como la siguiente:

"the exterior is sleek"

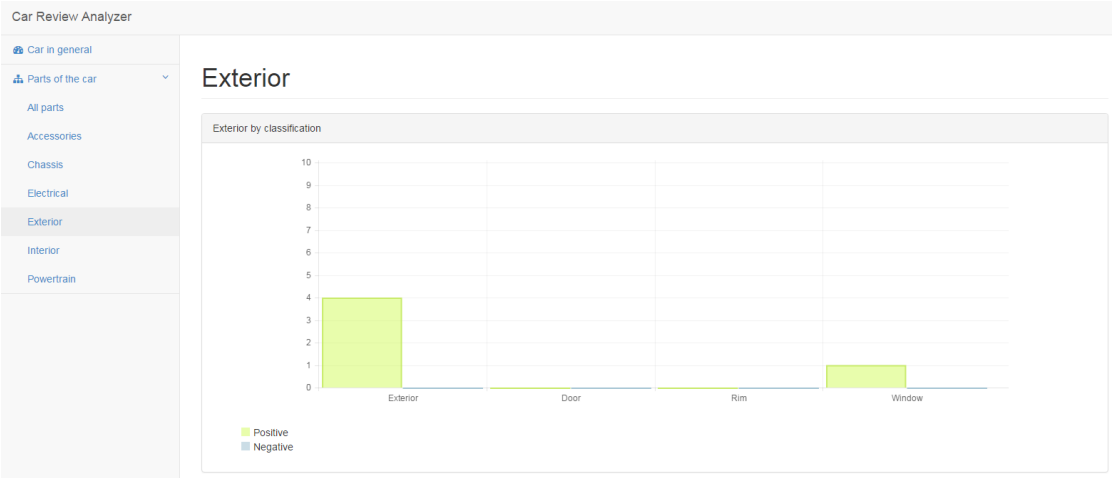

Figura 43. Gráfica de resultados de la categoría *exterior.*

6.3.7. Gráfica del interior

De igual forma que la categoría *exterior*, se habla directamente de *interior*. Por eso en la gráfica se tiene un par de barras para esta categoría, así como las categorías que contiene: *floor* y *seat*. (Ver Figura 44)

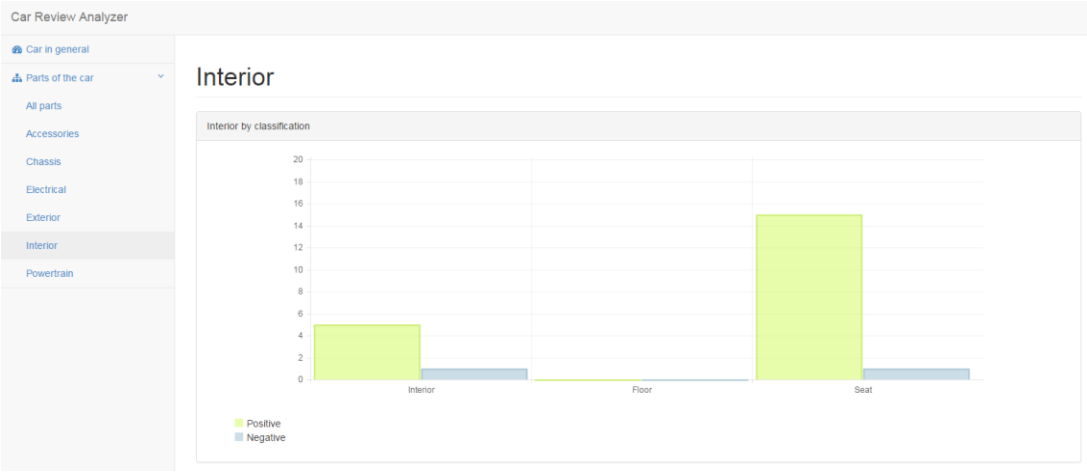

Figura 44. Gráfica de resultados de la categoría *interior.*

6.3.8. Gráfica del tren motor

Finalmente en la categoría *powertrain* se encuentran las categorías *engine*, *engine cooling system*, *enfine oil system*, *exhaust system*, *fuel supply system* y *transmission*. (Ver Figura 45)

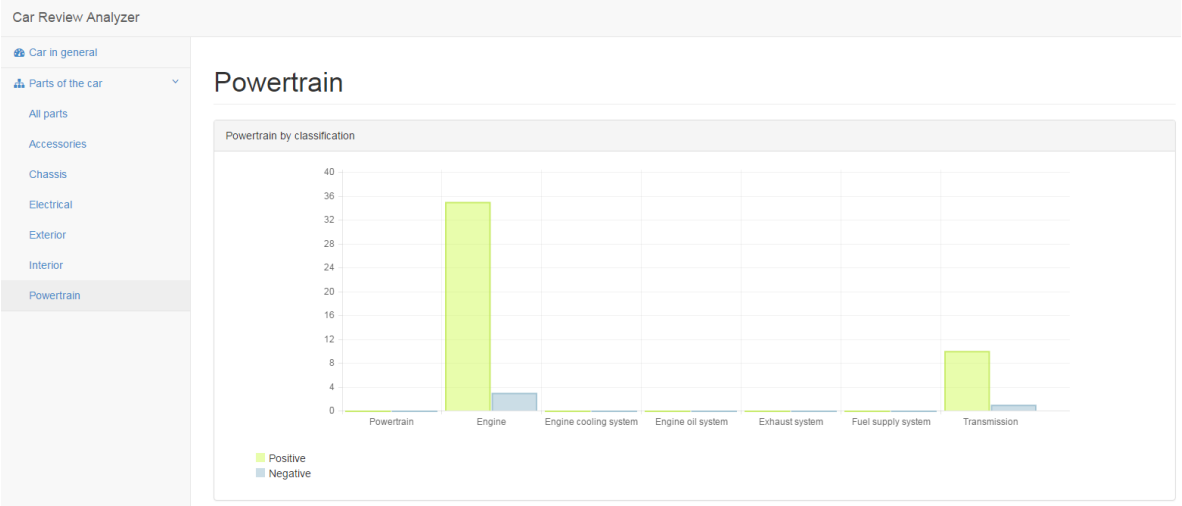

Figura 45. Gráfica de resultados de la categoría *powertrain.*

6.4. Prueba de análisis de los comentarios

Una vez que se ha mostrado cuales son las gráficas a obtener y cómo se interpretan se muestra en la Tabla 18 los nombres y la cantidad de comentarios de los archivos probados.

Tabla 18. Nombres de los archivos y cantidad de comentarios.

Jetta 2006		Spark 2013	
Nombre del archivo	Número de comentarios	Nombre del archivo	Número de comentarios
jettaCars.com.txt	57	**sparkCars.com.txt**	87
jettaEdmunds.com.txt	503	**sparkEdmunds.com.txt**	21
jettaKbb.com.txt	257	**sparkKbb.com.txt**	97

Como se mencionó anteriormente, no importa la cantidad de comentarios que se van a analizar porque solo se depende de la forma en la que escriben las personas. A continuación se muestran los resultados obtenidos con cada uno de los archivos.

6.4.1. Resultados del análisis de los comentarios del Jetta 2006

6.4.1.1 Resultados de los comentarios obtenidos de Cars.com

Para esta prueba se generó un archivo con 57 comentarios, se determinó que 306 sentencias podrían ser útiles, por lo tanto hay 306 pares aspecto-sentimiento pero algún sentimiento o aspecto puede ser duplicado por lo que en los catálogos hay 203 aspectos y 135 sentimientos. Sin embargo, al momento de asignar polaridad a los sentimientos sólo 48 son válidos, de la misma forma al asignar categoría a los aspectos sólo 25 aspectos son válidos. Al revisar las sentencias que tienen un aspecto con categoría y un sentimiento con polaridad, se tiene un total de 41 aspectos que serán contados para mostrar las gráficas de la Figura 46. Se puede observar que en la gráfica general de la izquierda hay 22 opiniones positivas del auto en general y 2 negativas. En la gráfica de la derecha muestra cómo se critican 9 aspectos como positivos y 8 negativos.

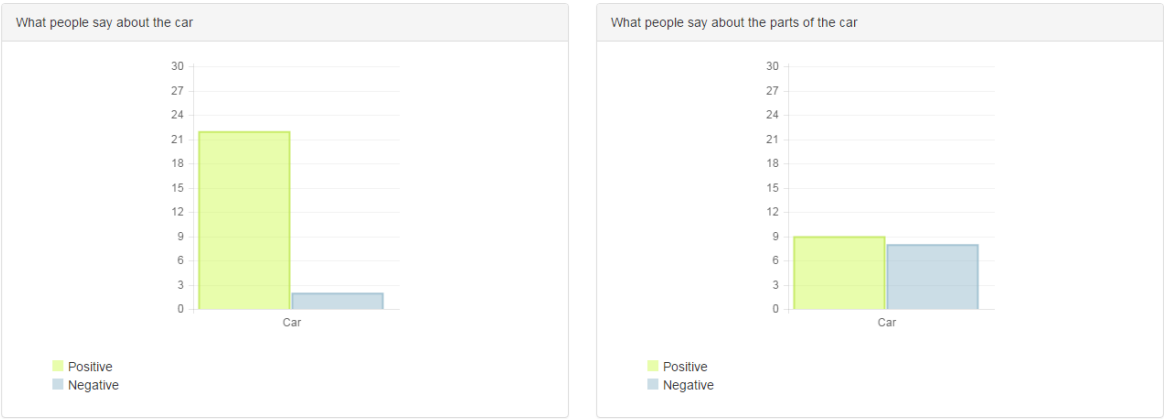

Figura 46. Gráficas generales del archivo jettaCars.com.txt.

Para saber información a detalle de los aspectos se muestra en la Figura 47 la gráfica agrupando los aspectos en las categorías del primer nivel de la ontología, en esta gráfica se habló negativamente de 1 aspecto de *accessories*, 1 positivo y 1 negativo de *electrical*, 2

aspectos buenos de *interior* y 3 malos mientras que en la categoría *powertrain* hay 6 aspectos positivos y 3 negativos.

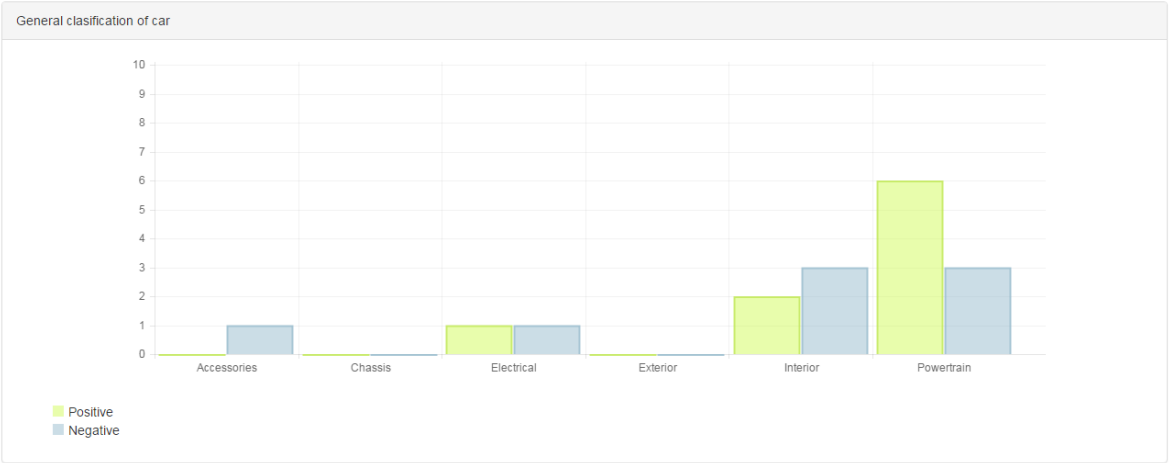

Figura 47. Gráfica de primer nivel del archivo jettaCars.com.txt**.**

Para tener mayor nivel de detalle, es decir, subcategorías de alguna de las categorías mencionadas se consulta la gráfica correspondiente, por ejemplo en la Figura 48 se muestra la gráfica de los aspectos de *powertrain*, con la cual puede concluirse que la mayoría de las opiniones sobre el tren motor es acerca del motor y la transmisión.

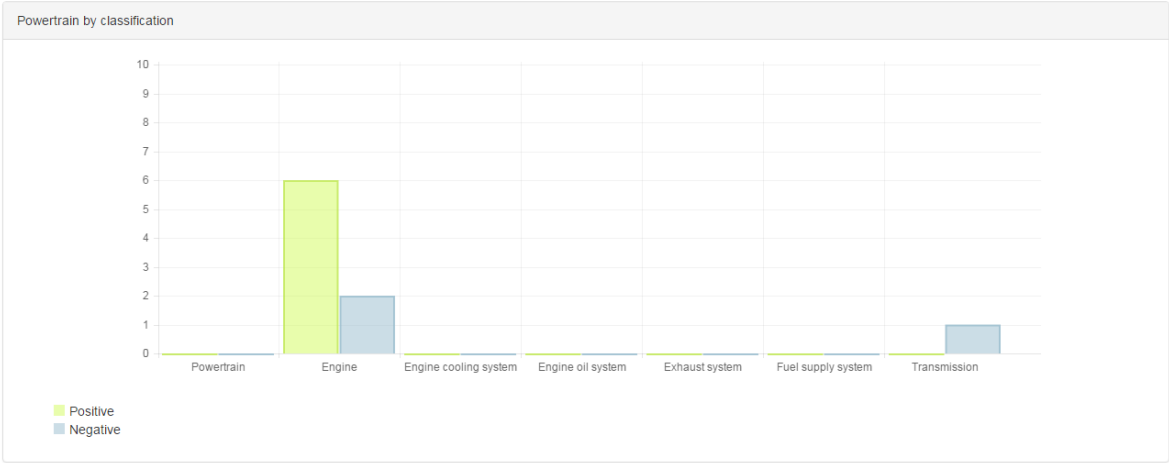

Figura 48. Gráfica de *powertrain* del archivo jettaCars.com.txt.

Finalmente se presentan en la Tabla 19 los resultados de los aspectos analizados del archivo jettaCars.com.txt.

Tabla 19. Resultados de las estadísticas del archivo jettaCars.com.txt.

Categoría (total de aspectos)	Subcategorías (total de aspectos)	Número de aspectos positivos	Número de aspectos negativos
Accessories (1)	-	-	1
Chassis (0)	Chassis (0)	-	-
	Braking (0)	-	-
	Steering (0)	-	-
	Suspension (0)	-	-
Electrical (2)	Electrical (2)	1	1
	Audio (0)	-	-
	Charging system (0)	-	-
	Electrical supply (0)	-	-
	Gauge (0)	-	-
	Ignition system (0)	-	-
	Lighting(0)	-	-
	Miscellaneous (0)	-	-
	Sensors (0)	-	-
	Starting system (0)	-	-
	Switches (0)	-	-
	Video accessories (0)	-	-
Exterior (0)	Exterior (0)	-	-
	Door (0)	-	-
	Rim (0)	-	-
	Window (0)	-	-
Interior (5)	Interior (2)	1	1
	Floor (0)	-	-
	Seat (3)	1	2
Powertrain (9)	Powertrain (0)	-	-
	Engine (8)	6	2
	Engine cooling system (0)	-	-
	Engine oil system (0)	-	-
	Exhaust system (0)	-	-
	Fuel supply system (0)	-	-
	Transmission (1)	-	1

6.4.1.2 Resultados de los comentarios obtenidos de Edmunds.com

El archivo contiene 503 comentarios de donde se extrajeron 2283 sentencias que pueden ser útiles, hay 820 aspectos almacenados en catálogo y 491 sentimientos. Las categorías en las que se agrupan los aspectos son 24.

En la Figura 49 están las gráficas de las estadísticas generales obtenidas, 164 fueron opiniones relacionadas directamente con el producto (gráfica izquierda) y otras 96 opiniones se hicieron para alguna de las partes del automóvil (gráfica derecha).

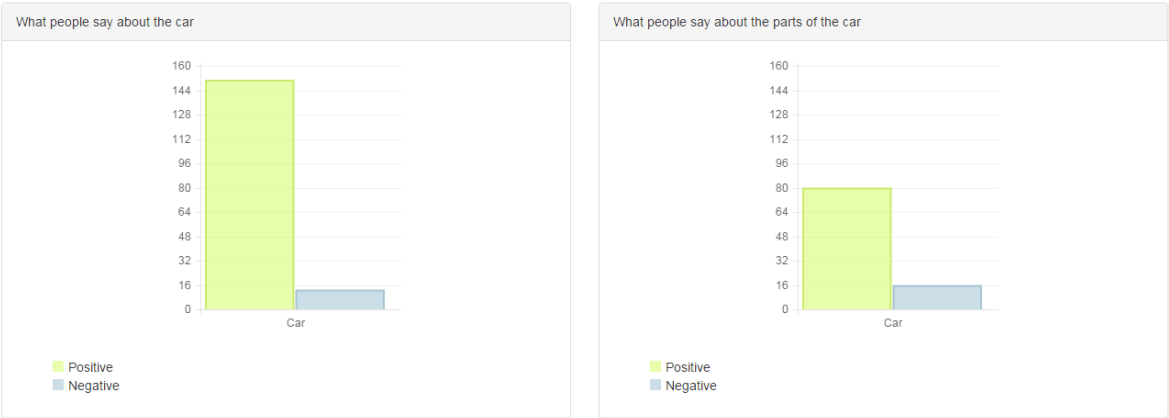

Figura 49. Gráficas generales del archivo jettaEdmunds.com.txt.

En la Figura 50 se encuentran los aspectos agrupados en las 6 categorías del primer nivel. Hay 2 opiniones malas sobre *accessories*, 7 buenas del *chassis*, 3 positivas y 8 negativas de *electrical*, 5 positivas de *exterior*, 20 positivas y 2 negativas de *interior* y por último *powertrain* con 45 positivas y 4 negativas.

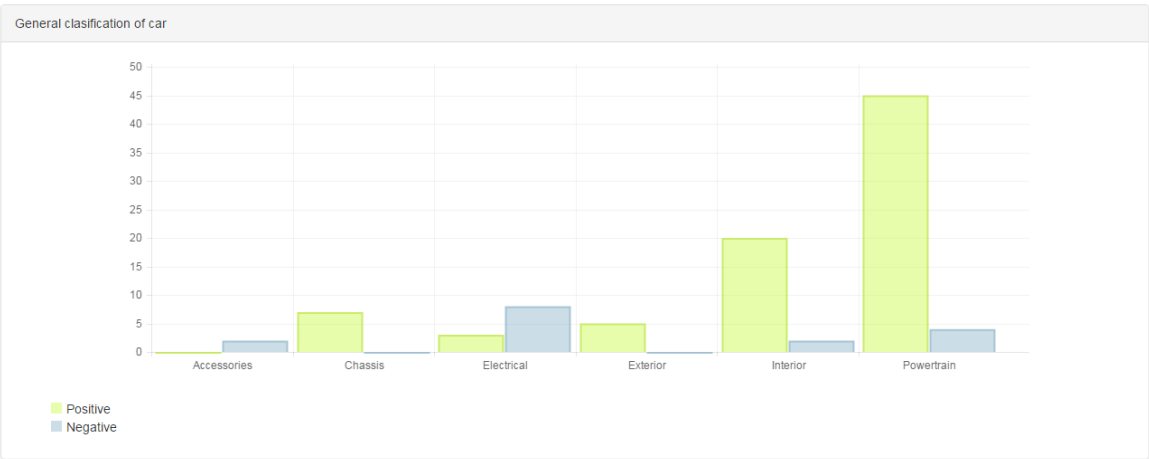

Figura 50. Gráfica de primer nivel del archivo jettaEdmunds.com.txt.

Los resultados de las gráficas de primer nivel (subcategorías de las 6 principales) se muestran en la Tabla 20.

Tabla 20. Resultados de las estadísticas del archivo jettaEdmunds.com.txt.

Categoría (total de aspectos)	Subcategorías (total de aspectos)	Número de aspectos positivos	Número de aspectos negativos
Accessories (2)	-	-	2
Chassis (7)	Chassis (0)	-	-
	Braking (0)	-	-
	Steering (4)	4	-
	Suspension (3)	3	-
Electrical (11)	Electrical (6)	-	6
	Audio (5)	3	2
	Charging system (0)	-	-
	Electrical supply (0)	-	-
	Gauge (0)	-	-
	Ignition system (0)	-	-
	Lighting(0)	-	-
	Miscellaneous (0)	-	-
	Sensors (0)	-	-
	Starting system (0)	-	-
	Switches (0)	-	-
	Video accessories (0)	-	-
Exterior (5)	Exterior (4)	4	-
	Door (0)	-	-
	Rim (0)	-	-
	Window (1)	1	-
Interior (22)	Interior (6)	5	1
	Floor (0)	-	-
	Seat (16)	15	1
Powertrain (49)	Powertrain (0)	-	-
	Engine (38)	35	3
	Engine cooling system (0)	-	-
	Engine oil system (0)	-	-
	Exhaust system (0)	-	-
	Fuel supply system (0)	-	-
	Transmission (11)	10	1

6.4.1.3 Resultados de los comentarios obtenidos de Kbb.com

Este archivo tiene 271 para su análisis. Una vez finalizado el procesamiento con el archivo se determinó que hay 1351 sentencias que podrían ser útiles. En los catálogos hay 611 aspectos y 377 sentimientos. Y las categorías que ayudan a clasificar los aspectos son 24.

Las gráficas muestran los resultados del análisis, una vez que se asignaron categorías y polaridades. En total se extrajeron datos de 139 opiniones divididas en 97 que hablan directamente del auto y 42 que lo hacen hablando de una parte en especial, como se muestra en la Figura 51.

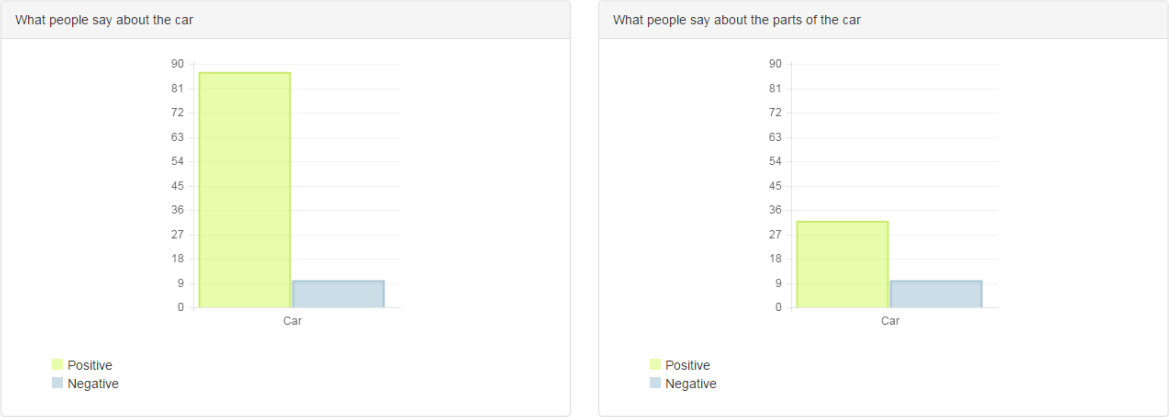

Figura 51. Gráficas generales del archivo jettaKbb.com.txt.

La mayoría de las opiniones hablan acerca del tren motor siendo 22 aspectos para esta categoría (17 positivos y 5 negativos). (Ver Figura 52)

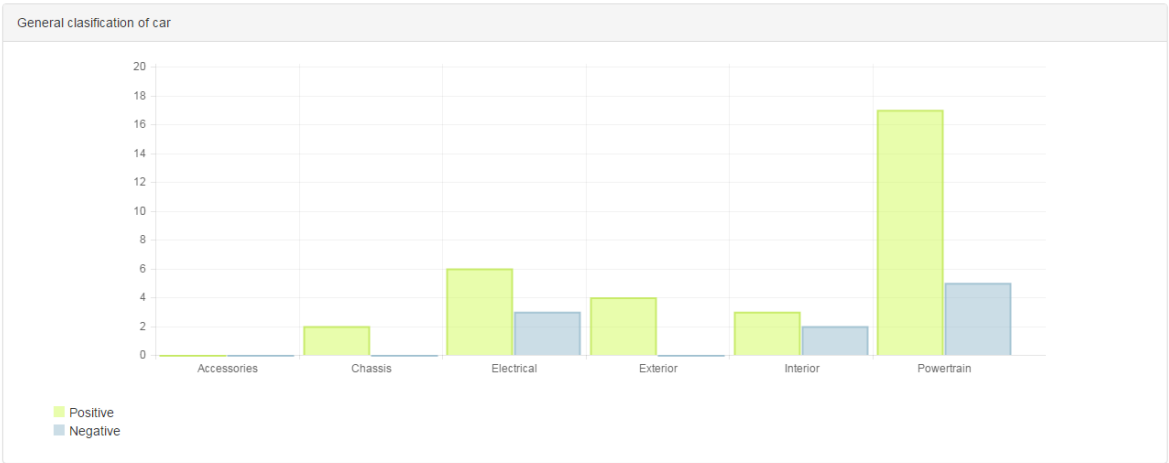

Figura 52. Gráfica de primer nivel del archivo jettaKbb.com.txt.

En la Tabla 21 están los resultados obtenidos a través de las gráficas.

Tabla 21. Resultados de las estadísticas del archivo jettaKbb.com.txt.

Categoría (total de aspectos)	Subcategorías (total de aspectos)	Número de aspectos positivos	Número de aspectos negativos
Accessories (0)	-	0	0
Chassis (2)	Chassis (0)	-	-
	Braking (1)	1	-
	Steering (0)	-	-
	Suspension (1)	1	-
Electrical (9)	Electrical (0)	-	-
	Audio (8)	6	2
	Charging system (0)	-	-
	Electrical supply (1)	-	1
	Gauge (0)	-	-
	Ignition system (0)	-	-
	Lighting(0)	-	-
	Miscellaneous (0)	-	-
	Sensors (0)	-	-
	Starting system (0)	-	-
	Switches (0)	-	-
	Video accessories (0)	-	-
Exterior (4)	Exterior (3)	3	-
	Door (0)	-	-
	Rim (0)	-	-
	Window (1)	1	-
Interior (5)	Interior (3)	2	1
	Floor (0)	-	-
	Seat (2)	1	1
Powertrain (22)	Powertrain (1)	-	1
	Engine (15)	13	2
	Engine cooling system (0)	-	-
	Engine oil system (0)	-	-
	Exhaust system (0)	-	-
	Fuel supply system (0)	-	-
	Transmission (6)	4	2

6.4.2. Resultados del análisis de los comentarios del Spark 2013

6.4.2.1 Resultados de los comentarios obtenidos de Cars.com.txt

Este archivo contiene 87 comentarios y los datos obtenidos son los siguientes:

- 540 sentencias posiblemente útiles
- 328 aspectos en catálogo
- 217 sentimientos en catálogo
- 20 categorías que agrupan a los aspectos

Las gráficas de los resultados generales se encuentran en la Figura 53.

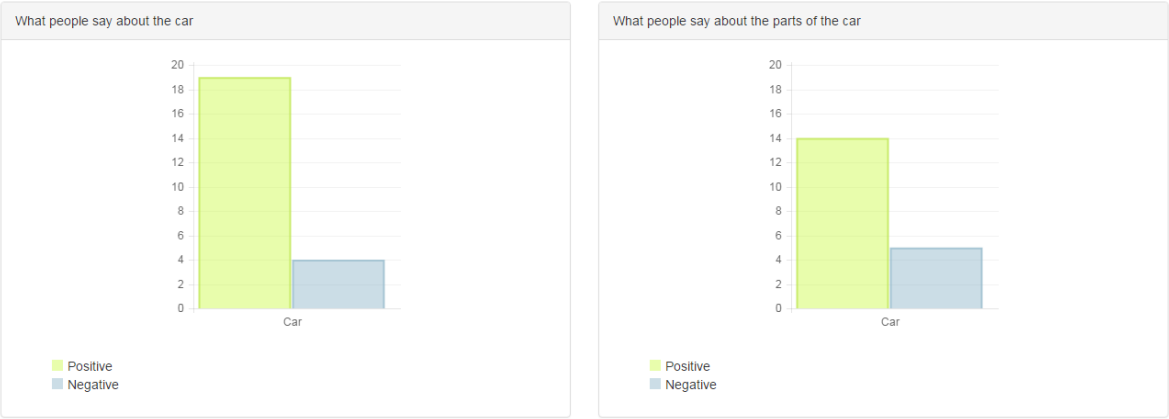

Figura 53. Gráficas generales del archivo sparkCars.com.txt.

Nuevamente con este modelo de auto se observa una gran cantidad de opiniones respecto al tren motor. (Ver Figura 54)

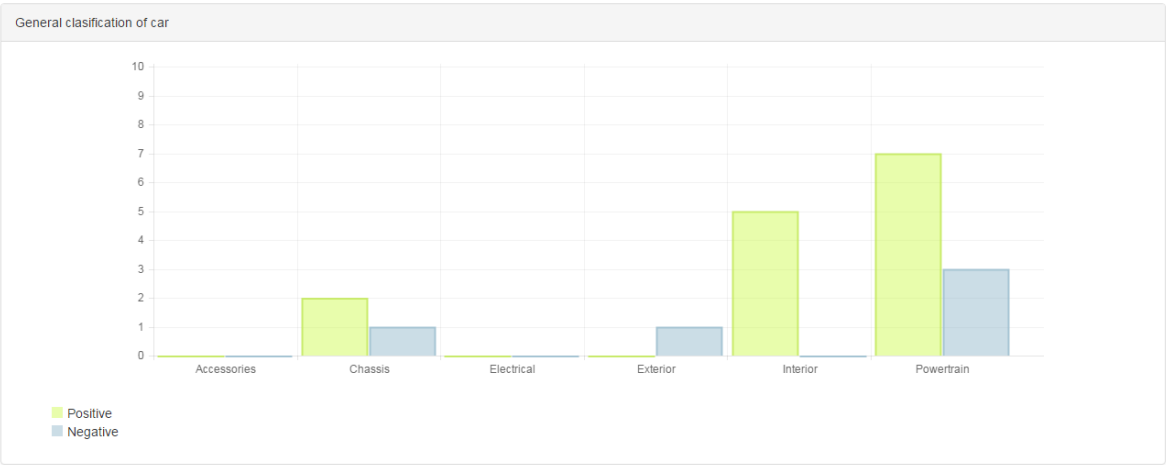

Figura 54. Gráficas de primer nivel del archivo sparkCars.com.txt.

Los resultados de las gráficas de primer nivel se encuentran en la Tabla 22.

Tabla 22. Resultados de las estadísticas del archivo sparkCars.com.txt.

Categoría (total de aspectos)	Subcategorías (total de aspectos)	Número de aspectos positivos	Número de aspectos negativos
Accessories (0)	-	-	-
Chassis (3)	Chassis (0)	-	-
	Braking (1)	1	-
	Steering (2)	1	1
	Suspension (0)	-	-
Electrical (0)	Electrical (0)	-	-
	Audio (0)	-	-
	Charging system (0)	-	-
	Electrical supply (0)	-	-
	Gauge (0)	-	-
	Ignition system (0)	-	-
	Lighting(0)	-	-
	Miscellaneous (0)	-	-
	Sensors (0)	-	-
	Starting system (0)	-	-
	Switches (0)	-	-
	Video accessories (0)	-	-
Exterior (1)	Exterior (0)	-	-
	Door (1)	-	1
	Rim (0)	-	-
	Window (0)	-	-
Interior (5)	Interior (1)	1	-
	Floor (0)	-	-
	Seat (4)	4	-
Powertrain (10)	Powertrain (0)	-	-
	Engine (8)	7	1
	Engine cooling system (0)	-	-
	Engine oil system (0)	-	-
	Exhaust system (0)	-	-
	Fuel supply system (0)	-	-
	Transmission (2)	-	2

6.4.2.2 Resultados de los comentarios obtenidos de Edmunds.com.txt

Este archivo es el más pequeño de todos, tiene 21 comentarios y los datos obtenidos son los siguientes:

- 144 sentencias posiblemente útiles
- 104 aspectos en catálogo
- 90 sentimientos en catálogo
- 9 categorías que agrupan a los aspectos

En la Figura 55 se muestran las gráficas generales de los resultados en donde se muestra que a pesar de ser una cantidad pequeña de comentarios se obtienen resultados de los aspectos agrupados en categorías.

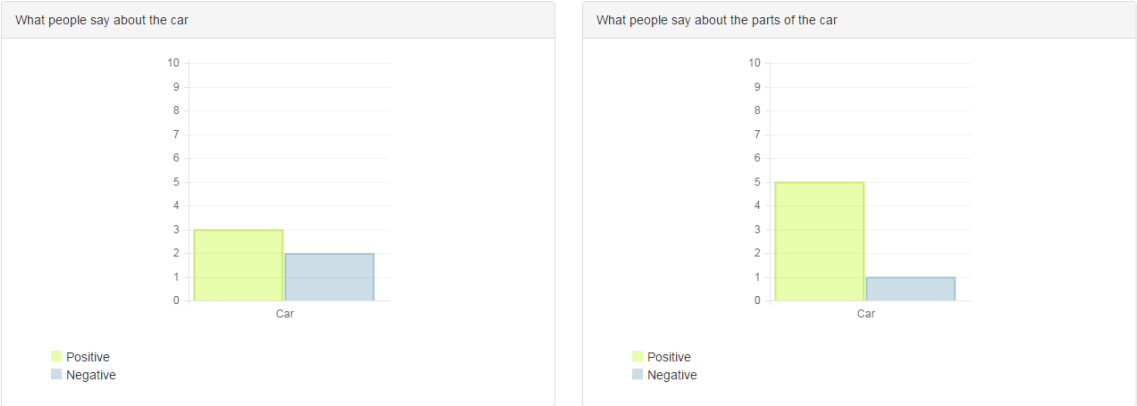

Figura 55. Gráficas generales del archivo sparkEdmunds.com.txt.

Los aspectos contados en la gráfica general de la Figura 56 (derecha) corresponden a las categorías *interior* y *powertrain*.

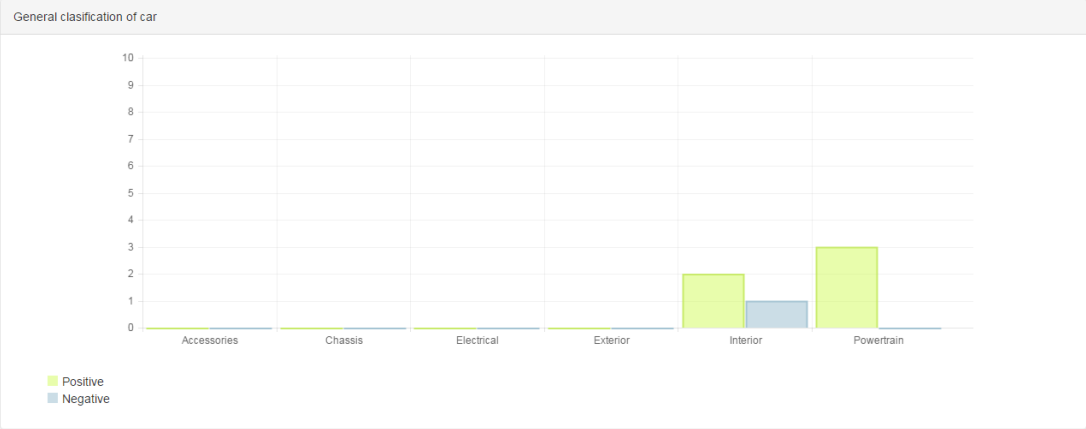

Figura 56. Gráfica de primer nivel del archivo sparkEdmunds.com.txt.

Los resultados del resto de las gráficas se muestran en la Tabla 23.

Tabla 23. Resultados de las estadísticas del archivo sparkEdmunds.com.txt.

Categoría (total de aspectos)	Subcategorías (total de aspectos)	Número de aspectos positivos	Número de aspectos negativos
Accessories (0)	-	-	-
Chassis (0)	Chassis (0)	-	-
	Braking (0)	-	-
	Steering (0)	-	-
	Suspension (0)	-	-
Electrical (0)	Electrical (0)	-	-
	Audio (0)	-	-
	Charging system (0)	-	-
	Electrical supply (0)	-	-
	Gauge (0)	-	-
	Ignition system (0)	-	-
	Lighting(0)	-	-
	Miscellaneous (0)	-	-
	Sensors (0)	-	-
	Starting system (0)	-	-
	Switches (0)	-	-
	Video accessories (0)	-	-
Exterior (0)	Exterior (0)	-	-
	Door (0)	-	-
	Rim (0)	-	-
	Window (0)	-	-
Interior (3)	Interior (1)	1	-
	Floor (0)	-	-
	Seat (2)	1	1
Powertrain (3)	Powertrain (0)	-	-
	Engine (3)	3	-
	Engine cooling system (0)	-	-
	Engine oil system (0)	-	-
	Exhaust system (0)	-	-
	Fuel supply system (0)	-	-
	Transmission (0)	-	-

6.4.2.3 Resultados de los comentarios obtenidos de Kbb.com

Este archivo contiene 97 comentarios y los datos obtenidos son los siguientes:

- 600 sentencias posiblemente útiles
- 317 aspectos en catálogo
- 213 sentimientos en catálogo
- 15 categorías que agrupan a los aspectos

Los resultados generales se muestran en la Figura 57.

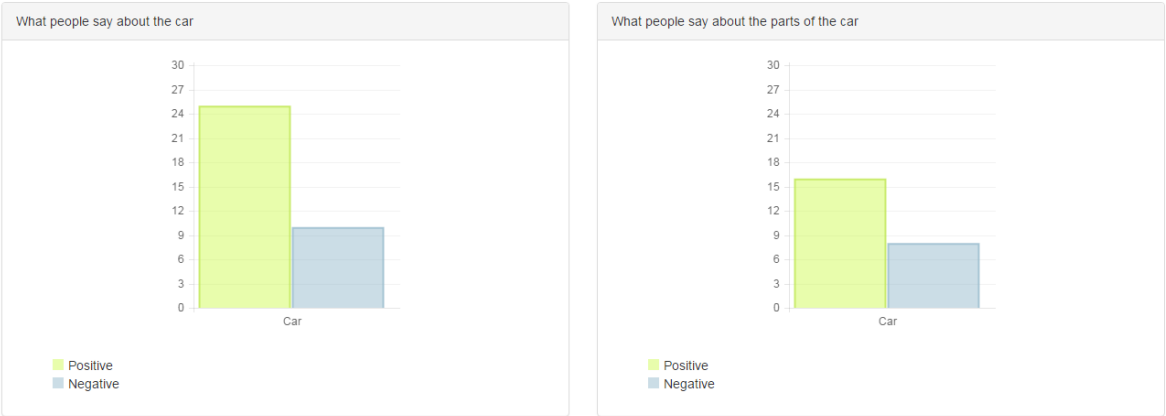

Figura 57. Gráficas generales del archivo sparkKbb.com.txt.

Las opiniones de los aspectos agrupados en las 6 principales categorías se encuentran en la Figura 58.

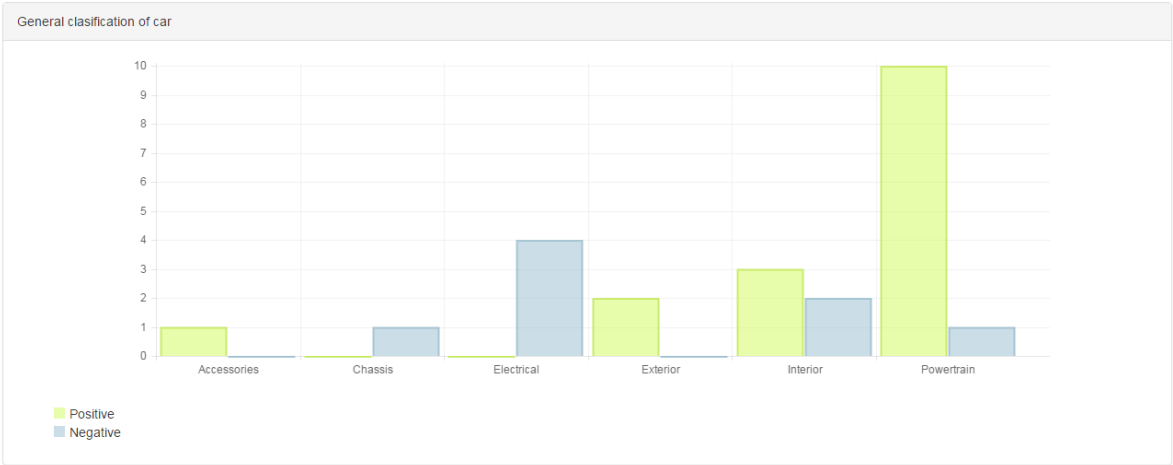

Figura 58. Gráfica de primer nivel del archivo sparkKbb.com.txt.

En la Tabla 24 están los resultados de todas las gráficas obtenidas en el análisis de este archivo.

Tabla 24. Resultados de las estadísticas del archivo sparkKbb.com.txt.

Categoría (total de aspectos)	Subcategorías (total de aspectos)	Número de aspectos positivos	Número de aspectos negativos
Accessories (1)	-	1	-
Chassis (1)	Chassis (0)	-	-
	Braking (0)	-	-
	Steering (0)	-	-
	Suspension (1)	-	1
Electrical (4)	Electrical (0)	-	-
	Audio (3)	-	3
	Charging system (0)	-	-
	Electrical supply (0)	-	-
	Gauge (0)	-	-
	Ignition system (0)	-	-
	Lighting(1)	-	1
	Miscellaneous (0)	-	-
	Sensors (0)	-	-
	Starting system (0)	-	-
	Switches (0)	-	-
	Video accessories (0)	-	-
Exterior (2)	Exterior (2)	2	-
	Door (0)	-	-
	Rim (0)	-	-
	Window (0)	-	-
Interior (5)	Interior (0)	-	-
	Floor (0)	-	-
	Seat (5)	3	2
Powertrain (11)	Powertrain (0)	-	-
	Engine (10)	9	1
	Engine cooling system (0)	-	-
	Engine oil system (0)	-	-
	Exhaust system (0)	-	-
	Fuel supply system (0)	-	-
	Transmission (1)	1	-

Conclusiones

De acuerdo con los resultados obtenidos en el Capítulo 6 se puede concluir que la cantidad de comentarios que se encuentran en el archivo a ser analizado no es de relevancia, es decir, no es necesario tener demasiados comentarios para obtener un resultado. Es claro que con mayor cantidad de comentarios, mayor información será extraída, sin embargo, a pesar de tener pocos comentarios se obtendrá una respuesta.

Las gráficas generales ayudan a comparar las opiniones que fueron dadas directamente al automóvil y aquellas que lo hicieron de forma indirecta, por ejemplo, mencionando que el motor falló. A pesar de que muchas personas describen su auto en general, otras más se esmeran en detallar las partes del auto con las que tuvieron una experiencia buena o mala. Esto ayuda a reforzar la importancia de la detección automática de aspectos.

Este sistema comenzó con la idea de ayudar a personas especializadas en *marketing* para analizar su mercado, sin embargo, la información extraída es muy útil para personas que desean hacer una futura compra, estas personas están en búsqueda de recomendaciones de modelos de automóvil y al contar con demasiada información, ésta herramienta es de gran ayuda.

Las herramientas que se utilizan actualmente para llevar a cabo el análisis de sentimientos en comentarios, implementan algoritmos de clasificación, esto obliga a realizar un entrenamiento de la máquina que requiere millones de datos que se encuentren validados; en comparación con dichas herramientas este trabajo no necesita ningún tipo de entrenamiento, para entregar resultados.

El profesor Bing Liu ha escrito artículos sobre la extracción de aspectos, donde menciona que al tener una gran cantidad de textos, puede asumirse que las palabras que más se repiten son los aspectos de los cuales hablan las personas. Comparando el sistema desarrollado puede mencionarse que no es necesario tener una gran cantidad de comentarios para saber algunas características del producto.

Para el procesamiento del lenguaje natural existen varias herramientas que proveen de grandes vocabularios, como los diccionarios, sin embargo, al intentar utilizar el recurso *WordNet* para la asignación de categorías se observó que es más complicado porque es necesario realizar más pasos en el pre-procesamiento como una correcta lematización de las palabras y esto no puede garantizarse porque los comentarios no siempre estarán bien escritos. Por esto se concluye que el uso de la ontología fue un buen recurso para que los aspectos (palabras aisladas) comprendieran su relación con otros conceptos.

Al comenzar con la búsqueda de las estructuras gramaticales que servirían para el análisis se tenía la hipótesis de que éstas serían difíciles de encontrar, debido a los diversos estilos que tienen las personas al momento de escribir. Finalmente se puede concluir que a pesar de que las personas escriben diferente unas de otras, respetan las reglas gramaticales, algunas palabras pueden estar mal escritas pero para transmitir ideas similares lo hacen siguiendo estructuras gramaticales simples. Gracias a esto se lograron definir las reglas gramaticales que permitieron extraer los aspectos.

Trabajos a futuro

Una vez finalizado el sistema se observó que hay varias tareas que pueden compleméntalo y mejorarlo. Estas tareas se detallan a continuación:

- Analizar más estructuras gramaticales

  Este sistema se enfoca en la búsqueda de sentencias que tengan una de las dos estructuras gramaticales con las que se trabaja. Estas estructuras no son las únicas que proporcionan información de los aspectos de un producto; algunas veces cuando se describe un objeto se realiza un listado de adjetivos separados por comas, o pueden llegar a realizarse comparaciones entre productos mencionando que uno es mejor que otro.

- Ampliar y detallar la ontología

  La ontología que pretendía utilizarse es *Volkswagen Vehicles Ontology* [8] sin embargo no tenía la profundidad deseada ni la cantidad de conceptos que se necesitaban, por eso se creó la ontología *Car* que describe cada una de las partes del automóvil. Por eso el uso de la ontología está abierto para agregar más conceptos y sinónimos que ayuden a la generalización de la ontología a varias marcas de autos que utilicen un nombre en especial para una parte del auto.

- Crear y agregar un *crawler*[6]

  Los comentarios se buscan y limpian manualmente, por eso un gran aporte sería crear un *crawler* que extraiga reseñas de sitios Web, para ser posteriormente analizados por el sistema.

- Detección automática del producto y comparación de modelos

  Cuando el usuario proporciona su archivo con comentarios está seguro que estos comentarios hablan del modelo que desea, sin embargo, detectar el auto del que se habla ayudaría para que posteriormente llegue a realizarse una comparación entre varios modelos de automóviles.

[6] *Crawler*: Es un programa que visita sitios Web para leer y extraer su información.

- Manejo de sesiones

  El sistema tiene una base de datos que almacena aspectos, sentimientos, categorías, comentarios y sentencias. Hasta ahora, cada que el análisis de un archivo comienza se eliminan los datos del análisis anterior, así que podría realizarse el manejo de sesiones para que un usuario registrado pueda consultar los resultados de los análisis que ha realizado. Con el uso frecuente de los usuarios se tendía una base de datos cada vez más extensa que daría pie a crear un corpus.

Referencias

[1] C. Acevedo, R. Clorio, R. Zagal, C. García,"Arquitectura Web para análisis de sentimientos en Facebook con enfoque semántico", Research in Computing Science, vol. 75, pp 56-69, Agosto 2014.

[2] S. Bird, E. Klein, E. Loper, "Natural Languaje Processing with Python", J. Steele, O'Reilly, USA, Junio 2009.

[3] G. Booch, J. Rumbaugh e I. Jacobson, "El lenguaje unificado de modelado", España: Addison-Wesley, 2006.

[4] A. Gelbukh y G. Sidorov, "Procesamiento automático del español con enfoque en regursos léxicos grandes", México: SEP, CONACyT, 2006.

[5] R. González Rojas y P. Pineda Barreto, "Sistema de análisis del impacto de productos a través de las redes sociales Facebook y Twitter desde la perspectiva del usuario", México, 2014.

[6] J. E. Gómez Balderas. "Resolución automática de la homonimia morfológica para el español". Tesis de maestría. CIC-IPN México D.F. 2007.

[7] A. Guzmán Aguilar y L. Mancilla Escamilla, "Generación automática de resúmenes de opinión sobre entidades nombradas en español", México, 2013.

[8] M. Hepp, "Volkswagen Vehicles Ontology" [En línea]. Disponible en: http://www.volkswagen.co.uk/. [Último acceso: Mayo 2015].

[9] P. Hitzlet, M. Krötzsch, S. Rudolph, "Foundations of Semantic Web Technologies", CRC Press, 2010.

[10] B. Lui, "A list of positive and negative opinion words or sentiment words for English". [En línea]. Disponible en: http://www.cs.uic.edu/~liub/. [Último acceso: Junio 2015].

[11] B. Liu, "Sentiment Analysis and Opinion Mining", Morgan & Claypool Publishers, 2012.

[12] C. Muñoz Razo. "Como elaborar y asesorar una investigación de tesis". Ciudad de México: Prentice Hall, 2011.

[13] D. Salgado «Reputación online. Una metodología» 2013. [En línea]. Disponible en: http://www.effectivetalk.es/blog/. [Último acceso: Febrero 2015].

[14] G. Silverman, "Los secretos del marketing boca a boca", New York: Grupo Norma Editorial, 2012.

[15] R. Socorro, A. Simón, R. Valdés, F. O. Fernández, A. Rosete, M. Moreno, E. Leyva, J. Pina, "Las ontologías en la representación del conocimiento". [En línea]. Disponible en: https://nopiedra.files.wordpress.com/. [Último acceso: Mayo 2015].

[16] AMIPCI, "Estudio de marketing digital y social media". Mente digital, 2014.

[17] Apache Jena [En línea]. Disponible en: https://jena.apache.org/. [Último acceso: Mayo 2015].

[18] Cars.com [En línea]. Disponible en https://www.cars.com. [Último acceso: Diciembre 2015].

[19] Edmunds.com [En línea]. Disponible en http://www.edmunds.com. [Último acceso: Diciembre 2015].

[20] «El Poder de la Recomendación Boca a Boca en las Decisiones de Compra del Consumidor Mexicano» 2007. [En línea]. Disponible en: http://www.leventergroup.com/. [Último acceso: Enero 2015].

[21] Google Think Insights, "¿Cómo deciden los mexicanos comprar un auto nuevo?" México, 2013.

[22] I.C. Society, «IEEE Recomendation Practices for Software Requirements Specification ANSI/IEEE 830, » 2009.

[23] Kelley blue book oficial gide [En línea]. Disponible en http://www.kbb.com. [Último acceso: Diciembre 2015].

[24] Merca2.0. [En línea]. Disponible en: http://www.merca20.com. [Último acceso: Marzo 2015].

[25] NLTK 3.0 documentation [En línea]. Disponible en http://www.nltk.org. [Último acceso: Agosto 2015].

[26] Python [En línea]. Disponible en: https://www.python.org. [Último acceso: Agosto 2015].

[27] W3C, "Ontologies: What is a vocabulary?" [En línea]. Disponible en: http://www.w3.org/standards/semanticweb/ontology

[28] W3C, "OWL Ontology Web Language Overview" [En línea]. Disponible en: http://www.w3.org/TR/owl-features/. [Último acceso: Mayo 2015].

[29] WordNet a lexical database for English [En línea]. Disponible en: https://wordnet.princeton.edu. [Último acceso: Agosto 2015].

[30] WordNet Domains WordNet-Affects [En línea]. Disponible en: http://wndomains.fbk.eu/wnaffect.html. [Último acceso: Agosto 2015].